1 Regression analysis - basics

Regression, one of the most widely used statistical technique, estimates relationships among variables. Regression models provide a very flexible framework for describing and testing hypotheses about relationships between explanatory variables and a response variable.

Typically, a regression analysis is used for the following purposes:

(1) modeling the relationship between variables.

(2) prediction of the target variable (forecasting).

(3) testing of hypotheses.

The basis of regression analysis is the linear model. The model can be characterized as follows. We have \( n \) sets of observations \( \{X_{1i}, X_{2i}, \ldots, X_{pi}, Y_i\}, i = 1, \ldots, n \), which represent a random sample from a larger population. It is assumed that these observations satisfy a linear relationship

\[
Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_p X_{pi} + \epsilon_i, \quad i = 1, \ldots, n
\]

where the \( \beta \) coefficients are unknown parameters, and the \( \epsilon_i \) are random error terms.

Note: By a linear model, it is meant that the model is linear in the parameters.

Linear models may not represent a true representation of reality; but perhaps it provides a useful representation of reality and works well for a wide range of circumstances.

George Box: “All models are wrong, but some are useful.”

John Tukey: “Embrace your data, not your models.”

Ref. deterministic (or functional) relationships vs statistical relationships

eg. Fahrenheit = \( \frac{9}{5} \) Celsius + 32

Circumference = \( \pi \times \) diameter

\( E = ms^2 \), \( E \): Energy, \( m \): mass, \( s \): speed of light
Height and weight, Driving speed and gas mileage

I. Simple linear regression:

a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.

One variable \((X)\) is regarded as the predictor, explanatory, or independent variable.

The other variable, denoted \((Y)\), is the response, outcome, or dependent variable.

Simple linear model (SLM):

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim \text{uncorrelated } (0, \sigma^2) \]

The first step in any analysis is to look at the data; in the regression context, that means looking at histograms and a scatter plot. Estimating the unknown parameters \(\beta_0\) and \(\beta_1\) corresponds to putting a straight line through the point cloud in the scatter plot.

Q. “What is the best fitting line?”

The standard approach is the **least squares** regression, where the estimates are chosen to minimize

\[ \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2. \]

This is a standard calculus problem, and was solved for the first time either by Legendre in 1805, or by Gauss in 1794 (Legendre published first). It can be shown that the least squares estimates satisfy

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \]
\[
\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}}
\]

- Special notation for simplicity:

\[
S_{yy} = \sum_i (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}
\]
\[
S_{xx} = \sum_i (X_i - \bar{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}
\]
\[
S_{xy} = \sum_i^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_iY_i - \frac{(\sum X_i)(\sum Y_i)}{n}
\]
Terminology

• The least squares parameter (regression coefficients) estimates: \( \hat{\beta}_0, \hat{\beta}_1 \)

• The predicted (or fitted) values: \( \hat{Y}_i = \hat{\beta}_0 - \hat{\beta}_1 X_i \) for \( i = 1, \ldots, n \)

• The residuals for data point \( i \): \( e_i = Y_i - \hat{Y}_i \)

• The squared prediction error for data point \( i \): \( e_i^2 = (Y_i - \hat{Y}_i)^2 \)

• Sum of Squared Error (SSE):

\[
SSE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2
\]

cf. MSE = \( \frac{SSE}{n-2} \): Mean Squared Error

• Because the formulas for \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are derived using the least squares criterion, the resulting equation often referred to as the “least squares regression line”, or simply the “least squares line”. It is also sometimes called the estimated regression equation.

We need a few assumptions in order to justify using the least squares regression:

(a) the expected value of the errors is zero. i.e. \( E(\epsilon_i) = 0 \) for all \( i \). (or the mean of the response \( E(Y_i) \) at each value of the predictor \( X_i \) is a linear function of \( X_i \).)

(b) errors are uncorrelated with each other (independent).

(c) errors are normally distributed (confidence/prediction intervals or hypothesis tests).

(d) the variance of the errors is constant (equal). i.e. \( \text{Var}(\epsilon_i) = \sigma^2 \) for all \( i \).

Gauss-Markov theorem: under the above conditions of the simple linear model (SLM) the least squares estimators are unbiased and have minimum variance among all unbiased linear estimators.
Example. Effects of ozone pollution on soybean yield

Four dose levels of ozone and the resulting mean seed yield of soybeans are given. The dose of ozone is the average concentration (parts per million, ppm) during the growing season. Yield is reported in grams per plant.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ozone (ppm)</td>
<td>Yield (gm/plt)</td>
</tr>
<tr>
<td>.02</td>
<td>242</td>
</tr>
<tr>
<td>.07</td>
<td>237</td>
</tr>
<tr>
<td>.11</td>
<td>231</td>
</tr>
<tr>
<td>.15</td>
<td>201</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>$\hat{Y}_i$</th>
<th>$e_i$</th>
<th>$e_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>242</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>237</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>231</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>201</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1.1 INTERVAL ESTIMATES and HYPOTHESES TESTS

In order to construct confidence intervals or test hypotheses it is necessary to make some assumptions about the distribution of $\epsilon_i, i = 1, ..., n$. The usual assumption is that

$$\epsilon_i \sim \text{Normal}(0, \sigma^2),$$

i.e. $\epsilon_i$ is normally distributed, independent of all other $\epsilon_j$ and all have the same variance, $\sigma^2$, and mean zero.

Since the random errors are unknown, the residuals can be used to get an estimate of the error variance:

$$\frac{1}{n-p} \sum_{i=1}^{n} \epsilon_i^2 = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \hat{Y})^2 = \text{MSE}$$

Properties of least squares estimates

We consider the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for } i = 1, ..., n$, where $\epsilon_i \sim \text{i.i.d. Normal (0, } \sigma^2\text{)}$.

(a) $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$; unbiased

(b) $\text{Var}(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{X^2}{S_{xx}}\right)\sigma^2$

(c) $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

(d) $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{X}{S_{xx}}$

(e) $\text{E(MSE)} = \sigma^2$, where $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2}$

(f) Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed. (MSE is independent of both $\hat{\beta}_1$ and $\hat{\beta}_1$.)

Since $\hat{\beta}_1 \sim \text{Normal}(\beta_1, \sigma^2/S_{xx})$,

$$t^* = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{S_{xx}}/\sigma}{\sqrt{\frac{(n-2)\text{MSE}}{\sigma^2} / (n-2)}} \sim t_{n-2}.$$

Then, the $100(1 - \alpha)\%$ confidence interval for $\beta_1$ is $\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}} \sqrt{\text{MSE}/S_{xx}}$. 
Under the normality assumption, in general, the $100(1 - \alpha)\%$ confidence interval for $\beta_k$ is

$$\hat{\beta}_k \pm t_{1-\frac{\alpha}{2}} \frac{s(\hat{\beta}_k)}{\sqrt{n}}, \quad k = 1, \ldots, p$$

where $s(\hat{\beta}_k)$ is a standard error (estimate) of $\hat{\beta}_k$.

To test a one-sided hypothesis: $H_0 : \beta_k \geq 0$ vs. $H_a : \beta_k < 0$

Reject $H_0$ if

$$t^* = \frac{\hat{\beta}_k - 0}{s(\hat{\beta}_k)} < t_{\alpha,n-p};$$

otherwise do not reject $H_0$.

A $100(1 - \alpha)\%$ confidence interval for $\sigma^2$ is given by $\left(\frac{(n-2)\text{MSE}}{\chi^2_{n-2,1-%frac{\alpha}{2}}}, \frac{(n-2)\text{MSE}}{\chi^2_{n-2,%frac{\alpha}{2}}}\right)$.

CONFIDENCE INTERVALS FOR THE MEAN OF $Y$ : One very useful application of the preceding result is to the problem of estimating the mean value of $Y$ at a fixed value of $X$, say, $X_0$. In our straight-line regression setting,

$$E(Y) \text{ at } x = X_0 = \beta_0 + \beta_1 X_0$$

which is just a linear combination of $\beta_0$ and $\beta_1$ with $l' = (1 \ X_0)$. Then, it is estimated by

$$l'\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

and the variance of $l'\hat{\beta}$ is given by

$$\sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right)$$

(verify!)

Thus, a $100(1 - \alpha)\%$ confidence interval for the mean response of $Y$ at a fixed value $X_0$ is given by

$$(\hat{\beta}_0 + \hat{\beta}_1 X_0) \pm t_{1-\frac{\alpha}{2},n-2} \sqrt{\text{MSE} \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right)}$$

EXTRAPOLATION: It is sometimes desired to estimate $E(Y) \text{ at } x = X_0 = \beta_0 + \beta_1 X_0$ based on the fit of the straight line for values of $X_0$ outside the range of $X$ values used in the data. This is called extrapolation, and can be quite dangerous. In order for our inferences to be valid, we must believe that the straight line relationship holds for $X$ values outside the range where we have observed data. In some situations, this may be reasonable; in others, we may have no theoretical basis for making such a claim without data to support it.
1.1 INTERVAL ESTIMATES and HYPOTHESES TESTS

PREDICTION INTERVALS FOR a FUTURE Y: Sometimes we are interested in the actual value of \( Y \) we might observe when \( X = X_0 \). On the surface, this may sound like the same problem as above, but they are, indeed, very different. For example, consider a stockbroker who would like to learn about the value of a stock based on previous data. In this setting, the stockbroker would like to predict or forecast the actual value of the stock, say, \( Y_0 \) that might be observed when \( X = X_0 \). On the other hand, the stockbroker probably does not care about what might happen “on the average” at some future time; that is, he/she is probably not concerned with estimating \( E(Y) \) at \( X = X_0 \). In the context of our model, we are interested in the future observation \( Y_0 = \beta_0 + \beta_1X_0 + \epsilon_0 \) where \( \epsilon_0 \) is the error associated with \( Y_0 \) that makes it differ from the mean \( \beta_0 + \beta_1X_0 \). Note also that we wish to predict a random quantity, not a fixed parameter.

The error in prediction \( \hat{Y}_0 - Y_0 \) is normally distributed; more precisely,

\[
\hat{Y}_0 - Y_0 \sim N\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right]\right)
\]

Then, a 100\((1 - \alpha)\)% prediction interval for \( Y_0 \) is

\[
\hat{Y}_0 \pm t_{1-\frac{\alpha}{2},n-2} \sqrt{\text{MSE} \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right)}
\]

REMARK: In the simple linear regression case, the \( F^* \) test using \( F^* = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\text{MSE}} \) is the same as the \( t^* \) test for \( H_0 : \beta_1 = 0 \) vs. \( H_a : \beta_1 \neq 0 \). In general, the \( t^* \) statistic is the square root of \( F^* \) for one degree of freedom (two-sided) hypotheses tests. So if a two-sided test is desired, either \( t^* \) or \( F^* \) give identical results (Numerator of \( F^* \) has one degree of freedom).

The rejection rule is: Reject \( H_0 \) if \( F^* > F_{1-\alpha, df_1, df_2} \); otherwise do not reject \( H_0 \).

\( (df_1 = p - 1, \ df_2 = n - p) \), \( p = \) number of regression parameters.

Example. Effects of ozone pollution on soybean yield (- continued)

a) Give the 95% confidence interval for \( \beta_0 \).

b) Test \( H_0 : \beta_1 = 0 \) vs. \( H_0 : \beta_1 \neq 0 \).

c) Find the 95% prediction interval for the soybean yield at \( X_0 = .02 \).
1.2 Analysis of Variance approach

Sums of Squares

In a regression framework we have a set of values of $Y$, whose variability we want to explain. We are interested in how much the variables in $X$ help to explain $Y$.

The total variation of $Y$, corrected only for the mean (SST) is a measure of the variation of $Y$ around the mean $\bar{Y}$. The total variation can be split up into two component sums of squares, sum of squares due to regression (SSR) and sum of squares due to errors (SSE). SSR is the variation of the fitted points around the mean and SSE is the residual variation in $Y$ that is not explained by the regression curve.

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y})^2$$

This kind of partitioning is called the analysis of variance because the total sum of squares (SST) is the sum of squares that is used to compute the variance of $Y$ if no model is fit. It is this variance that is being “analyzed” or partitioned in the name of analysis of variance (ANOVA).

Typically this breakdown of sums of squares is put into an analysis of variance table: Let $p$ denote the number of parameters to be estimated in a regression model.

<table>
<thead>
<tr>
<th>Source</th>
<th>degrees of freedom</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$p - 1$</td>
<td>$\sum_{i=1}^{n}(\hat{Y}<em>i - \bar{Y})^2 = \sum</em>{i=1}^{n}\hat{Y}_i^2 - n(\bar{Y})^2$</td>
<td>$\frac{SSR}{p-1} = MSR$</td>
<td>MSR</td>
</tr>
<tr>
<td>Error</td>
<td>$n - p$</td>
<td>$\sum_{i=1}^{n}(Y_i - \hat{Y})^2$</td>
<td>$\frac{SSE}{n-p} = MSE$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\hat{Y}_i^2 - n(\bar{Y})^2$</td>
<td>$\frac{MSR}{MSE}$</td>
<td></td>
</tr>
</tbody>
</table>

The Coefficient of Determination: $R^2$

This immediately implies that a good regression is one with a large $R^2$, where

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$
The $R^2$ value (called the coefficient of determination) measures the proportion of variability in $Y$ accounted for by the regression. Values closer to 1 indicate a strong regression, while values closer to 0 indicate a weaker one. When it is small, it may be that there is a lot of random inherent variation in the data.

Sometimes a slightly adjusted value of $R^2$, which is designed to offset an upward bias in it, is reported; the **adjusted** $R^2$ has the form

$$R^2_a = R^2 - \frac{1}{n-2}(1 - R^2).$$

**The Coefficient of Correlation**

In observational data situations, often, we do not choose fixed values of $X$; rather, we merely observe the pair of random variables $(X, Y)$. Then, there may exist a bivariate distribution of the random vector $(X, Y)$ with the (population) correlation coefficient $\rho$ indicating their linear association:

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

The sample correlation coefficient defined as

$$r \text{ (or } r_{XY}) = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

is $\pm \sqrt{R^2}$ with the sign depending on the sign of the estimated slope coefficient.

**Warnings**

1. A large $R^2$ value does not imply that the estimated regression line fits the data well. Another function might better describe the trend in the data.

2. One data point (or a few data points) can greatly affect the $R^2$ value.

3. Association (or correlation) does not imply causation.

4. Correlations that are based on rates or averages tend to overstate the strength of an association.

5. A “statistically significant” $R^2$ value does not imply that the slope is meaningfully different from 0. A large r-squared value does not necessarily imply useful predictions.
1.3 Matrix approach to Simple Linear Models

Reminder about Matrices

i) \( AB \neq BA \) ii) \((AB)' = B'A'\) (transpose)

iii) Rank and dependence

In general, \( k \) columns (of a matrix) \( U_1, U_2, \ldots, U_k \) are linearly dependent if there exist \( \lambda_1, \lambda_2, \ldots, \lambda_k \) such that 1) \( \lambda_1 U_1 + \lambda_2 U_2 + \ldots + \lambda_k U_k = \Phi \) where \( \Phi \) is a column of 0’s and 2) at least one of the \( \lambda \)'s is not 0. Thus, \( k \) columns are linearly independent if the only linear combination of them which will produce the zero vector is the linear combination with all \( \lambda \)'s 0.

The rank of a matrix is the maximum number of linearly independent columns which can be selected from the columns of the matrix. Note that if the rank of a matrix is 1, then there is one column such that all other columns are direct multiples.

For any matrix \( A \), the rank of \( A \) is the same as the rank of \( A'A \). The row rank of any matrix is always equal to the column rank.

iv) Inverse of a matrix \((A^{-1})\)

The inverse of an \( n \times n \) matrix \( A \) is an \( n \times n \) matrix \( B \) such that \( AB = I_n \) where \( I_n \) is an \( n \times n \) identity matrix. Such a matrix \( B \) will exist only if \( A \) is of rank \( n \). In this case it is also true that \( BA = I_n \).

RANDOM VECTORS

Let \( \mathbf{Y} = (Y_1, Y_2, ..., Y_n)' \) be a random vector with the probability density function (pdf) denoted by \( f(y) \) (describing how \( Y_1, Y_2, ..., Y_n \) are jointly distributed).

Then, \( E(\mathbf{Y}) = (E(Y_1), E(Y_2), ..., E(Y_n))' \) and the variance-covariance matrix of \( \mathbf{Y} \) is

\[
\begin{bmatrix}
\text{Var}(Y_1) & \text{Cov}(Y_1, Y_2) & \cdots & \text{Cov}(Y_1, Y_n) \\
\text{Cov}(Y_2, Y_1) & \text{Var}(Y_2) & \cdots & \text{Cov}(Y_2, Y_n) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(Y_n, Y_1) & \text{Cov}(Y_n, Y_2) & \cdots & \text{Var}(Y_n)
\end{bmatrix}
\]
1.3 Matrix approach to Simple Linear Models

defined by $E[(\mathbf{Y} - E(\mathbf{Y}))(\mathbf{Y} - E(\mathbf{Y}))'] = \text{Var}(\mathbf{Y})$. Since $\text{Cov}(Y_i, Y_j) = \text{Cov}(Y_j, Y_i)$, it follows that the variance-covariance matrix is symmetric.

The covariance of two random vectors, $\mathbf{Y}_{n \times 1}$ and $\mathbf{Z}_{m \times 1}$ is given by

$$
\text{Cov}(\mathbf{Y}, \mathbf{Z}) = \begin{bmatrix}
\text{Cov}(Y_1, Z_1) & \text{Cov}(Y_1, Z_2) & \cdots & \text{Cov}(Y_1, Z_m) \\
\text{Cov}(Y_2, Z_1) & \text{Cov}(Y_1, Z_2) & \cdots & \text{Cov}(Y_2, Z_m) \\
\vdots & \ddots & \ddots & \vdots \\
\text{Cov}(Y_n, Z_1) & \text{Cov}(Y_n, Z_2) & \cdots & \text{Cov}(Y_n, Z_m)
\end{bmatrix}_{n \times m}.
$$

**BASIC RESULTS**

Suppose that $A, B$ are $n \times m$ matrices of constants and that $\mathbf{c}$ is a vector of constants. Let $V$ be the variance-covariance matrix of $\mathbf{Y}$.

i) $E(A\mathbf{Y}) = A E(\mathbf{Y})$

ii) $\text{Var}(A\mathbf{Y} + \mathbf{c}) = A \text{Var}(\mathbf{Y}) A' = AVA'$.

iii) $\text{Cov}(A\mathbf{Y}, B\mathbf{Z}) = A \text{Cov}((\mathbf{Y}, \mathbf{Z}) B'$

iv) $\mathbf{Y}'A\mathbf{Y} = \sum \sum a_i Y_i Y_j$ is a quadratic form and

$$
E(\mathbf{Y}'A\mathbf{Y}) = \mu' A \mu + \text{tr}(AV),
$$

where $E(\mathbf{Y}) = \mu$ and $\text{tr}(W) = \sum_{i=1}^{n} w_{ii}$ is the trace of $W_{n \times n}$.

Recall that expectation of a linear combination of random variables, $U_k$, $k = 1, \cdots, m$ : $E(\sum_{k=1}^{m} a_k U_k) = \sum_{k=1}^{m} a_k E(U_k)$

The variance of a linear combination of random variables is

$$
\text{Var}(\sum_{k=1}^{m} a_k U_k) = \sum_{k=1}^{m} \sum_{l=1}^{m} a_k a_l \text{Cov}(U_k, U_l)
$$

Definition: The random vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$ is said to have a multivariate normal distribution with mean $\mu$ and variance-covariance matrix $V$ if its (joint) probability density function is
given by
\[ f_Y(y) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(y - \mu)'V^{-1}(y - \mu)\right\}, \]
for all \( y \in \mathbb{R}^n \). Shorthand notation for this statement is \( Y \sim \text{MVN} (\mu, V) \).

- If \( Y \sim \text{MVN}(\mu, V) \), then marginally, each \( Y_i \sim \text{Normal}(\mu_i, \text{Var}(Y_i)) \).

- If \( A \) is a matrix of constants and \( c \) is a vector of constants,

\[
AY + c \sim \text{MVN}(A\mu + c, AVA').
\]

**Regression Model in Matrix Notation**

For more complicated models it is useful to write the model in matrix notation. When we expand out the simple linear model,
\[
Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n
\]

\[
Y_1 = \beta_0 + \beta_1 X_1 + \epsilon_1
\]
\[
Y_2 = \beta_0 + \beta_1 X_2 + \epsilon_2
\]
\[
\vdots
\]
\[
Y_n = \beta_0 + \beta_1 X_n + \epsilon_n
\]

This can be expressed as

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix} = \begin{bmatrix} 1 & X_1 \\
1 & X_2 \\
\vdots & \vdots \\
1 & X_n
\end{bmatrix} \begin{bmatrix} \beta_0 \\
\beta_1
\end{bmatrix} + \begin{bmatrix} \epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{bmatrix}
\]

or

\[
Y = X\beta + \epsilon
\]

The normal equations can be written in matrix notation as well. They are

\[
X'X\beta = X'Y
\]
or

\[
\begin{bmatrix}
  n & \sum X_i \\
  \sum X_i & \sum X_i^2 \\
\end{bmatrix}
\begin{bmatrix}
  \beta_0 \\
  \beta_1 \\
\end{bmatrix} =
\begin{bmatrix}
  \sum Y_i \\
  \sum X_i Y_i \\
\end{bmatrix}
\]

Thus, if the design matrix $X$ has full column rank, then the normal equations can be solved by pre-multiplying both sides by $(X'X)^{-1}$.

\[
\hat{\beta} = (X'X)^{-1}X'Y
\]

You should verify for yourselves that in the simple linear regression case

\[
\hat{\beta} = (X'X)^{-1}X'Y = 
\begin{bmatrix}
  Y - \hat{\beta}_1 X \\
  \frac{s_{xy}}{s_{xx}} \\
\end{bmatrix}
\]

Recall that the vector $e$ is what makes the above model a peculiar statistical model. One of the assumption of the ordinary regression model is that $\epsilon_i \sim$ uncorrelated $(0, \sigma^2)$.

We also see that $Y$ is a random vector because $X_i$ is a known constant.

The least squares estimates of the parameters are random vectors since they are linear combinations of $Y$.

So we can find the mean and covariance matrix of $\hat{\beta}$ as

\[
E(\hat{\beta}) = E\{(X'X)^{-1}X'Y\} = (X'X)^{-1}X'E\{Y\} \\
= (X'X)^{-1}X'X\beta \\
= \beta
\]

Thus, the least squares estimates for $\beta$ are unbiased.

The variance-covariance matrix of $\hat{\beta}$ is

\[
\text{Var}(\hat{\beta}) = \text{Var}\{(X'X)^{-1}X'Y\} \\
= (X'X)^{-1}X'\text{Var}\{Y\}X(X'X)^{-1} \\
= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1}, \text{ when Cov}(\epsilon_i, \epsilon_j) = 0 \text{ and Var}(\epsilon_i) = \sigma^2 \text{ for all } i, j \\
= (X'X)^{-1}\sigma^2
\]
The predicted values or estimated values of $Y$ are called $\hat{Y}$:

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY \text{ where } H = X(X'X)^{-1}X' \text{ called the hat matrix.}$$

- $H$: symmetric, idempotent ($HH = H$)

The residuals are deviations of the observed points from the fitted regression line. They are important for assessing the fit of the model and for computing an estimate of the variance.

$$e = Y - \hat{Y} = Y - X\hat{\beta} = Y - HY = (I - H)Y$$

Properties of the residuals that are used in checking the model assumptions:

1. The sum of squares of the residuals is minimized by the criterion for least squares.
2. $X'e = 0$ (The residuals are orthogonal to the columns of X matrix). This means that
   a) For a model with an intercept (a column of ones) the sum of residuals is zero because
   $$1'e = \sum_{i=1}^{n} e_i = 0$$
   b) The residuals are orthogonal to every independent variable in the model
   $$\sum_{i=1}^{n} X_{ik}e_i = 0$$
   This means that if you tried to regress the residuals on $X$ you would get a slope of zero and the scatter plot of $e$ vs $X$ would be centered about zero.
3. The residuals are also orthogonal to $\hat{Y}$ since $\hat{Y}'e = \hat{\beta}'X'e = 0$. Notice that the residuals are not orthogonal to the observed values of $Y$.

Note that the ANOVA sums of squares are all quadratic forms.

i) $SST = Y'Y - \frac{1}{n}Y'11'Y = Y'[(I - \frac{1}{n}11')Y$

ii) $SSR = \hat{\beta}'X'Y - \frac{1}{n}Y'11'Y = Y'[(H - \frac{1}{n}11')]Y$

iii) $SSE = Y'Y - \hat{\beta}'X'Y = Y'[I - H]Y$
We can also test a hypothesis about a linear combination of parameters, $l^T\beta$.

$$l^T\beta \sim \text{Normal}(l^T\beta, l^T(X'X)^{-1}l\sigma^2).$$

A two sided test: $H_0: l^T\beta = M$ vs. $H_a: l^T\beta \neq M$ can be done as follows.

Reject $H_0$ if $|t^*| = \frac{|l^T\hat{\beta} - M|}{s(l^T\hat{\beta})} > t_{1-\frac{\alpha}{2}, n-p}$, $s(l^T\hat{\beta}) = \sqrt{l^T(X'X)^{-1}l\text{MSE}}$.

(Recap) Under the normality assumption, the $100(1-\alpha)$% confidence interval for $\beta_k$ is

$$\hat{\beta}_k \pm t_{1-\frac{\alpha}{2}} s(\hat{\beta}_k), \quad k = 1, 2 \quad s(\hat{\beta}_k) = \sqrt{(X'X)^{-1}kk\text{MSE}}.$$

Simple example. DATA (X,Y): (4,5), (7,7), (5,6), (4,6) Assume that $\sigma^2 = 0.25$ known.

$$\begin{bmatrix} 5 \\ 7 \\ 6 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 1 & 7 \\ 1 & 5 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}$$

We find the estimated parameters using the formula $\hat{\beta} = (X'X)^{-1}X'Y = (3.5, 0.5)'$ (Verify!)

The variance-covariance matrix of parameter estimates is

$$\begin{bmatrix} -4.4167 & -0.8333 \\ -0.8333 & 0.1667 \end{bmatrix} (0.25) = \begin{bmatrix} 1.1042 & -0.2083 \\ -0.2083 & 0.0417 \end{bmatrix}$$

(1) Test that the slope is 0: $t = 0.5/\sqrt{0.0417} = 0.5/0.2041 = 2.45$ (2 degrees of freedom)

(2) Test that the intercept is 0: (2 degrees of freedom)

(3) Estimate the mean value of Y at X = 6 and give a 95% confidence interval.

We are estimating: $\beta_0 + 6\beta_1 = [1, 6] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ and $l^T\hat{\beta} = \begin{bmatrix} 1.1042 & -0.2083 \\ -0.2083 & 0.0417 \end{bmatrix} \begin{bmatrix} 1 \\ 6 \end{bmatrix} = 0.1042$.

The 95% confidence interval for our little example is

Note that the 97.5% quantile of the $t$ distribution with 2 d.f. = 4.30.
1.4 DIAGNOSTICS and MODEL EVALUATION

When a simple linear model is selected for an application, one cannot be certain in advance that the model is appropriate to the dataset in question. Note that all the tests, intervals, predictions, etc., are based on believing that the assumptions of the regression hold. To assess whether the assumptions underlying the model seem reasonable, we study the residuals with graphical analyses. That is, we analyze the residuals to see if they support the assumptions of linearity, independence, normality and equal variances.

When conducting a residual analysis, we make the following plots;

(1) plot of the residuals versus the fitted values.
   - The residuals should be scattered randomly around the 0 line.
   - The residuals should roughly form a horizontal band around the 0 line, suggesting that the variances of the error terms are equal.

(2) Residuals vs. predictor plot
   - The interpretation is identical to that for a “residuals vs. fits plot.”
   - If the predictor on the $X$ axis is a new and different predictor, the residuals vs. predictor plot can help to determine whether the predictor should be added to the model (and hence a multiple regression model used instead).

(3) Normal probability plot of the residuals
   - constructed by plotting the $n$ ordered residuals against the $n$ ordered quantiles from the standard normal distribution.
   - To assesses the apparent normality of the residuals and should look like a straight line (roughly).
   - Isolated points represent unusual observations, while a curved line indicates that the errors are probably not normally distributed, and tests and intervals might not be trustworthy.
(4) Residuals versus time

- To check if your data has a time structure to it. If not, there should be no apparent pattern.

OUTLIERS: An outlier is an observation that is unusually small or large; i.e. outliers are the data points that do not fit well with the pattern of the rest of the data. In straight-line regression, an outlier might be an observation that falls far off the apparent approximate straight line trajectory followed by the remaining observations. Practitioners often “toss out” such anomalous points, which may or may not be a good idea. If it is clear that an outlier is the result of a mishap or a gross recording error, then this may be acceptable. On the other hand, if no such basis may be identified, the outlier may, in fact, be a genuine response; in this case, it contains information about the process under study, and may be reflecting a legitimate phenomenon. Then, “throwing out” an outlier may lead to misleading conclusions, because a legitimate feature is being ignored.

STRATEGIES TO DEAL WITH OUTLIERS: What should we do if an outlier (or outliers) are identified? Unfortunately, there is no clear-cut answer! However, here are some suggestions:

a. Delete outliers and redo the analysis (new outliers may surface).
b. Sometimes the purpose of the experiment is just to identify the outliers. In this case, there is no need to redo the analysis.

c. Check the experimental circumstances surrounding the data collection for the outlying cases.

d. Report the analysis both with and without the analysis and let the reader decide.

LEVERAGEs: To identify outliers, we should consider first looking at the residual plot of $e_i$ versus $\hat{Y}_i$. Recall the property of residuals:

$$e_i = Y_i - \hat{Y}_i \sim N(0, \sigma^2(1 - h_{ii}))$$

where

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}},$$

which is called the leverage of $i$th case.

So the residuals (unlike the errors) do not have constant variance and are slightly correlated! Observations where $X_i$ is far away from $\bar{X}$ will have large values of $h_{ii}$. However, not all observations with large leverages are necessarily outliers.

Leverage is an index of the importance of an observation to a regression analysis: large deviations from mean are influential.

STUDENTIZED RESIDUALS: To account for the different variances among residuals, we consider “studentizing” the residuals (i.e., dividing by an estimate of their standard deviation).

The (internally) studentized residuals are defined as:

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

have $E(r_i) = 0$, $\text{Var}(r_i) = 1$ so that the studentized residuals have a constant variance regardless of the location of the $X$’s. Values of $|r_i|$ larger than 3 or so should cause concern.

TRANSFORMATIONS: Transforming the data is a way of handling violations of the usual assumptions. In the regression context, this may be done in a number of ways. One way is to invoke an appropriate transformation, and then postulate a regression model on the
transformed scale. Sometimes it may be that, although the data do exhibit constant variance on the original scale, they may be analyzed better on some transformed scale. It is also important to remember that if a transformation is used, the resulting inferences apply to this transformed scale (and no longer to the original scale). One should remember that transforming the data may fix one problem, but it may create other violations of the model.

Another approach is to proceed with a regression method known as weighted-least squares. In a weighted regression analysis, different responses are given different weights depending on their variances (to be discussed later).

In general, transforming your data almost always involves lots of trial and error.

**BOX-COX TRANSFORMATION:** The power transformation

\[
g(Y) = \begin{cases} 
\log(Y), & \lambda = 0 \\
y^\lambda, & \lambda > 0 
\end{cases}
\]

was suggested by Box and Cox (1964). The log and square root transformations are special cases with \( \lambda = 0 \) and \( \lambda = 1/2 \), respectively. Approximate \( 100(1 - \alpha) \) percent confidence intervals for \( \lambda \) are available.

**OTHER TRANSFORMATIONS** (mainly for simple linear regression models)

Some guidelines: You may want to try transformations of the \( Y \)-variable when there is evidence of non-normality and/or non-constant variance problems in one or more residual plots. Try transformations of the \( X \)-variable(s) (e.g., \( X^{-1}, X^2, \ln(X) \)) when there are strong nonlinear trends in one or more residual plots. If either of the above trials does not work, consider transforming both the response \( Y \) and the predictor \( X \) (e.g., \( \ln(Y), \ln(X) \)).

If the error variances are unequal, try ‘stabilizing the variance’ by transforming \( Y \):

- When the response is a Poisson count, take the \( \sqrt{Y} \) transformation.

- For a binomial proportion, use the ‘arcsine transformation’ \( \sin^{-1}(\sqrt{y}) \) or more commonly do the logistic regression.

- When the response isn’t anything special but showing unequal variances, use \( \ln(Y) \), or \( 1/Y \).
2 Multiple Regression

While the straight-line model serves as an adequate description for many situations, more often than not, researchers who are engaged in model building consider more than just one predictor variable $X$. In fact, it is often the case that the researcher has a set of $p - 1$ candidate predictor variables, say, $X_1, X_2, ..., X_{p-1}$, and desires to model $Y$ as a function of one or more of these $p - 1$ variables. To accommodate this situation, we must extend our linear regression model to handle more than one predictor variable.

MULTIPLE REGRESSION SETTING: Consider an experiment in which $n$ observations are collected on the response variable $Y$ and $p - 1$ predictor variables $X_1, X_2, ..., X_{p-1}$.

<table>
<thead>
<tr>
<th>Individual</th>
<th>$Y$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>...</th>
<th>$X_{p-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_1$</td>
<td>$X_{11}$</td>
<td>$X_{12}$</td>
<td>...</td>
<td>$X_{1p-1}$</td>
</tr>
<tr>
<td>2</td>
<td>$Y_2$</td>
<td>$X_{21}$</td>
<td>$X_{22}$</td>
<td>...</td>
<td>$X_{2p-1}$</td>
</tr>
<tr>
<td>$n$</td>
<td>$Y_n$</td>
<td>$X_{n1}$</td>
<td>$X_{n2}$</td>
<td>...</td>
<td>$X_{np-1}$</td>
</tr>
</tbody>
</table>

To describe $Y$ as a function of the $p - 1$ independent variables $X_1, X_2, ..., X_{p-1}$, we posit the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{ip-1} + \epsilon_i$$

for $i = 1, 2, ..., n$, where $n > p$ and $\epsilon_i \sim N(0, \sigma^2)$. The values $\beta_0, \beta_1, ..., \beta_{p-1}$ are regression coefficients as before and we assume that $X_1, X_2, ..., X_{p-1}$ are all fixed. The random errors $\epsilon_i$’s are still assumed to be independent and have a normal distribution with mean zero and a common variance $\sigma^2$. Then,

$$Y = X\beta + \epsilon,$$

where $\epsilon \sim \text{MVN } (0, \sigma^2I)$.

In the multiple regression setting, because of the potentially large number of predictors, it is more efficient to use matrices to define the regression model and the subsequent analyses.
Note that the matrix $X$ is called the design matrix, since it contains all of the predictor variable information.

The least squares method is used to estimate the regression parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ for which we can easily find closed-form solutions in terms of matrices and vectors.

Provided that $X'X$ is full rank, the least-squares estimator of $\beta$ is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

NOTE: For the least-squares estimator $\beta$ to be unique, we need $X$ to be of full column rank; i.e., $r(X) = p$. That is, there are no linear dependencies among the columns of $X$. If $r(X) < p$, then $r(X) = r(X'X)$ and $X'X$ does not have a unique inverse. In this case, the normal equations cannot be solved uniquely.

To avoid the more technical details of working with non-full rank matrices we will assume that $X$ is of full rank, unless otherwise stated.

Example. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. Overall taste scores were obtained by combining the scores from several tasters. Data were collected on the following variables:

- $Y$ = taste test score (TASTE)
- $X_1$ = concentration of acetic acid (ACETIC)
- $X_2$ = concentration of hydrogen sulfide (H2S)
- $X_3$ = concentration of lactic acid (LACTIC)

Variables ACETIC and H2S are both on the (natural) log scale. The variable LACTIC has not been transformed. Table below contains concentrations of the various chemicals in $n = 30$ specimens of mature cheddar cheese and the observed taste score.

Suppose that the researchers postulate that each of the three chemical composition covariates $X_1, X_2, \text{ and } X_3$ are important in describing the taste. In this case, they might
initially consider the following regression model

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \]

for \( i = 1, 2, \ldots, 30 \). Are there other predictor variables that influence taste not considered here? Alternatively, what if not all of \( X_1, X_2, \) and \( X_3 \) are needed in the model? For example, if the acetic acid concentration (\( X_1 \)) is not helpful in describing taste, then we might consider a smaller model which excludes it; i.e.,

\[ Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \]

for \( i = 1, 2, \ldots, 30 \). The goal of any regression modeling problem should be to identify each of the important predictors, and then find the simplest model that does the best job.

The full model, \( Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \) in matrix notation is \( Y = X\beta + \epsilon \),
where

\[
Y = \begin{pmatrix}
12.3 \\
20.9 \\
39.0 \\
5.5 \\
\vdots \\
\end{pmatrix},
X = \begin{pmatrix}
1 & 4.543 & 3.135 & 0.86 \\
1 & 5.159 & 5.043 & 1.53 \\
1 & 5.366 & 5.438 & 1.57 \\
1 & 6.176 & 4.787 & 1.25 \\
\vdots & \vdots & \vdots & \vdots \\
\end{pmatrix}
\text{and } \beta = \begin{pmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3 \\
\end{pmatrix}.
\]

We compute

\[
X'X = \begin{pmatrix}
30 & 164.941 & 178.254 & 43.260 \\
164.941 & 916.302 & 1001.806 & 240.879 \\
178.254 & 1001.806 & 1190.343 & 269.113 \\
43.260 & 240.879 & 269.113 & 65.052 \\
\end{pmatrix}
\quad \text{and } X'Y = \begin{pmatrix}
736.000 \\
4194.442 \\
5130.932 \\
1162.065 \\
\end{pmatrix}.
\]

Thus, the least squares estimate of \( \beta \) for these data is given by

\[
\hat{\beta} = (X'X)^{-1}X'Y = (-28.877 0.328 3.912 19.670)'.
\]

The least-squares regression equation becomes

\[
\hat{Y}_i = -28.877 + 0.328X_{i1} + 3.912X_{i2} + 19.670X_{i3}
\]

or, in terms of the variable names,

\[
\text{TASTE}_i = -28.877 + 0.328 \text{ACETIC}_i + 3.912 \text{H2S}_i + 19.670 \text{LACTIC}_i.
\]

Using the first model, the ANOVA table for the cheese data is shown below. The F statistic is used to test \( H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \) vs. \( H_1 : \beta_1, \beta_2, \beta_3 \) not all zero. Since the \( P \)-value for the test is so small, we would conclude that at least one of \( X_1, X_2, \) or \( X_3 \) is important in describing taste. The coefficient of determination is \( R^2 \approx 0.652 \). Thus, about 65 percent of the variability in the taste data is explained by \( X_1, X_2, \) and \( X_3 \) assuming the model is true.
2.1 Sampling Distributions and Inference for the parameters

Sampling Distribution of $\hat{\beta}$

Since $\hat{\beta}$ is a linear combination of $Y$, $\hat{\beta}$ has a multivariate Normal distribution as $\epsilon$ and $Y$.

$$\hat{\beta} \sim \text{MVN}(\beta, \sigma^2(X'X)^{-1}).$$

Now $(X'X)^{-1}$ is a $p \times p$ matrix. So is $\text{Var}(\hat{\beta})$, the variance-covariance matrix of $\hat{\beta}$.

**IMPLICATIONS:**

(a) $E(\hat{\beta}_k) = \beta_k$, for $k = 0, 1, ..., p - 1$; unbiased.

(b) $\text{Var}(\hat{\beta}_k) = (X'X)^{-1}_{kk}\sigma^2$, for $k = 0, 1, ..., p - 1$.

The value $(X'X)^{-1}_{kk}$ represents the $k$th diagonal element of the $(X'X)^{-1}$ matrix.

(c) $\text{Cov}(\hat{\beta}_k, \hat{\beta}_l) = (X'X)^{-1}_{kl}\sigma^2$, for $k \neq l$.

The value $(X'X)^{-1}_{kl}$ is the entry in the $k$th row and $l$th column of the $(X'X)^{-1}$ matrix.

(d) Marginally, $\hat{\beta}_k \sim N(\beta_k, (X'X)^{-1}_{kk}\sigma^2)$, for $k = 0, 1, ..., p - 1$.

(e) $E(\text{MSE}) = \sigma^2$ where

$$\frac{\text{SSE}}{\sigma^2} = \frac{\text{MSE}(n - p)}{\sigma^2} \sim \chi^2_{n-p}.$$

**CONFIDENCE INTERVALS:**

The 100(1 $- \alpha$)% confidence interval for $\beta_k$ is $\hat{\beta}_k \pm t_{1-\alpha/2, n-p} s(\hat{\beta}_k)$, $k = 0, \ldots, p - 1$ where $s(\hat{\beta}_k)$ is the standard error of $\hat{\beta}_k$, i.e., $\sqrt{(X'X)^{-1}_{kk}} \text{MSE}$.

Since

$$\frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{\sigma^2} \sim \chi^2_p$$

is independent of $\text{MSE}$,

$$\frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{p\text{MSE}} \sim \frac{\chi^2_p/p}{\chi^2_{n-p}/(n-p)} \equiv F_{p,n-p}$$

Thus, a 100(1 $- \alpha$)% (simultaneous) confidence region for $\beta$ can be formed as

$$(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \leq p \text{ MSE } F_{1-\alpha/p(n-p)}.$$ 

These regions are ellipsoidally shaped and cannot be easily visualized due to high-dimensionality.
CONFIDENCE INTERVAL FOR THE MEAN RESPONSE: Given a new set of predictors $X_0$, what is the estimated mean response? We call it $\hat{Y}_0$, which is $X_0'\hat{\beta}$. Var($X_0'\hat{\beta}$) = $X_0'(X'X)^{-1}X_0\sigma^2$. So a $100(1 - \alpha)$% confidence interval for the average of the responses with given $X_0$ is

$$\hat{Y}_0 \pm t_{1 - \frac{\alpha}{2}, n-p} \sqrt{\text{MSE} \cdot X_0'(X'X)^{-1}X_0}.$$ 

PREDICTION INTERVAL OF A SINGLE FUTURE RESPONSE FOR $X_0$: Recall that a future observation is predicted to be $\hat{Y}_0 = X_0'\hat{\beta}$ (where we don’t know what the future response will turn out to be.) Then, a $100(1 - \alpha)$% prediction interval for a single future response is

$$\hat{Y}_0 \pm t_{1 - \frac{\alpha}{2}, n-p} \sqrt{\text{MSE} \cdot [1 + X_0'(X'X)^{-1}X_0]}.$$ 

HYPOTHESIS TESTS:

For testing $H_0 : \beta_k = \beta_{k,0}$ vs. $H_1 : \beta_k \neq \beta_{k,0}$, we use

$$|t^*| = \frac{\hat{\beta}_k - \beta_{k,0}}{\sqrt{(X'X)^{-1}_{kk} \text{MSE}}} > t_{1 - \frac{\alpha}{2}, n-p}.$$ 

to reject $H_0$; Otherwise, do not reject $H_0$.

Example (cheese data continued). Considering the full model,

a) Find the estimate of the variance-covariance matrix of $\hat{\beta}$.

We first compute $(X'X)^{-1} = \begin{pmatrix}
3.795 & -0.760 & 0.087 & -0.071 \\
-0.760 & 0.194 & -0.020 & -0.128 \\
0.087 & -0.020 & 0.015 & -0.046 \\
-0.071 & -0.128 & -0.046 & 0.726
\end{pmatrix}$

b) To assess the importance of the hydrogen sulfide concentration and its influence on taste of cheese, we can test $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$.

c) $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_1 : \beta_1, \beta_2, \beta_3$ not all zero.

Note that if $H_0$ is rejected, we do not know which one or how many of the $\beta_k$’s are nonzero; we only know that at least one is.
SUMS OF SQUARES AS QUADRATIC FORMS

Recall that in page 8 the corrected sums of squares were partitioned and expressed in terms of random vectors. To enhance the understanding of these sums of squares, we often express them in terms of quadratic forms.

\[
\begin{align*}
\text{SST} &= Y'(I - \frac{1}{n}11')Y = Y'(I - \frac{1}{n}J)Y \\
\text{SSR} &= Y'(H - \frac{1}{n}J)Y \\
\text{SSE} &= Y'(I - H)Y
\end{align*}
\]

Recall that any \(n \times n\) real symmetric matrix \(A\) determines a quadratic form such that

\[
Y'AY = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}Y_iY_j.
\]

Def. The trace of a \(n \times n\) square matrix \(A\) is \(tr(A) = \sum_{i=1}^{n} A_{ii}\)

Lemma: \(E(Y'AY) = tr(A\Sigma) + \mu'A\mu\), where \(E(Y) = \mu\) and \(Var(Y) = \Sigma\).

EXPECTED MEAN SQUARES

We have seen that MSE is unbiased for \(\sigma^2\) and this holds whether or not \(H_0: \beta_0 = \ldots = \beta_{p-1} = 0\) is true. On the other hand,

\[
E(\text{SSR}) = E[Y(H - n^{-1}11')Y] \\
= tr[(H - n^{-1}11')\sigma^2 I] + (X\beta)'(H - n^{-1}11')X\beta \\
= (p - 1)\sigma^2, \text{ only when } H_0 \text{ is true.}
\]

\[
E(\text{MSR}) = E\left(\frac{\text{SSR}}{p-1}\right) = \sigma^2, \text{ only when } H_0 \text{ is true.}
\]

NOTE. When \(H_0\) is true, both MSR and MSE are estimating the same quantity, and the \(F\) statistic should be close to one. When \(H_0\) is not true, the \((X\beta)'(H - n^{-1}11')X\beta > 0\), and hence, MSR is estimating something larger than \(\sigma^2\). In this case, we would expect \(F\) to be larger than one. This gives an intuitive explanation of why \(F\) statistic is computed large when \(H_0\) is not true. Also, the name of ‘Analysis of Variance’ is coined with this kind of study because we are conducting hypothesis tests by comparing different estimators for the variance. The extra term \((X\beta)'(H - n^{-1}11')X\beta\) is called a noncentrality parameter.
2.1 Sampling Distributions and Inference for the parameters ©HYON-JUNG KIM, 2017

<table>
<thead>
<tr>
<th>Source</th>
<th>degrees of freedom</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$p - 1$</td>
<td>$\hat{\beta}'X'Y - \frac{1}{n}Y'11'Y$</td>
<td>$\frac{SSR}{p-1}$</td>
</tr>
<tr>
<td>Error</td>
<td>$n - p$</td>
<td>$Y'Y - \hat{\beta}'X'Y$</td>
<td>$\frac{SSE}{n-p}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$Y'Y - \frac{1}{n}Y'11'Y$</td>
<td></td>
</tr>
</tbody>
</table>

Example (cheese data continued). Considering the full model,

d) Find a 95 percent confidence interval for the mean taste rating when concentration of acetic acid, hydrogen sulfide, lactic acid are 5,6,1, respectively.

e) Find a 95 percent prediction interval for a particular taste rating score when concentration of acetic acid, hydrogen sulfide, lactic acid are 5,6,1, respectively.

Coefficient of Determination, R-squared, and Adjusted R-squared

As in simple linear regression, $R^2 = \frac{SSR}{SSE}$, represents the proportion of variation in $Y$ (about its mean) ‘explained’ by the multiple linear regression model with predictors. However, unlike in simple linear regression $R^2$ always increases (or stays the same) as more predictors are added to a multiple linear regression model, even if the predictors added are unrelated to the response variable. Thus, by itself, $R^2$ cannot be used to help us identify which predictors should be included in a model. An alternative measure, adjusted $R^2$, does not necessarily increase as more predictors are added, and can be used to help us identify which predictors should be excluded in a model.

Interpretation of the Model Parameters

Each $\beta_i$ coefficient represents the change in the mean response, $E(Y)$, per unit increase in the associated predictor variable when all the other predictors are held constant.

In the case of two predictors, the estimated regression equation yields a plane (as opposed to a line in the simple linear regression setting). For more than two predictors, the estimated regression equation yields a hyperplane.
2.2 General linear tests - REDUCED vs. FULL MODEL

Often we want to test a hypothesis about several parameters simultaneously (but not all of them as in the case of testing regression relation). For example, in the model

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \beta_4 X_{i2}^2 + \epsilon_i \]

we might want to test the hypothesis that all second order terms are zero; i.e. \( \beta_3 = \beta_4 = 0 \).

It is not a good idea to do \( t^* \) test for each parameter because of the correlations among the regressor (independent) variables. These \( t^* \) tests for each parameter test the following hypotheses:

- \( H_0 : \beta_3 = 0 \) given that \( \beta_0, \beta_1, \beta_2 \) and \( \beta_4 \) are not zero, or
- \( H_0 : \beta_4 = 0 \) given that \( \beta_0, \beta_1, \beta_2 \) and \( \beta_3 \) are in the model.

If \( X_{i1}X_{i2} \) and \( X_{i2}^2 \) are highly correlated, then dropping \( X_{i2}^2 \) may not have a noticeable impact. Likewise, if \( X_{i2}^2 \) is kept in the model, dropping \( X_{i1}X_{i2} \) may not have a big effect. It is entirely possible to do two individual \( t^* \) tests and conclude that neither variable is required. However, it can be that the pair would significantly improve the fit of the model, if you test the two variables simultaneously.

A common way to test \( H_0 : \beta_3 = \beta_4 = 0 \) vs. \( H_a: \) not both are zero, is to fit the “reduced” model

**REDUCED:** \( Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \)

Then compare the residual sums of squares from reduced model with the residual sum of squares from the full model. The test statistic is

\[ F^* = \frac{(\text{SSE}(\text{reduced}) - \text{SSE}(\text{full}))/\text{df}_\text{red} - \text{df}_\text{full}}{\text{MSE}(\text{full})} \]

The test criterion is: reject \( H_0 \) if \( F^* > F_{1-\alpha, \text{df}_\text{red} - \text{df}_\text{full}, \text{df}_\text{full}} \).

This type of general linear test includes the F test for a regression relation and is a very flexible method of testing hypotheses. It is also possible to test any linear hypotheses where the reduced model is nested within the full model. For example, if

**FULL:** \( Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i \)
and the hypothesis to be tested is $H_0 : \beta_3 = 0$ and $\beta_1 = \beta_2$, then we can write the reduced model as

$$\text{REDUCED: } Y_i = \beta_0 + \beta_1 (X_{i1} + X_{i2}) + \epsilon_i$$

so that the design matrix has two columns: $X = \begin{bmatrix} 1 & X_{i1} + X_{i2} \\ \vdots & \vdots \\ 1 & X_{n1} + X_{n2} \end{bmatrix}$

The parameters of this model are obtained by putting restrictions on the parameters of the full model, so we say that this model is nested within the full model. The test statistic has $2$ and $n - 4$ degrees of freedom.

**GENERAL LINEAR TESTS** in another way

The general linear test statistic can be computed another way which does not require fitting the reduced model. Any linear hypothesis can be stated in the form $H_0 : C\beta = h$. For example to test the hypothesis in the above model, write

$$C = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix} \text{ and } h = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Then

$$H_0 : C\beta = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \beta_3 \\ \beta_1 - \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The (extra) sum of squares due to fitting the full model over the reduced model is

$$\text{SSE(reduced)} - \text{SSE(full)} = (C\hat{\beta} - h)'(C(X'X)^{-1}C')^{-1}(C\hat{\beta} - h)$$

The $F^*$ test statistic is then

$$F^* = \frac{(C\hat{\beta} - h)'(C(X'X)^{-1}C')^{-1}(C\hat{\beta} - h)/d f}{\text{MSE(full)}}$$

The degrees of freedom($df$) for the numerator is the rank of $C(X'X)^{-1}C'$. This way of writing the $F^*$ statistic gives a different perspective on the test. Notice that

$$\hat{\text{Var}}(C\hat{\beta} - h) = C(X'X)^{-1}C \text{ MSE}$$
Therefore, \(F^*\) is \((C\hat{\beta} - h)'\) times the inverse of its estimated covariance matrix times \((C\hat{\beta} - h)\).

This is a similar idea to a \(t^*\) statistic. We are essentially standardizing the difference \(C\hat{\beta} - h\) based on its covariance matrix. If \(df = 1\), so that \(C\hat{\beta}\) is a scalar, then \(F^* = (t^*)^2\).

In this test, we are testing \(H_0: C\beta = h\) using a variance estimate that theoretically includes only random error. The numerator and denominator of the \(F^*\) ratio must be independent quantities that both estimate \(\sigma^2\) under \(H_0\) but estimate different quantities under \(H_a\). Sometimes if another estimate of \(\sigma^2\) is available, this may be used in the denominator instead of MSE(full) with a corresponding degrees of freedom.

**EXTRA SUMS OF SQUARES**

The general linear test is very flexible and can be used to test all sorts of hypotheses. Certain hypotheses are standard and are often included in an analysis of variance table. That is, we can partition the total sums of squares into sums of squares that test different hypotheses. The difference in SSE between a full model and a reduced model is called the “extra sum of squares” due to fitting the additional parameters.

Example: For a model with three independent variables, \(X_i, i = 1, 2, 3\), an \(n \times 1\) vector each,

\[
Y = \begin{bmatrix} 1 \vert X_1 \vert X_2 \vert X_3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \epsilon.
\]

Let \(SSE(X_1, X_2, X_3)\) be the error sum of squares for full model, i.e. \(SSE(\text{FULL})\). Define

- \(SSE(X_1) = \text{error sum of squares for model } Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i\)
- \(SSE(X_2) = \text{error sum of squares for model } Y_i = \beta_0 + \beta_2 X_{i2} + \epsilon_i\)
- \(SSE(X_3) = \text{error sum of squares for model } Y_i = \beta_0 + \beta_2 X_{i3} + \epsilon_i\)
- \(SSE(X_1, X_2) = \text{error sum of squares for model } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i\)
- \(SSE(X_1, X_3) = \text{error sum of squares for model } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i3} + \epsilon_i\)
- \(SSE(X_2, X_3) = \text{error sum of squares for model } Y_i = \beta_0 + \beta_1 X_{i2} + \beta_2 X_{i3} + \epsilon_i\)
If you want to test whether $\beta_3$ is needed, given that $\beta_1$ and $\beta_2$ are already in the model, then the **extra sums of squares** are defined as

$$R(X_3|X_1, X_2) = \text{SSE}(X_1, X_2) - \text{SSE}(X_1, X_2, X_3).$$

The “R” stands for the reduction in the error sums of squares due to fitting the additional term. The $F^*$ test for $\beta_3 = 0$ is

$$F^* = \frac{R(X_3|X_1, X_2)/(df_{\text{red}} - df_{\text{full}})}{\text{MSE}(\text{full})} = \frac{(\text{SSE(}\text{reduced}) - \text{SSE(}\text{full}))/ (df_{\text{red}} - df_{\text{full}})}{\text{MSE}(\text{full})}$$

which is the same test as we had before but expressed in different notation.

The extra sums of squares can also be written as a difference of regression sums of squares. For example,

$$R(X_3|X_1, X_2) = \text{SSE}(X_1, X_2) - \text{SSE}(X_1, X_2, X_3)$$

$$= \text{SST} - \text{SSR}(X_1, X_2) - \text{SST} - \text{SSE}(X_1, X_2, X_3)$$

$$= \text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_1, X_2)$$

**Sequential Sums of Squares** or “Type I sums of squares” (in SAS)

They are extra sums of squares for adding each term given that preceding terms in the model statement are already in the model. For the model with 3 independent variables: the Type I sums of squares are (in SAS MODEL Y= X_1 X_2 X_3;) These sequential sums of squares

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>df</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_1</td>
<td>1</td>
<td>R(X_1)</td>
</tr>
<tr>
<td>X_2</td>
<td>1</td>
<td>R(X_2</td>
</tr>
<tr>
<td>X_3</td>
<td>1</td>
<td>R(X_3</td>
</tr>
<tr>
<td>error</td>
<td>n-4</td>
<td>SSE(X_1, X_2, X_3)</td>
</tr>
</tbody>
</table>

add up to SSR(X_1, X_2, X_3).

$$R(X_1) + R(X_2|X_1) + R(X_3|X_1, X_2)$$

$$= \text{SSR}(X_1) + \text{SSR}(X_1, X_2) - \text{SSR}(X_1) + \text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_1, X_2) = \text{SSR}(X_1, X_2, X_3)$$
2.2 General linear tests - REDUCED vs. FULL MODEL

The sequential sums of squares depend on the order that you write the independent variables in the model statement. For example, the model statement: MODEL Y = X_3 X_2 X_1 would give

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>df</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_3</td>
<td>1</td>
<td>R(X_3)</td>
</tr>
<tr>
<td>X_2</td>
<td>1</td>
<td>R(X_2</td>
</tr>
<tr>
<td>X_1</td>
<td>1</td>
<td>R(X_1</td>
</tr>
<tr>
<td>error</td>
<td>n-4</td>
<td>SSE(X_1, X_2, X_3)</td>
</tr>
</tbody>
</table>

Partial Sums of Squares or “Type II” of sums of squares (in SAS)

These are extra sums of squares for adding a term to the model given that all the other independent variables in the model statement are already in the model. The order of terms in the model makes no difference. For either of the following model statements

MODEL Y = X_1 X_2 X_3  
MODEL Y = X_3 X_2 X_1

the partial sums of squares are the same; But these sums of squares do not add up to the total regression sum of squares. In other words, this is not a partition of the regression sum of squares. The only case that the Type II sums of squares do add up to SSR(X_1, X_2, X_3) is when the three independent variables are orthogonal to each other. This does not happen often in the regression setting, but designed experiments are usually set up to take advantage of this.

Note that if X_1 and X_2 are uncorrelated, then SSR(X_1) = R(X_1|X_2) and SSR(X_2) = R(X_2|X_1).
2.2 General linear tests - REDUCED vs. FULL MODEL

Example: DATA

<table>
<thead>
<tr>
<th>Y</th>
<th>X₀</th>
<th>X₁</th>
<th>X₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>1</td>
<td>-2</td>
<td>1</td>
</tr>
<tr>
<td>-3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ Y'Y = 17 \text{, } \text{SST} = 17 - 9/4 = 14.75 \]

<table>
<thead>
<tr>
<th>(X'X)</th>
<th>(X'X)^{-1}</th>
<th>X'Y</th>
<th>( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regress Y on X₀ only</td>
<td>4</td>
<td>1/4</td>
<td>-3</td>
</tr>
<tr>
<td>(SSE= 14.75)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regress Y on X₀, X₁</td>
<td>[ \begin{bmatrix} 4 &amp; 2 \ 2 &amp; 10 \end{bmatrix} ]</td>
<td>[ \begin{bmatrix} 10/36 &amp; -2/36 \ -2/36 &amp; 4/36 \end{bmatrix} ]</td>
<td>[ \begin{bmatrix} -3 \ 3 \end{bmatrix} ]</td>
</tr>
<tr>
<td>(SSE= 12.50)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regress Y on X₀, X₁, X₂</td>
<td>[ \begin{bmatrix} 4 &amp; 2 &amp; 4 \ 2 &amp; 10 &amp; 1 \ 4 &amp; 1 &amp; 10 \end{bmatrix} ]</td>
<td>[ \begin{bmatrix} 0.467 &amp; -0.076 &amp; -0.179 \ -0.076 &amp; 0.113 &amp; 0.019 \ -0.179 &amp; 0.019 &amp; 0.170 \end{bmatrix} ]</td>
<td>[ \begin{bmatrix} -3 \ 3 \ 4 \end{bmatrix} ]</td>
</tr>
<tr>
<td>(SSE = 2.9481)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SSE(X₂) =6.5833

R(X₂|X₁) = 9.552 = SSR(X₁, X₂) - SSR(X₁)

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>TYPE I SSq(sequential)</th>
<th>TYPE II SSq(partial)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>R(X₁) = 2.25</td>
<td>R (X₁</td>
</tr>
<tr>
<td>X2</td>
<td>R(X₂</td>
<td>X₁) = 9.552</td>
</tr>
</tbody>
</table>

? = 3.6353

(last row: type I and II are equal)
2.2 General linear tests - REDUCED vs. FULL MODEL

Example: (cheese data)

a) We want to know whether or not the variables $X_1$ (ACETIC) and $X_2$ (H2S) should be added to the model. Does the smaller model do just as well at describing the data as the full model?

### Analysis of Variance: Reduced Model

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>3800.4</td>
<td>3800.4</td>
<td>27.55</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>3862.5</td>
<td>137.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>7662.887</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Analysis of Variance: Full Model

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>4994.509</td>
<td>1664.836</td>
<td>16.22</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>26</td>
<td>2668.378</td>
<td>102.629</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>7662.887</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b) The ‘lm’ in R produces the following sequential sums of squares for the cheese data:

```r
> ch.lm=lm(TASTE ~ ACETIC+H2S+LACTIC, data= cheese)
```

Obtain the following sequential sums of squares to check whether or not each variable should be added to the model (given the preceding terms): $R(X_1)$, $R(X_2|X_1)$, $R(X_3|X_1, X_2)$

State the appropriate test hypotheses for each F-value computed in the ANOVA table.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACETIC</td>
<td>1</td>
<td>2314.14</td>
<td>2314.14</td>
<td>22.55</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>H2S</td>
<td>1</td>
<td>2147.11</td>
<td>2147.11</td>
<td>20.92</td>
<td>0.0001</td>
</tr>
<tr>
<td>LACTIC</td>
<td>1</td>
<td>533.26</td>
<td>533.26</td>
<td>5.20</td>
<td>0.0311</td>
</tr>
</tbody>
</table>
c) Suppose that we had used the different ordering of model:

```R
ch.lm1=lm(TASTE ~ H2S+LACTIC+ACETIC, data=cheese)
```

Compute $R(X_1)$, $R(X_2|X_1)$, $R(X_3|X_1,X_2)$. What is the major difference between results of b) and c) due to different ordering?

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2S</td>
<td>1</td>
<td>4376.8</td>
<td>4376.8</td>
<td>42.65</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>LACTIC</td>
<td>1</td>
<td>617.1</td>
<td>617.1</td>
<td>6.01</td>
<td>0.02123</td>
</tr>
<tr>
<td>ACETIC</td>
<td>1</td>
<td>0.6</td>
<td>0.6</td>
<td>0.0054</td>
<td>0.94193</td>
</tr>
</tbody>
</table>

d) In R, `drop1(ch.lm, test = "F")` gives the partial sums of squares.

After adjusting for the effects of acetic and lactic concentrations, do we have significant evidence that the hydrogen sulfide concentration is important in describing taste?

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACETIC</td>
<td>1</td>
<td>0.56</td>
<td>0.56</td>
<td>0.0054</td>
<td>0.9419</td>
</tr>
<tr>
<td>H2S</td>
<td>1</td>
<td>1007.69</td>
<td>1007.69</td>
<td>9.8187</td>
<td>0.0042</td>
</tr>
<tr>
<td>LACTIC</td>
<td>1</td>
<td>533.26</td>
<td>533.26</td>
<td>5.1959</td>
<td>0.0311</td>
</tr>
</tbody>
</table>

The Lack of Fit Test

When we want to determine whether a specified regression function adequately fits the data, we can conduct a lack of fit test. However, it is important to note that the test requires repeated observations (replications) for at least one of the values of the predictors ($X$). This test is also based on decomposing sums of squares (due to Errors) and the test procedure can be derived in the same way as testing the full vs. reduced model.

$$F^* = \frac{(SSLF - SSPE)/(n - p - (n - c))}{MSPE}$$

where $p$ is the number of regression parameters and $c$ is the number of distinct $X$ values, SSLF denotes the lack of fit sum of squares, and SSPE (thus, MSPE for mean squares) stands for the sum of squares due to pure error.
2.3 Qualitative Independent Variables

Types of variables

- Qualitative variables: Numerical measurements on the phenomena of interest are not possible. Rather, the observations are categorical.
  
  e.g. gender (female, male), Company status (private, public), Treatment (yes, no), blood pressure rating (low, average, high)

- Quantitative variables: The observations are in the form of numerical values.
  
  e.g. age, income, temperature, number of defectives, etc.

Qualitative or “classification” variables can be included as explanatory variables in regression models by using indicator variables (also called ‘dummy’ or ‘binary’ variables).

Example: On average, do smoking mothers have babies with lower birth weight?

Response (Y): birth weight in grams of baby

$X_1$: length of gestation in weeks, $X_2$: Smoking status of mother (smoker or non-smoker)

Then, a first order model with one binary and one quantitative predictor appears to be a natural model to formulate for these data:

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad \text{where} \quad X_{i2} = \begin{cases} 
1 & \text{if mother } i \text{ smokes} \\
0 & \text{otherwise} 
\end{cases} \]

Q. Why not just fit two separate regression functions one for the smokers and one for the non-smokers?

The combined regression model assumes that the slope for the two groups are equal and that the variances of the error terms are equal. Then, it is better to use as much data as possible to estimate standard errors of regression coefficients for testing and confidence intervals.

Pooling your data and fitting the combined regression function allows you to easily and efficiently answer research questions concerning the binary predictor variable.
Example. If we are interested in quantifying the relationship between total population (of a metropolitan area) and number of active physicians it may be important to take (4) geographic regions into account. The indicator variables can be easily used to identify each of 4 regions as follows:

\[
X_1 = \begin{cases} 
1 & \text{if region 1} \\
0 & \text{otherwise}
\end{cases}, \quad X_2 = \begin{cases} 
1 & \text{if region 2} \\
0 & \text{otherwise}
\end{cases}, \\
X_3 = \begin{cases} 
1 & \text{if region 3} \\
0 & \text{otherwise}
\end{cases}, \quad X_4 = \begin{cases} 
1 & \text{if region 4} \\
0 & \text{otherwise}
\end{cases}
\]

Then, we can fit a model for the effect of total population on the number of physicians with a separate intercept for each region:

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 (\text{Pop})_i + \epsilon_i
\]

For \( n = 10 \) observations, the design matrix would look something like this:

\[
X = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & X_{15} \\
1 & 1 & 0 & 0 & 0 & \text{region 1} \\
1 & 0 & 1 & 0 & 0 & X_{35} \\
1 & 0 & 1 & 0 & 0 & X_{45} & \text{region 2} \\
1 & 0 & 0 & 1 & 0 & X_{55} \\
1 & 0 & 0 & 1 & 0 & X_{65} & \text{region 3} \\
1 & 0 & 0 & 1 & 0 & X_{75} \\
1 & 0 & 0 & 0 & 1 & X_{85} \\
1 & 0 & 0 & 0 & 1 & X_{95} & \text{region 4} \\
1 & 0 & 0 & 0 & 1 & X_{105} \\
\end{bmatrix}
\]

\[
\beta = \begin{bmatrix} 
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5 
\end{bmatrix}
\]

This is not of full rank, so in order to compute parameter estimates we need to either put restrictions on the parameters or redefine some of the independent variables. When the
design matrix is not of full rank, we are trying to estimate too many parameters. One simple way to deal with this problem is to drop one column. In the above, columns 2 through 5 of X matrix add up to the first column. If we drop the last of these columns, $X_4$, the result is the following reparameterized model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 (\text{Pop})_i + \epsilon_i$$

this model is essentially the same as before, but parameters have different meanings. For each region there is a separate intercept and all regions have a common slope.

region1: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_4 (\text{Pop})_i + \epsilon_i$

region2: $Y_i = \beta_0 + \beta_2 X_{i2} + \beta_4 (\text{Pop})_i + \epsilon_i$

region3: $Y_i = \beta_0 + \beta_3 X_{i3} + \beta_4 (\text{Pop})_i + \epsilon_i$

region4: $Y_i = \beta_0 + \beta_4 (\text{Pop})_i + \epsilon_i$

In this parameterization $\beta_0$ is not the average intercept for all regions. Instead it is the intercept for region 4. The other parameters such as $\beta_1$ can be interpreted as the difference between the intercept for that region and the intercept for region 4.

To test whether all regions have the same intercept we simply test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ using a general linear test. Reject $H_0$ if

$$F^* = \frac{R(X_1, X_2, X_3|\text{Pop})/3}{\text{MSE(full)}} > F_{1-a}(3, n - 5)$$

The model with separate intercepts but common slopes may not be realistic. You may need separate slopes as well. You can fit a model that has entirely distinct lines for each region using indicator variables as well. One parameterization for this type of model just builds on our separate intercepts model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 (\text{Pop})_i + \beta_5 X_{i1}(\text{Pop})_i + \beta_6 X_{i2}(\text{Pop})_i + \beta_7 X_{i3}(\text{Pop})_i + \epsilon_i$$

In this model $\beta_0$ is the intercept for region 4 and $\beta_4$ is the slope for region 4. The terms $\beta_5, \beta_6$ and $\beta_7$ give the difference between each region’s slope and the slope for region 4.

side notes: Fitting the full model this way gives the same slopes and intercepts for the four regions as if you had fit four separate regression lines. The MSE from the full model is the
pooled variance estimate from the individual regressions: \( \text{MSE} = \frac{\text{SSE}_1 + \text{SSE}_2 + \text{SSE}_3 + \text{SSE}_4}{n_1 + n_2 + n_3 + n_4 - 8} \) If the variances are different for the different groups, then it is best to do separate regressions so that you get separate variance estimates. Sometimes pointing out that the variances are different is just as important as determining whether the slopes are different.

If the variances are similar for the different groups, or if a transformation will make the variance fairly homogeneous, then it is more efficient to fit the single regression model than to fit four separate regressions.

There are other ways of dealing with the non-full rank design matrix. In particular, if we code the data with 1, 0, or -1 for each region as follows,

\[
Z_1 = \begin{cases} 
1 & \text{if region 1} \\
0 & \text{if region 2 or 3} \\
-1 & \text{region 4}
\end{cases} \\
Z_2 = \begin{cases} 
1 & \text{if region 2} \\
0 & \text{if region 1 or 3} \\
-1 & \text{region 4}
\end{cases} \\
Z_3 = \begin{cases} 
1 & \text{if region 3} \\
0 & \text{if region 1 or 2} \\
-1 & \text{region 4}
\end{cases}
\]

The resulting design matrix for this model is

\[
X = \begin{bmatrix} 
1 & 1 & 0 & 0 & X_{15} \\
1 & 1 & 0 & 0 & X_{25} \\
1 & 0 & 1 & 0 & X_{35} \\
1 & 0 & 1 & 0 & X_{45} \\
1 & 0 & 0 & 1 & X_{55} \\
1 & 0 & 0 & 1 & X_{65} \\
1 & 0 & 0 & 1 & X_{75} \\
1 & -1 & -1 & -1 & X_{85} \\
1 & -1 & -1 & -1 & X_{95} \\
1 & -1 & -1 & -1 & X_{105}
\end{bmatrix}
\]

The equations for each region are

region 1: \( Y_i = \beta_0 + \beta_1 + \beta_4 (\text{Pop})_i + \epsilon_i \)
2.3 Qualitative Independent Variables

region 2:  \[ Y_i = \beta_0 + \beta_2 + \beta_4(\text{Pop})_i + \epsilon_i \]

region 3:  \[ Y_i = \beta_0 + \beta_3 + \beta_4(\text{Pop})_i + \epsilon_i \]

region 4:  \[ Y_i = \beta_0 - (\beta_1 + \beta_2 + \beta_3) + \beta_4(\text{Pop})_i + \epsilon_i \]

Interaction Effects

A regression model contains interaction effects if the response function is not additive and cannot be separated into distinct functions of each of the individual predictors. Two predictors interact if the effect on the response variable of one predictor depends on the value of the other.

Example. Some researchers were interested in comparing the effectiveness of three treatments (A, B, C) for severe depression and collected a random sample of \( n = 36 \) severely depressed individuals.

\[ Y_i = \text{measure of the effectiveness of the treatment for individual } i \]

\[ X_{i1} = \text{age (in years) of individual } i \]

\[ X_{i2} = 1 \text{ if individual } i \text{ received treatment A and 0, if not.} \]

\[ X_{i3} = 1 \text{ if individual } i \text{ received treatment B and 0, if not.} \]

<table>
<thead>
<tr>
<th>( Y )</th>
<th>56</th>
<th>41</th>
<th>40</th>
<th>28</th>
<th>55</th>
<th>25</th>
<th>46</th>
<th>71</th>
<th>48</th>
<th>63</th>
<th>52</th>
<th>62</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>45</td>
<td>58</td>
<td>46</td>
<td>58</td>
<td>34</td>
<td>65</td>
<td>55</td>
<td>57</td>
<td>59</td>
<td>64</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>62</td>
<td>36</td>
<td>69</td>
<td>47</td>
<td>73</td>
<td>64</td>
<td>60</td>
<td>62</td>
<td>71</td>
<td>62</td>
<td>70</td>
<td>71</td>
</tr>
<tr>
<td>AGE</td>
<td>21</td>
<td>23</td>
<td>30</td>
<td>19</td>
<td>28</td>
<td>23</td>
<td>33</td>
<td>67</td>
<td>42</td>
<td>33</td>
<td>33</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>43</td>
<td>38</td>
<td>37</td>
<td>43</td>
<td>27</td>
<td>43</td>
<td>45</td>
<td>48</td>
<td>47</td>
<td>48</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>58</td>
<td>29</td>
<td>53</td>
<td>29</td>
<td>58</td>
<td>66</td>
<td>67</td>
<td>63</td>
<td>59</td>
<td>51</td>
<td>67</td>
<td>63</td>
</tr>
<tr>
<td>TRT</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>C</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>B</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>C</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
</tr>
</tbody>
</table>

PAGE 40
As the data indicates that $X_1$ and $X_2$ interact and that $X_1$ and $X_3$ interact, we formulate the model as:

$$Y_i = \beta_0 + \beta_1X_{i1} + \beta_2X_{i2} + \beta_3X_{i3} + \beta_4X_{i1}X_{i2} + \beta_5X_{i1}X_{i3} + \epsilon_i.$$  

The estimated regression function is

$$Y = 6.21 + 1.0334\text{AGE} + 41.30X_2 + 22.71X_3 - 0.703\text{AGE}X_2 - 0.510\text{AGE}X_3$$

When we plug the possible values for $X_2$ and $X_3$ into the estimated regression function, we obtain the three best fitting lines one for each treatment (A, B and C) through the data.
2.3 Qualitative Independent Variables

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>1</td>
<td>3424.4</td>
<td>3424.4</td>
<td>222.29</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$X_2$</td>
<td>1</td>
<td>803.8</td>
<td>803.8</td>
<td>52.18</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$X_3$</td>
<td>1</td>
<td>1.2</td>
<td>1.2</td>
<td>0.077</td>
<td>0.7831</td>
</tr>
<tr>
<td>AGE $X_2$</td>
<td>1</td>
<td>375.0</td>
<td>375.0</td>
<td>24.34</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>AGE $X_3$</td>
<td>1</td>
<td>328.4</td>
<td>328.4</td>
<td>21.32</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

a) For every age, is there a difference in the mean effectiveness for the three treatments? i.e. test whether the regression functions are identical for each age, regardless of three treatments. $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ vs. $H_1$: at least one of these slope parameters is not 0.

b) Does the effect of age on the treatment’s effectiveness depend on treatment? $H_0 : \beta_4 = \beta_5 = 0$ vs. $H_1$: at least one of these parameters are not zero.

**Piecewise Linear Regression Models**

The basic idea behind the piecewise linear regression is that if the data follow different linear trends over different regions of the data, then we should model the regression function in “pieces.” The pieces can be connected or not connected. Here, we’ll fit a model in which the pieces are connected. We also use a dummy variable and an interaction term to define a piecewise linear regression model.

Example. The data set contains information on compressive strength ($Y$) of $n = 18$ batches of concrete against the proportion of water ($X$) mixed in with the cement:

A straight line estimated by the least squares may fit the data fairly well in some overall sense, but we could do better if we use a piecewise linear function. We instead split our original scatter plot into two pieces (where the water-cement ratio is 70%) and fit two separate, but connected lines, one for each piece.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{i1} - 70)X_{i2} + \epsilon_i.$$  

Alternatively, we can write the formulated piecewise model above as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^* + \epsilon_i,$$

PAGE 42
2.3 Qualitative Independent Variables

where $Y_i$ is the comprehensive strength, in pounds per square inch, of concrete batch $i$

$X_{i1}$ is the water-cement ratio, in $\%$, of concrete batch $i$

$X_{i2}$ is a dummy variable (0, if $X_{i1} \leq 70$ and 1, if $X_{i1} > 70$) of concrete batch $i$

$X_{i2}^*$ denotes the interaction term $(X_{i1} - 70)X_{i2}$

Incidentally, the $X$-value at which the two pieces of the model connect is called the “knot value.” The regression equation is ‘strength = 7.79 − 0.0663 ratio − 0.101X_{i2}^*.’

which yields two estimated regression lines, connected at $X = 70$ fitting the data quite well.

---

Regression with a straight line

with a piecewise linear function
2.3 Qualitative Independent Variables

Polynomial Regression

In many cases the usual linear relationship between the dependent variable and any independent variable we have assumed may not be adequate. One way to account for such a nonlinear relationship is through a polynomial or trigonometric regression model. Such models are linear in the parameters and the least squares method can be used for estimation of the parameters as long as the usual assumptions on the errors are appropriate. For example, a model for a single predictor, $X$, is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p + \epsilon,$$

where $p$ is the degree of the polynomial. For lower degrees, the relationship has a specific name (i.e., $p = 2$ is called quadratic, $p = 3$ is called cubic, $p = 4$ is called quartic, and so on). However, polynomial regression models may have other predictor variables in them as well, which could lead to interaction terms and the model can grow depending on your application.

The design matrix for the second-degree polynomial model is:

$$X = \begin{bmatrix}
1 & X_1 & X_1^2 \\
1 & X_2 & X_2^2 \\
\vdots & \vdots & \vdots \\
1 & X_n & X_n^2
\end{bmatrix}.$$

For the polynomial regression in one independent variable, an important aspect that distinguishes it from other multiple regression models is that the mean of the dependent variable is a function of a single independent variable. The fact that the independent variables in a simple polynomial model are functions of a single independent variable affects the interpretation of the parameters. In the second-degree model, the parameter $\beta_1$ is the slope only at $X = 0$. The parameter $\beta_2$ is half the rate of change in the slope of $E(Y)$.

The higher-degree polynomial models provide greatly increased flexibility in the response surface. Although it is unlikely that any complex process will be truly polynomial in form, the flexibility of the higher-degree polynomials allows any true model to be approximated to any desired degree of precision.
The full second-degree polynomial model in two variables is

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i1}X_{i2} + \beta_5 X_{i2}^2 + \epsilon_i. \]

The degree (or order) of an individual term in a polynomial is defined as the sum of the powers of the independent variables in the term. The degree of the entire polynomial is defined as the degree of the highest-degree term.

The squared terms allow for a curved response in each variable and the product term allows for the surface to be twisted. A quadratic response surface will have a maximum, a minimum, or a saddle point, depending on the coefficients in the regression equation. (Refer to Box and Draper (1987) for a further discussion of the analysis of the properties of quadratic response surfaces.)

Due to its extreme flexibility, some caution is needed in the use of polynomial models; it is easy to overfit a set of data with polynomial models. Extrapolation is particularly dangerous when higher-degree polynomial models are being used, since minor extrapolations can have serious errors. Nevertheless, polynomial response models have proven to be extremely useful for summarizing relationships.

The polynomial model is built sequentially, starting either with a first-degree polynomial and adding progressively higher-order terms as needed. The lowest-degree polynomial that accomplishes the degree of approximation needed or warranted by the data is adopted. It is common practice to retain in the model all lower-degree terms, regardless of their significance, that are contained in (or are subsets of) any significant term.

Example. Data are from a growth experiment with blue-green algae Spirulina platensis for the treatment where \( CO_2 \) is bubbled through the culture. The measure of algae density is the dependent variable. Consider a cubic polynomial model with the data for the first replicate. The (ordinary) least squares fit of the model is given by

\[ Y_i = 0.00948 + 0.53074X_i + 0.00595X_i^2 - 0.00119X_i^3 + \epsilon_i, \]

\[ (0.16761) \quad (0.09343) \quad (0.01422) \quad (0.00062) \]
2.3 Qualitative Independent Variables

<table>
<thead>
<tr>
<th>Day</th>
<th>rep1</th>
<th>rep2</th>
<th>Day</th>
<th>rep1</th>
<th>rep2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.530</td>
<td>.184</td>
<td>8</td>
<td>4.059</td>
<td>3.892</td>
</tr>
<tr>
<td>2</td>
<td>1.183</td>
<td>.664</td>
<td>9</td>
<td>4.349</td>
<td>4.367</td>
</tr>
<tr>
<td>3</td>
<td>1.603</td>
<td>1.553</td>
<td>10</td>
<td>4.699</td>
<td>4.551</td>
</tr>
<tr>
<td>4</td>
<td>1.994</td>
<td>1.910</td>
<td>11</td>
<td>4.983</td>
<td>4.656</td>
</tr>
<tr>
<td>5</td>
<td>2.708</td>
<td>2.585</td>
<td>12</td>
<td>5.100</td>
<td>4.754</td>
</tr>
<tr>
<td>6</td>
<td>3.006</td>
<td>3.009</td>
<td>13</td>
<td>5.288</td>
<td>4.842</td>
</tr>
<tr>
<td>7</td>
<td>3.867</td>
<td>3.403</td>
<td>14</td>
<td>5.374</td>
<td>4.969</td>
</tr>
</tbody>
</table>

where the standard errors of the estimates are given in parentheses. Assuming that a cubic model is adequate, we can test the hypotheses

a) that a quadratic polynomial model is adequate.

b) that a linear trend model is adequate. SSE(full) = 0.013658, SSE(Red) = 1.458.

c) With two replicates per each day, we can test the adequacy of a quadratic polynomial model. We fit the quadratic model with whole data (including rep1 and rep2) to obtain

\[
\text{SSE(reduced)} = .7984 \text{ and pure-error sum of squares} = .6344.
\]

Note that the natural polynomials with terms \(X_i, X_i^2,\) and \(X_i^3,\) etc. give (nearly) linearly dependent columns of the \(X\) matrix leading to multicollinearity problems. For the cubic polynomial model (as in the above example), a set of orthogonal polynomials can be defined:

\[
\begin{align*}
O_{0i} & = 1, \\
O_{1i} & = 2X_i - 15, \\
O_{2i} & = .5X_i^2 - 7.5X_i + 20, \\
O_{3i} & = \frac{5}{3}X_i^3 - 37.5X_i^2 + \frac{698.5}{3}X_i - 340.
\end{align*}
\]

where \(O_{1i}, O_{2i},\) and \(O_{3i}\) are linear combinations of the natural polynomials \(X_i, X_i^2,\) and \(X_i^3.\)

For the data above, \(Y_i = 3.48164 + .19198O_{1i} - .04179O_{2i} - .00072O_{3i} + \epsilon_i,\)

\[
(0.03123) (0.00387) (0.00433) (0.00037)
\]

PAGE 46
The orthogonal polynomials can be obtained using the Gram Schmidt orthogonalization procedure or with a function ‘poly’ in R.

2.4 Data Transformations

There are many situations in which transformations of the dependent or independent variables are helpful in least squares regression. Transformations of the dependent variable are often used as remedies for nonnormality and for heterogeneous variances of the errors.

Note that data transformation definitely requires a ‘trial and error’ approach. In building the model, we try a transformation and then check to see if the transformation eliminated the problems with the model. If it doesn’t help, we try another transformation until we have an adequate model.

Previously, we learned tools for detecting problems with a linear regression model. Once we’ve identified problems with the model, we have a number of options:

- If important predictor variables are omitted, see whether adding the omitted predictors improves the model.
- If the mean of the response is not a linear function of the predictors, try a different function. e.g. polynomial regression or applying a logarithmic transformation to the response variable allowing for a nonlinear relationship between the response and the predictors while remaining within the multiple linear regression framework.
- If there are unequal error variances, try transforming the response and/or predictor variables or use “weighted least squares regression” (or generalized least squares).
- If an outlier exists, try using a “robust estimation procedure.”
- If error terms are not independent, try fitting a “time series model.”

Another reason for making transformations is to simplify the relationship between the dependent variable and the independent variables. Curvilinear relationships between two variables
frequently can be simplified by a transformation on either one or both of the variables. Many models nonlinear in the parameters can be linearized, reexpressed as a linear function of the parameters, by appropriate transformations. For example, the relationship

\[ Y = \alpha X^\beta \]

is linearized by taking the logarithm of both sides of the equality giving

\[ \ln(Y) = \ln(\alpha) + \beta \ln X \quad \text{or} \quad Y^* = \alpha^* + \beta X^*. \]

Logarithms are often used in data transformations because they are connected to common power curve (as above) and exponential growth \( Y = \alpha \exp(X\beta) \) relationships.

It should be remembered that it may not be possible to find a set of transformations that will satisfy all objectives. A transformation on the dependent variable will change the variance of the \( Y \)-variable and the errors, or, a transformation to stabilize variance may cause nonnormality. Fortunately, transformations for homogeneity of variance and normality often tend to go hand-in-hand so that both assumptions are more nearly satisfied after an appropriate transformation (Bartlett, 1947). The logit, arcsin, and probit transformations of the \( Y \)-variable that are used to stabilize variance and straighten relationships also make the distribution more normal-like by stretching the tails of the distribution, values near zero or one, to give a more bell-shaped distribution.

Transformations of the \( X \)-variable(s) (e.g., \( X^{-1}, X^2, \ln(X) \)) are recommended in practice when there are strong nonlinear trends in one or more residual plots. Box-Cox transformations are a family of power transformations on \( Y \) (refer to page 16), where \( \lambda \) is a parameter to be determined using the data.

It may be that no transformation adequately stabilized the variances, or a transformation made to simplify a relationship left heterogeneous variances. The weighted least squares and generalized least squares provide minimum variance linear unbiased (least squares) estimates when the variance-covariance matrix of the errors is an arbitrary symmetric positive definite matrix.
2.5 Weighted Least Squares

The least squares regression method discussed so far was based on the assumptions that the errors are normally distributed independent random variables with a common variance. Least squares estimators based on these assumptions is referred to as ordinary least squares estimates and they have the desirable property of being the best (minimum variance) among all possible linear unbiased estimators by Gauss-Markov theorem. When the normality assumption is satisfied, the least squares estimators are also maximum likelihood estimators.

If the variance of $Y$ is not constant, there may be a transformation of $Y$ that will stabilize the variance. This is the most common approach, but under certain circumstances there is another method that is very attractive. If the variance heterogeneity is known up to a constant (and the observations $Y_i$ are uncorrelated), then we can use the method of weighted least squares:

$$\text{Var}(\epsilon_i) = \sigma^2_i = \frac{\sigma^2}{w_i} = \text{Var}(Y_i) \text{ and Cov}(\epsilon_i, \epsilon_j) = 0$$

Let $W = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w_n \end{bmatrix}$

What we do here is minimizing a sum of squares, $(Y - X\beta)'W(Y - X\beta)$, where we weight each observation according to its variability. If an observation is highly variable it gets low weight, and if an observation has small variance it gets greater weight.

The weighted least squares estimates of $\beta$ are

$$b = (X'WX)^{-1}X'WY$$

where the weighting matrix, $W$, is proportional to the inverse of the variance-covariance matrix of $Y$. Also,

$$\text{Var}(b) = (X'WX)^{-1}X'W\text{Var}(Y)WX(X'WX)^{-1} = \sigma^2(X'WX)^{-1}$$
When you compare the weighted least squares (WLS) results with those from ordinary least squares it often happens that the estimated coefficients do not change very much. However, the weighted least squares is more efficient than the ordinary least squares; i.e. the parameters are estimated more precisely. Note, however, that if you use ordinary least squares when heteroscedasticity is present (non-constant error variances), the estimated variances of the regression coefficients that come out of a standard regression program will not actually be the correct variances, so they may look smaller than those under WLS. Also, if you compute confidence bands or prediction intervals, the ones computed using WLS will get wider as \( X \) increases, as they should. This is probably the most important benefit of using the weighted least squares. We want the confidence and prediction intervals to be narrow when \( \text{Var}(Y_i) \) is small.

When you do the weighted least squares, it is important to use studentized residuals rather than raw residuals for diagnostic plots. The raw residuals do not have homogeneous variances and will not show the results of the weighting, but the studentized residuals do have homogeneous variances.

The weighted least squares is a special case of generalized least squares (GLS) where the error terms not only may have different variances but pairs of error terms may also be correlated.

The generalized least squares (GLS) estimator

The generalized least squares estimates of \( \beta \) is

\[
\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}Y
\]

where \( V \) is the variance-covariance matrix of \( Y \). When \( V \) is a diagonal matrix, \( \hat{\beta}_{GLS} = \hat{\beta}_{WLS} \).

Note that there exists a \( n \times n \) nonsingular matrix \( K \) such that \( KK' = V, \text{Var}(\epsilon) = V \).

In practice, \( V \) is not known and need to be estimated by \( \hat{V} \). Then, the estimated generalized least squares estimates of \( \beta \) is obtained by an iterative algorithm:

\[
\hat{\beta}_{EGLS} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y.
\]
2.6 Model Building

In many practical applications, a researcher has a large set of candidate predictor variables from which he/she tries to identify the most appropriate predictors to include in his regression model. However, for example, if a researcher has (only) 10 candidate predictor variables, then there are $2^{10} = 1024$ possible regression models from which to choose.

Some models have a natural hierarchy (polynomial models/Models with interactions). In this case, lower order terms should not be removed from the model before higher order terms in the same variable.

In general, we want to explain the data in the simplest way, since collinearity is caused by having too many variables or unnecessary predictors will add noise to the estimation of other quantities that we are interested in. However, an underspecified model is a model in which important predictors are missing and such a model yields biased regression coefficients and biased predictions of the response.

Here we aim to learn some variable selection methods such as stepwise regression, best subsets regression and LASSO regression so that the resulting model is simple and useful.

**Stepwise Regression**

We build our regression model from a set of candidate predictor variables by entering and removing predictors in a stepwise manner into our model until there is no justifiable reason to enter or remove any more. we start with no predictors in our “stepwise model.” Then, at each step along the way we either enter or remove a predictor based on the partial F-tests that is, the t-tests for the slope parameters.

(a) **Backward Elimination**

1. Start with all the predictors in the model
2. Remove the predictor with highest $p$-value greater than $\alpha_R$
3. Refit the model and goto 2
4. Stop when all $p$-values are less than $\alpha_R$
(b) Forward Selection

1. Start with no variables.

2. For all predictors not in the model, check their $p$-value if they are added to the model. Choose the one with lowest $p$-value less than $\alpha_E$

3. Continue until no new predictors can be added.

Note that $\alpha_R$ and $\alpha_E$ will typically be greater than the usual 0.05 level so that it is not too difficult to enter predictors into the model.

(c) Stepwise Regression: combination of backward elimination and forward selection.

1. Fit each of the one-predictor models. Of those predictors, the predictor that has the smallest t-test $p$-value than $\alpha_E$ is put in the stepwise model. If no predictor has a $p$-value than $\alpha_E$, stop.

2. Fit each of the two-predictor models that include the predictor chosen in 1. Again, of those predictors, the predictor that has the smallest t-test $p$-value than $\alpha_E$ is put in the model. Step back and see if entering the newly chosen variable into the model somehow affects the significance of the previously chosen predictor in 1. If $p$-value for the previously chosen variable is greater than $\alpha_R$, remove the variable.

3. Fit each of the three-predictor models with two predictors chosen in 2 and so on.

4. Continue the same procedure as above with more variables.

5. Stop when adding an additional predictor does not yield a $p$-value less than $\alpha_E$

Warnings:

- The procedure yields a single final model, although there are often several equally good models. The final model is not guaranteed to be optimal in any specified sense: it does not mean that all the important predictor variables for predicting $Y$ have been identified, nor that all the unimportant predictor variables have been eliminated.

- Stepwise regression does not take into account a researcher’s knowledge about the predictors. It may be necessary to force the procedure to include important predictors.
- One should not over-interpret the order in which predictors are entered into the model.

- It is possible that we may have committed a Type I or Type II error along the way.

Best subsets regression

We select the subset of predictors that does the best at meeting some well-defined objective criterion, such as having the largest \( R^2 \) (adjusted \( R^2 \)) value or the smallest MSE.

1. Identify all of the possible regression models derived from all of the possible combinations of the candidate predictors.

2. Determine the one-predictor models that do the ‘best’ at meeting some well-defined criteria, the two-predictor models that do the ‘best’, the three-predictor models that do the ‘best’, and so on.

3. Further evaluate and refine models by residual analyses, transforming variables, adding interaction terms, etc. until the model meets necessary conditions allowing you to answer your research question.

The adjusted \( R^2 \)-value is defined as:

\[
R_a^2 = 1 - \left( \frac{n - 1}{n - p} \right) \frac{SSE}{SST} = 1 - \left( \frac{n - 1}{n - p} \right) (1 - R^2) = 1 - \left( \frac{n - 1}{SST} \right) MSE.
\]

Criterion-based procedures

1. information criterion statistics:

Akaike’s Information Criterion (AIC), Bayesian Information Criterion (BIC) (or called Schwartz’s Bayesian Criterion (SBC)), and Amemiya’s Prediction Criterion (APC)

\[
AIC = n \ln(SSE) - n \ln n + 2p
\]

\[
BIC = n \ln(SSE) - n \ln n + p \ln(n)
\]

\[
APC = \frac{n + p}{n(n - p)} SSE
\]
where \( n \): sample size and \( p \): number of regression coefficients in the model being evaluated including the intercept.

Notice that the only difference between AIC and BIC is the multiplier of \( p \). Each of the information criteria is used in a similar way in comparing two models, the model with the lower value is preferred.

The BIC places a higher penalty on the number of parameters in the model so will tend to reward more parsimonious (smaller) models. This stems from one criticism of AIC in that it tends to overfit models.

2. PRESS (prediction residual sum of squares):

\[
PRESS = \sum_{i=1}^{n} (e_{i(i)})^2
\]

where \( e_{i(i)} \)'s are the residuals calculated omitting \( i \)th observation in the regression fit. The model with the lowest PRESS value is selected. This tends to pick larger models (which may be desirable if prediction is the objective).

3. Mallows’ \( C_p \)-statistic

It estimates the size of the bias that is introduced into the predicted responses by having an underspecified model.

The bias in predicted response:

\[
Bias_i = E(\hat{Y}_i) - E(Y_i)
\]

and the average MSE of prediction:

\[
\frac{1}{\sigma^2}E(\hat{Y}_i - E(Y_i))^2
\]

which can be estimated by the \( C_p \) statistic:

\[
C_p = p + \frac{(MSE_p - MSE_{all})(n - p)}{MSE_{all}} = \frac{SSE_p}{MSE_{all}} + 2p - n
\]

where \( MSE_p \) is the mean squared error from fitting the model containing the subset of \( p - 1 \) predictors (\( p \) parameters with the intercept), and \( MSE_{all} \) is the mean squared error obtained from fitting the model containing all of the candidate predictors.
Using $C_p$ to identify ‘best’ models:

- Identify subsets of predictors for which the $C_p$ value is near $p$ (if possible). The full model always yields $C_p = p$, so don’t select the full model based on $C_p$.

- If all models, except the full model, yield a large $C_p$ not near $p$, it suggests some important predictor(s) are missing from the analysis.

- If a number of models have $C_p$ near $p$, choose the model with the smallest $C_p$ value, thereby insuring that the combination of the bias and the variance is at a minimum.

- When more than one model has a small value of $C_p$ value near $p$, in general, choose the simpler model or the model that meets your research needs.

**Cross-validation**

It is a model validation technique with which the regression equation of the model fit is used to the original dataset to make predictions for the new dataset. Then, we can calculate the prediction errors (differences between the actual response values and the predictions) and summarize the predictive ability of the model by the mean squared prediction error (MSPE). This gives an indication of how well the model will predict in the future.

The sample data is usually partitioned into a training (or model-building) set, which we can use to develop the model, and a validation (or prediction) set, which is used to evaluate the predictive ability of the model.

The $K$-fold cross-validation partitions the sample dataset into $K$ parts which are (roughly) equal in size. For each part, we use the remaining $K - 1$ parts to estimate the model of interest (i.e., the training sample) and test the predictability of the model with the remaining part (i.e., the validation sample). Then, the sum of squared prediction errors can be computed and combining $K$ estimates of prediction error produces a $K$-fold cross-validation estimate.

When $K = 2$, it is usually preferable to residual diagnostic methods and takes not much longer to compute.

When $K = n$, this is called **leave-one-out cross-validation**. That means that $n$
2.7 Multicollinearity

Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated.

Types of multicollinearity

- Structural multicollinearity: a mathematical artifact caused by creating new predictors from other predictors, such as creating the predictor $X_2$ from the predictor $X$.

- Data-based multicollinearity: a result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

When it exists, there can be one of the following problems:

- the estimated regression coefficient of any one variable depends on other predictors that are included in the model.

- the precision of the estimated regression coefficients decreases as more predictors are added to the model.

- the marginal contribution of any one predictor variable in reducing the error sum of squares depends on the other predictors that are included in the model.

- hypothesis tests for $\beta_k = 0$ may yield different conclusions depending on which predictors are in the model.

Example. The researchers were interested in determining if a relationship exists between

separate data sets are trained on all of the data (except one point) and then prediction is made for that one point. The evaluation of this method is very good, but often computationally expensive. Note that the $K$-fold cross-validation estimate of prediction error is identical to the PRESS statistic.
blood pressure and age, weight, body surface area, duration, pulse rate and/or stress level. (refer to the data set on the web)

\[ Y: \text{blood pressure (BP, in mm Hg)} \] - 20 individuals with high blood pressure

\[ X_1: \text{age (in years)}, X_2: \text{weight (in kg)}, X_3: \text{body surface area (BSA in square meter)} \]

\[ X_4: \text{duration of hypertension (in years)}, X_5: \text{basal pulse (in beats per minute)}, X_6: \text{stress index} \]

Q. What is the effect on the regression analysis if the predictors are perfectly uncorrelated?

When there is no apparent relationship at all between the predictors \( X_1 \) and \( X_2 \) so that the correlation between them is zero (\( X_1 \) and \( X_2 \) are orthogonal to each other), the regression parameter estimates (and their standard errors) remain the same regardless of the used model form and the sequential sum of squares are the same as the partial sum of squares.

Reducing Multicollinearity

- Remove one or more of the violating predictors from the regression model.

- Collect additional data under different experimental or observational conditions. (Data need to be collected in such a way to ensure that the correlation among the violating predictors is actually reduced.)

- When multicollinearity is a mathematical artifact caused by creating new predictors from other predictors (e.g. in polynomial regression), multicollinearity can be reduced by ‘centering the predictors’ i.e. centering a predictor merely entails subtracting the mean of the predictor values in the data set from each predictor value.

\section*{2.8 Ridge Regression}

In some cases, the least squares estimates can not be computed or is subject to a very high-variance:
1. $X$ is not of full rank (i.e., its columns are not linearly independent).

2. The number of predictors is larger than the number of responses ($p > n$).

3. **Multicollinearity**: $X$ is of full rank, but there exists strong correlation between variables. $\Rightarrow \hat{\beta}_j$’s can explode since $(X'X)^{-1}$ can grow very large in terms of matrix norm (leading to a high variance).

4. The number of predictors $p$ is very large ($p \approx n$) – estimates suffer from high-variance when more terms are included in the model.

To control the variance of $\hat{\beta}_j$’s, we can regularize the coefficients, i.e., control how large they can grow. Ridge regression (or Tikhonov Regularization) estimators can be computed in all 1-4 cases stated above (Hoerl and Kennard, 1970).

Ridge regression estimator $\hat{\beta}_\lambda$ minimizes the **penalized residual sum of squares criterion**:

$$J_\lambda(\beta) = \|Y - X\beta\|^2 + \lambda\|\beta\|^2$$

$$= \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda > 0$ is the **ridge (shrinkage) parameter**. Equivalently,

$$\hat{\beta}_\lambda = \arg \min_\beta \|Y - X\beta\|^2 \text{ subject to } \|\beta\|^2 \leq s$$

which makes explicit the size constraint on the parameters. There is 1-to-1 correspondence between $\lambda$ and $s$. $\lambda$ offers a tradeoff between the regularizing constraint and minimization of the sum of squared residuals. The bigger the $\lambda$ the greater is the amount of shrinkage of the coefficients toward zero.

Note that $J_\lambda(\beta)$ is strictly convex and hence has a unique solution $\hat{\beta}_\lambda$. Thus, setting the derivatives to zero, $\nabla_\beta J_\lambda(\beta) = 0$ yields that $\hat{\beta}_\lambda$ solves

$$(X'X + \lambda I)\beta = X'Y$$

Then, the ridge regression (RR) estimator has a closed-form solution:

$$\hat{\beta}_\lambda = (X'X + \lambda I)^{-1}X'Y$$

PAGE 58
The additional “ridge” down the diagonal elements guarantees that $X'X$ is invertible.

When $X'X = I$ (i.e., $X$ is orthonormal) and $n > p$: ($\hat{\beta}$ denotes OLS estimates.)

$$\hat{\beta}_\lambda = \frac{\hat{\beta}}{1 + \lambda},$$

- This form illustrates the essential feature of ridge regression: shrinkage towards zero
- The ridge penalty introduces bias but reduces the variance of the estimate (bias-variance tradeoff).

As $\lambda \to 0$, $\hat{\beta}_\lambda \to \hat{\beta}$, whereas $\lambda \to \infty$, $\hat{\beta}_\lambda \to 0$.

Theorem. If $\epsilon_i$’s are i.i.d. zero mean with variance $\text{Var}(\epsilon_i) = \sigma^2$, then the bias of the ridge regression estimator is

$$\text{Bias}(\hat{\beta}_\lambda) = E[\hat{\beta}_\lambda] - \beta = -\lambda R_\lambda \beta$$

and the variance-covariance matrix is

$$\text{Var}(\hat{\beta}_\lambda) = E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])'] = \sigma^2 R_\lambda (X'X) R_\lambda, \text{ where } R_\lambda = (X'X + \lambda I)^{-1}.$$ 

Recall that mean squared error (MSE) = variance + (bias)$^2$, and in multiparameter problems:

$$\text{MSE}(\hat{\beta}_\lambda) = E[(\hat{\beta}_\lambda - \beta)(\hat{\beta}_\lambda - \beta)^T] = \text{Var}(\hat{\beta}_\lambda) + \text{Bias}(\hat{\beta}_\lambda) \cdot [\text{Bias}(\hat{\beta}_\lambda)]'$$

which also implies that “total MSE”:

$$\sum_{i=1}^{p} \text{MSE}(\hat{\beta}_{\lambda,i}) = \sum_{i=1}^{p} \text{Var}(\hat{\beta}_{\lambda,i}) + \sum_{i=1}^{p} [\text{Bias}(\hat{\beta}_{\lambda,i})]^2$$

as $\lambda \uparrow$ as $\lambda \uparrow$

There always exists $\lambda$ such that the total MSE of $\hat{\beta}_\lambda$ is smaller than the total MSE of the least squares estimate $\hat{\beta}$.

Q. How to choose the shrinkage parameter $\lambda$?
• Hoerl and Kennard (1970) suggested using *ridge traces*: - Plot the components of $\hat{\beta}_\lambda$ against $\lambda$ or rather $DF(\lambda)$ due to 1-to-1 relation. - Choose a $\lambda$ for which the coefficients are not rapidly changing and have sensible signs.

• This is a heuristic approach (requiring extensive computations for different choices of $\lambda$) and can be criticized from many perspectives.

• Nowadays, more standard approach is to use cross-validation (CV) of information theoretic criterions such as Bayesian Information Criterion (BIC).

Regression with a straight line with a piecewise linear function

### 2.9 LASSO (Least Absolute Shrinkage and Selection Operator)

Another way to deal with variable selection is to use regularization (or penalization) - regularize the regression coefficients. LASSO (Tibshirani, 1996) has been a popular technique for simultaneous linear regression estimation and variable selection. Specifically, we define $\hat{\beta}$ to minimize the penalized sums of squares

$$\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Or

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 \text{ subject to } \|\beta\|_1 \leq s$$
As before (in the ridge regression), the shrinkage (penalty) parameter $\lambda$ offers a tradeoff between the regularizing constraint and minimization of the sum of squared residuals. The bigger the $\lambda$ the greater is the amount of shrinkage. With LASSO, some of the estimated coefficients are shrunk all the way to zero. Another important role of $\lambda$ is for automated variable selection.

Example: Prostate cancer data ($n = 97, p = 8$) used in many text-books.

To study relationship between $Y =$ the level of prostate-specific antigen (lpsa) and a number of clinical measures in men who were about to receive a radical prostatectomy:

- $X_1 = \log$ cancer volume (lcavol)
- $X_2 = \log$ prostate weight (lweight)
- $X_3 = \text{age}$
- $X_4 = \log$ of the amount of benign prostatic hyperplasia(lbph),
- $X_5 = \text{seminal vesicle invasion (svi, binary)}$
- $X_6 = \log$ of capsular penetration (lcp),
- $X_7 = \text{Gleason score (gleason, ordered categorical)}$
- $X_8 = \text{percent of Gleason scores 4 or 5 (pgg45)}$
2.10 Robust Regression

Recall that the ordinary least squares estimates for linear regression are optimal when all of the regression assumptions are valid. When the errors are nonnormal, or there are big

Note: The dashed vertical line depicts the minimum BIC value
outliers, least squares regression can perform poorly. Robust regression is not overly affected by violations of assumptions by the underlying data-generating process.

**Robust loss functions**

1. Least absolute deviance (LAD) with $L_1$-loss: $\rho(e) = |e|$

2. Huber’s loss function: convex and differentiable

$$\rho_k(e) = \begin{cases} 
\frac{1}{2}e^2, & \text{for } |e| \leq k \\
ke - \frac{1}{2}k^2, & \text{for } |e| > k 
\end{cases}$$

where $c$ is a user-defined *tuning constant* that affects robustness and efficiency of the method.

3. Tukey’s biweight loss function

$$\rho_k(e) = \min \left\{ 1, 1 - \left( 1 - \left( \frac{e}{k} \right)^2 \right)^{\frac{3}{2}} \right\}$$

4. Least trimmed squares (LTS)

$$\sum_{j=1}^{h} (r^2)_{(i)}$$

where $(r^2)_{(i)}$ are the order statistics of the squared residuals. LTS estimator chooses the regression coefficients minimize the sum of the smallest $h$ of the squared residuals.