STAT 373 Online Fall 2020

Lin Xuntian

September 2020

Contents

1 Simple Linear Regression

1.1 Introduction to Regression Modelling

Understanding and quantifying variability in data is the backbone of statistics and statistical inference. In regression modelling, we attempt to explain, or account for, variation in a response variable, y, by using a statistical model to describe the relationship between the response and one or more explanatory variables, x_1, x_2, ... We can then use the model to learn and answer questions about relationships between the explanatory variables and the response, and/or predict the value of the response for a given set of explanatory variables.

Consider the following examples:

Example 1.1 An auditor wishes to determine whether the cost of overhead claimed by offices in a certain group is consistent with the office's attributes, including size, age, number of clients, number of employees, and the cost of living index of the city in which the office is located. To this end, the auditor creates a regression model to describe the relationship between these attributes and the (claimed) overhead in order to estimate the expected overhead for each office. The auditor can then investigate any claim for which a large discrepancy exists between the observed overhead and the expected overhead estimated from the model.

Example 1.2 Is there systemic gender inequity in the salaries of Waterloo faculty members?

Solution: To answer this question, a Waterloo working committee obtained information on each faculty member, including rank, academic unit, years of service, gender, and annual salary, and fit a regression model to the data. Based on the model, they found that, after accounting for rank, academic unit, years of service, and several other variables, males were getting paid significantly more than females, on average. The results from this regression analysis resulted in an immediate increase of \$2905 to the annual salaries of all female faculty members.

Example 1.3 Before listing a house, a realtor wishes to estimate its market value based on recent selling prices of homes in the area. Information is obtained on attributes of these homes that may help to account for selling price, such as size, lot size, number of rooms, number of bathrooms, number of stories, whether the house has a garage, etc., and a regression model is created to describe the relationship between selling price and these variables. The realtor can now use the model to estimate market value and predict selling price of the house, based on its attributes.

Note that in each of these examples, the objective for fitting a regression model is different, illustrating the power and usefulness of regression modelling. In the first example, the investigator wishes to detect discrepancies between an office's (claimed) overhead and the expected overhead for that office estimated from the regression model. In the second example, investigators wish to determine whether there is a relationship between gender and salary (i.e. whether there is a difference in mean salaries between male and female faculty members) after accounting for potentially confounding explanatory variables such as academic unit, rank, etc. Whereas in the last example, the objective was to predict the value of the response (market value) for a given set of explanatory variates.

We will be looking at each of these examples in more detail throughout the course. First, however, we will begin with a review of **simple linear regression**, in which we model the relationship between a single explanatory variable and the response.

1.2 Graphical and Numerical Summaries for Bivariate Data

With bivariate data, $\{x, y\}$, such as we have here, a scatterplot is an essential tool in visualizing and understanding the nature and strength of the relationship between an explanatory variable and a response.

Definition 1.1 A quantitative measure of the strength of a linear relationship between two variables is given by the correlation coefficient, r, defined as

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Properties of r:

1. $-1 \leqslant r \leqslant 1$, where the closer r is to 1 (-1), the stronger the positive (negative) relationship.
2. r is unitless. (Note that the units of the numerator and denominator will cancel). We can thus compare the relative strength of linear relationships across different scales and datasets.

1.3 The Simple Linear Regression Models

Consider again the scatterplot of overhead, y, vs. office size, x, for the audit data

1. a **deterministic** component, that describes the variation in accounted for by the functional form of the underlying relationship between y and x. Based on the scatterplot, the deterministic component can be adequately described by the linear function $\mu = \beta_0 + \beta_1 x$, where μ is the mean value of y for a given value of x.
2. an **error** term, denoted by the random variable ϵ, that describes the random variation in y not accounted for by the underlying relationship with x.

Definition 1.2 Incorporating both the deterministic and error components into our model yields the **simple linear regression** (SLR) model, expressed as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

where

- β_0 denotes the intercept parameter
- β_1 denotes the slope parameter
- the index i denotes the observation number (e.g. $\{x_3, y_3\}$ denotes the size and overhead associated with the third office in the dataset).

We will see in future lessons that, in order to derive the distributions of estimators for statistical inference procedures (i.e. confidence intervals and hypothesis tests for model parameters), we require certain distributional assumptions about the error random variable, ϵ.

In linear regression, we typically assume that the errors, ϵ_i , follow a normal distribution, with mean $= 0$, and variance denoted by σ^2. We also must assume that the errors are independent (recall for a normal random variable, independent errors $\Leftrightarrow \mathrm{Cov}(\epsilon_j, \epsilon_j) = 0,\ j \neq k$).

Definition 1.3 Incorporating these assumptions into our SLR model yields the **normal** model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2) \text{ ind.} \quad i = 1, \ldots, n$$

Assumptions of the Normal Model

1. the functional form (e.g. linear) of the relationship between and is correctly specified by the deterministic component of the model
2. the errors follow a normal distribution
3. errors have a constant variance, denoted by σ^2(this property is sometimes referred to as **homoskedasticity**)
4. the errors are independent

For the normal model to be an appropriate model to use in investigating the relationship between y and x, these assumptions must hold. Otherwise, our model will be inappropriate and any conclusions we obtain from our regression analysis will be invalid. We will be examining these model assumptions in more detail in later sections.

1.4 Least Squares Estimation of Model Parameters

Definition 1.4 Solving these normal equations yields the **least squares estimates**

$$\begin{aligned}\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 &= \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}\end{aligned}$$

1.5 The Fitted Model

Definition 1.5 The fitted model, or **fitted line**, for the SLR model is expressed as $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{\mu}$ is the estimated mean value of the response y for a given value of x.

Note that the fitted model is sometimes expressed in terms of the predicted value of the response, $\hat{\mu} = \hat{y}_0 + \hat{\beta}_1 x$. While $\hat{\mu}$ and $\hat{y}$ are identical in terms of the value they represent, there are subtle differences in their interpretation that we will discuss in a later section.

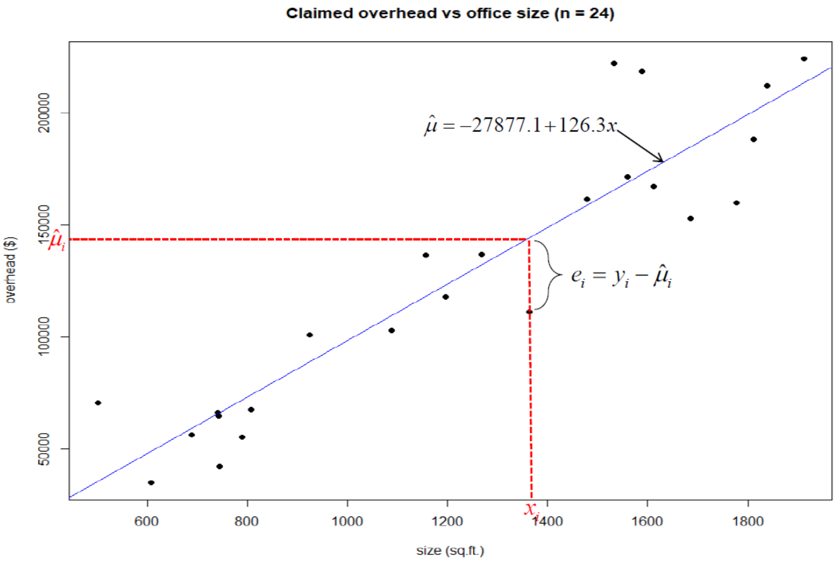
Definition 1.6 The **fitted residual** of the i^{th} observation, e_i, is the difference between the observed response, y_i, and the **fitted value**, $\hat{\mu}_i$, defined as

$$e_i = y_i - \hat{\mu}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

We will see in later sections that much of our statistical analysis from a regression model relies on the calculated value of the **sum of squares of the residuals**, $\sum e_i^2$.

Notes on the residuals:

Figure 1: From the audit data



- Understand the distinction between the residual, e_i, and the error, $e_i = y_i - \hat{\mu}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

  The error is the random variable, on which we impose certain distributional assumptions, we use to model the random variation in the response for a given value of x.

  The residual, $e_i = y - \hat{\mu}_i$, is the difference between the response and the estimated mean response, the values of which we calculate from the fitted line. We can think of the residuals as estimates of the errors.

- By taking the partial derivative with respect to each parameter and setting $= 0$ in our least squares estimation procedure, we have imposed two constraints on our residuals:

  $$\sum e_i = 0$$
  $$\sum x_i e_i = 0$$

  These constraints allow us to compute the remaining two residuals from $n-2$ observations. Thus, we say that the fitted model is associated with $n-2$ **degrees of freedom**.

1.6 Least Squares Estimation of σ^2

Recall the normal model given by

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2) \text{ ind.} \quad i = 1, \dots, n$$

Inference for model parameters requires not only the estimation of β_0 and β_1, but also on the estimation of the error variance, σ^2.

Definition 1.7

In any least squares regression model, this is obtained by dividing the sum of squares of the residuals by the degrees of freedom, giving the least squares estimate as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{n-2}$$

Note that $\hat{\sigma}^2$ is an **unbiased** estimate of σ^2 (i.e. $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$).

The **residual standard error** is the square root of the estimated variance, given by $\hat{\sigma} = \sqrt{\dfrac{\sum e_i^2}{n-2}}$

The residual standard error can be interpreted as the estimated standard deviation of the errors, and is a measure of the random variation in the response for a given value of x. The smaller the value of the residual standard error, the more variation in the response is explained by the relationship with x, and the better the fit of the model.

1.7 Interpretation of Parameter Estimates

Example 1.4

Let $\hat{\mu}_{x_0}$ be the estimated mean response at $x = x_0$, and let $\hat{\mu}_{x_0+1}$ be the estimated mean response at $x = x_0 + 1$.

Then

$$\begin{aligned} \hat{\mu}_{x_0+1} &= \hat{\beta}_0 + \hat{\beta}_1(x_0 + 1) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_1 \\ &= \hat{\mu}_{x_0} + \hat{\beta}_1 \end{aligned}$$

Thus we can see that, in general, $\hat{\beta}_1$ can be interpreted as

The estimated mean change in the response, y, associated with a change of one unit in x.

Example 1.5

Note that for $x = 0$, the estimated mean response reduces to $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1(0) = \hat{\beta}$.

Thus, $\hat{\beta}_0$ may be interpreted in certain situations as the estimated mean value y of at $x = 0$.

However, there is an important caveat: This interpretation may be nonsensical or meaningless in cases where $x = 0$ is not a relevant value, or where $x = 0$ is not in the range of values used in the fit of the model.

This serves as an important reminder:

Never extrapolate results to values of outside the range used to fit the model.

Example 1.6

Recall the least squares estimate of the standard deviation of the errors, σ, called the residual standard error and given by:

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

where $e_i = y_i - \hat{\mu}_i$ are the residuals of the fitted model.

Note that the residual standard error is similar to the (sample) standard deviation of the residuals (only with $n - 2$ degrees of freedom instead of $n - 1$), and is thus **a measure of the variability of the response about the fitted line**. The smaller the residual standard error, the closer the data are to the fitted line, and the better the fit of the model.

Similar to a standard deviation, the residual standard error can be roughly interpreted as a typical or 'standard' distance (or absolute difference) between the response, y_i, and the fitted value, $\hat{\mu}_i$.

1.8 Inference for the Slope Parameter

1.8.1 Confidence Interval

Definition 1.8 A $(1-\alpha)100\%$ confidence interval for β_1 is of the form:

$$\hat{\beta}_1 \pm t_{n-2,\,1-\alpha/2}\,\mathrm{SE}(\hat{\beta}_1)$$

Notes:

- $t_{n-2,\,1-\alpha/2}$ denotes the critical value from a t_{n-2} distribution corresponding to confidence level $(1-\alpha)100\%$. (Be sure you know how to obtain this value for a given confidence level from both R and the posted t-tables)
- $t_{n-2,\,1-\alpha/2}\,\mathrm{SE}(\hat{\beta}_1)$ is called the **margin of error** of the interval. It can be thought of as the bound on the difference between the value of the estimate and the actual (unknown) value of the parameter for the given confidence level.
- it should be obvious, both intuitively and from the form of the confidence interval that
  - the higher the confidence level, the wider the interval
  - the larger the standard error, $\mathrm{SE}(\hat{\beta}_1)$, the wider the interval

Example 1.7 Provide a 95% confidence interval for β_1 from the audit SLR model.

Solution: We need to obtain the interval:

$$\hat{\beta}_1 \pm t_{22,\,0.975}\,\mathrm{SE}(\hat{\beta}_1)$$

These values give us a 95% confidence interval for β_1 of

$$\begin{aligned} &= 126.33 \pm 2.074(10.88) \\ &= 126.33 \pm 22.57 \\ &= (103.76, 148.90) \end{aligned}$$

Example 1.8 How can we interpret this interval in the context of the study?

Solution: We are 95% confident that for every additional increase of one square foot in office size, the mean increase in overhead is between \$103.76 and \$148.90.

Example 1.9 What conclusions can be drawn from our confidence interval about whether a relationship exists between overhead and office size?

Solution: Since $\beta_1 = 0$ is not in the interval, $(103.76, 148.90)$, we conclude that there is a **significant** positive relationship between overhead and office size.

(Had 0 been in the interval, then 0 would be considered a plausible value for and we would thus conclude that there was **no significant relationship** between overhead and office size)

Summary (Hypothesis test for slope parameter β_1)

1. Present the null and alternative hypotheses H_0: $\beta_1 = 0$
2. Calculate the value of the test statistic $\dfrac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} \sim t_{n-2}$, under H_0, $\beta_1 = 0$
3. Obtain p-value $= \mathrm{P}(|T| \geqslant |t|) = 2\,\mathrm{P}(T \geqslant |t|)$, where $T \sim t_{22}$
4. Conclusion in context of study

1.8.2 Types of errors in hypothesis testing

Note that whenever we draw a conclusion from a hypothesis test regarding the significance of the parameter, we could be in error, since we are drawing conclusion based on a probabilistic criteria.

For example, we may reject H_0: $\beta_1 = 0$ and conclude there is a significant relationship, when no relationship exists ($\beta_1 = 0$). Conversely, we may accept H_0: $\beta_1 = 0$ when, in fact, $\beta_1 \neq 0$ and a relationship exists.

The possibility that we could have made one of these errors should always be kept in mind when drawing conclusions from a hypothesis test (as well as from a confidence interval). These two errors are called:

- **Type I error:** Rejecting the null hypothesis when it is true
- **Type II error:** Accepting (i.e. not rejecting the null hypothesis) when it is true.

Note that for any hypothesis test, P(Type I error) = .05 (Convince yourself of this. It will help in your understanding of p-values)

1.8.3 Two-sided vs one-sided tests

By default, we will use a two-sided alternative hypothesis when performing hypothesis tests, since, in most cases, we are concerned with discovering significant relationships in either direction (positive or negative), and have little or no prior reliable knowledge of the possible direction of the relationship. If a one-sided alternative seems appropriate, it will be specified.

Note that hypothesis tests for which a one-sided alternative (e.g. H_a: $\beta_1 > 0$, or H_a: $\beta_1 < 0$) is appropriate yield a p-value $= \mathrm{P}(t_{n-2} > |t|)$, half the p-value that one would obtain with the two-sided alternative, H_a: $\beta_1 \neq 0$.

1.8.4 The relationship between confidence interval and hypothesis tests

Note that the conclusions (i.e. whether not a significant relationship exists) drawn from a 95% confidence interval for will always be consistent with conclusions drawn from a test of H_0: $\beta_1 = 0$

- If the 95% confidence interval contains 0, then a (two-sided) test of H_0: $\beta_1 = 0$ would yield a p-value $\geqslant .05$

- If the 95% confidence interval does not contain 0, then a (two-sided) test of H_0: $\beta_1 = 0$ would yield a p-value $< .05$

2 Multiple Linear Regression

2.1 Multiple Regression Model

Definition 2.1

By extending the SLR model to include p explanatory variables, we obtain the **multiple linear regression model**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \quad i = 1, 2, \ldots, n$$

The multiple regression model can be expressed in matrix form as:

which we write as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

For the normal model,

Definition 2.2

for which we assume $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ ind., we write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \epsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

where $\mathrm{Var}(\epsilon) = \sigma^2\mathbf{I}$ is the **covariance matrix** of the error random vector ϵ with

- $\mathrm{Cov}(\epsilon_i, \epsilon_i) = \mathrm{Var}(\epsilon_i),\ i = 1, \ldots, n$ as the diagonal elements and
- $\mathrm{Cov}(\epsilon_j, \epsilon_k),\ j, k = 1, \ldots, n,\ j \neq k$ on the off-diagonals

Note that expressing the covariance matrix in this way captures both the constant variance assumption ($\mathrm{Var}(\epsilon_i) = \sigma^2$ for all i), and the independence assumption, since $\mathrm{Var}(\epsilon) = \sigma^2\mathbf{I} \rightarrow \mathrm{Cov}(\epsilon_j, \epsilon_k) = 0,\ j \neq k \rightarrow$ independent erroe for $\epsilon_i \sim$ Normal

2.2 Least squares estimation of β

These equations can be expressed in matrix form as:

$$(\mathbf{X}^T\mathbf{X})\hat{\beta} = \mathbf{X}^T\mathbf{y}$$

Solving for $\hat{\beta}$ by multiplying both sides of the equation by $(\mathbf{X}^T\mathbf{X})^{-1}$ yields the least squares estimate:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

(*Note that for $\mathbf{X}^T\mathbf{X}$ to be invertible, $\mathbf{X}$ must be of full rank. That is, all $p+1$ columns of $\mathbf{X}$ must be linearly independent. Otherwise a unique solution will not exist. We will explore this issue in more detail in a later topic.)

Notes:

- The fitted line can be represented by $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 \cdots + \hat{\beta}_p x_p = \mathbf{x}^T\hat{\beta}$, where $\mathbf{x}^T = \{1, x_1, \ldots, x_p\}$
- the vector of fitted values is given by $\hat{\mu} = \mathbf{X}^T\hat{\beta}$
- The residual vector is given by $\mathbf{e} = \mathbf{y} - \hat{\mu}$

2.3 The Hat Matrix

Recall the vector of fitted values, given by $\hat{\mu} = \mathbf{X}^T\hat{\beta}$, where $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Definition 2.3 We can thus express $\hat{\mu}$ as

$$\begin{aligned}\hat{\mu} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{Hy}\end{aligned}$$

where the **hat matrix**, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, maps the vector of response variables to the vector of fitted values.

Properties of $\mathbf{H}$:

- $\mathbf{H}$ is **symmetric** ($\mathbf{H}^T = \mathbf{H}$)
- $\mathbf{H}$ is **idempotent** ($\mathbf{HH} = \mathbf{H}$)

Definition 2.4 Note that the residual vector, e, can be expressed as

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \hat{\mu} \\ &= \mathbf{y} - \mathbf{Hy} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

We can thus express our response vector as

$$\begin{aligned}\mathbf{y} &= \hat{\mu} + \mathbf{e} \\ &= \mathbf{Hy} + (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

It can be shown, based on the symmetric and idempotent properties of $\mathbf{H}$ tells us that the response vector, $\mathbf{y}$, can be decomposed into its two orthogonal elements, the vector of fitted values, $\hat{\mu}$, and the vector of residuals, $\mathbf{e}$. This decomposition forms the basis of ANOVA methods, in which the variation in the response is partitioned into its two components - variation accounted for by the fitted model, and variation not accounted for by the model.

2.4 Least squares estimation of σ^2

Recall the least squares estimate of for the SLR model, given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{n-2}$$

Definition 2.5 The degrees of freedom for a p-explanatory variable multiple regression model with $p+1$ parameters (including the intercept, β_0) is thus $n-(p+1)$, yielding least squares estimate:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-(p+1)}$$

and the residual standard error

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)}}$$

2.5 Least squares vs maximum likelihood estimation

2.5.1 Gauss-Markov theorem and BLUE

We have shown that, for linear regression models with normal errors, the maximum likelihood and least squares methods yield identical estimators for β. How do these two (and other possible) estimation methods compare when the errors are not assumed to be normal?

The **Gauss-Markov theorem** states that the least squares estimator $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is the, 'best linear unbiased estimator' (BLUE) of β.

Theorem

Consider the model given by $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where $\mathrm{E}(\epsilon) = \mathbf{0}$, and $\mathrm{Var}(\epsilon) = \sigma^2\mathbf{I}$.

Among all unbiased linear estimators, $\hat{\beta}^* = \mathbf{M}^*\mathbf{Y}$, the least square estimator, given by $\hat{\beta} = \mathbf{M}Y$, where $\mathbf{M} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, has the smallest variance. That is

$$\mathrm{Var}(\hat{\beta}^*) = \mathrm{Var}(\hat{\beta}) + \sigma^2(\mathbf{M}^* - \mathbf{M})(\mathbf{M}^* - \mathbf{M})^T$$

where $(\mathbf{M}^* - \mathbf{M})(\mathbf{M}^* - \mathbf{M})^T$ is a positive semidefinite matrix (a matrix $\mathbf{A}$ is positive semidefinite if $\mathbf{a}^T\mathbf{A}\mathbf{a} \geqslant 0$ for any vector $\mathbf{a}$).

2.6 Inference: Distribution of Parameter Estimates

2.6.1 Distribution of $\hat{\beta}$

Definition 2.6

The distribution of $\hat{\beta}_j$ is given by

$$\hat{\beta}_j \sim \mathrm{N}(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

and the distribution of $\hat{\beta}_j$, the jth element of $\hat{\beta}$ by

$$\hat{\beta}_j \sim \mathrm{N}(\beta_j, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}) \quad j = 0, 1, 2, \ldots, p$$

where $(\mathbf{X}^T\mathbf{X})^{-1}_{jj})$ represents the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1})$.

We can conclude that $\dfrac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-(p+1)}$.

We also have the following results:

- $\mathrm{Var}\,\hat{\beta}_j = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}$
- $\mathrm{SE}(\hat{\beta}_j) = \hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}$
- $\mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jk} \neq 0$

2.6.2 Interpretation of $\hat{\beta}_j$

Since the parameter estimators are not independent, the value of $\hat{\beta}_j$, the estimate associated with the variable x_j will depend on the other variables in the model.

Example 2.1

The Fitted audit model:

```
> summary(audit.lm)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -198262.24  74354.09  -2.666   0.0157
size             31.26     21.47   1.456   0.1625
age             330.38    502.03   0.658   0.5188
employees      4695.73   5492.21   0.855   0.4038
col          178136.66  69013.51   2.581   0.0188
clients       38.52        33.03   1.166   0.2587
---
Residual standard error: 14430 on 18 degrees of freedom
Multiple R-squared: 0.9566, Adjusted R-squared: 0.9446
F-statistic: 79.41 on 5 and 18 DF, p-value: 1.261e-11
```

Interpretation of $\hat{\beta}_j$: $\hat{\beta}_2 = 330.38$: After accounting for size, # of employees, col, and # of clients, each additional year in the age of the office is associated with an estimated increase in overhead of \$ 330.38.

$\hat{\beta}_j$ can be thus be interpreted as: the estimated mean change in the response associated with a change of one unit in x_j after accounting for the other variables (i.e. while holding all other variables constant).

2.7 Confidence Intervals and Hypothesis Tests for Model Parameters

2.7.1 Confidence intervals for β_j

Definition 2.7

A $(1-\alpha)100\%$ confidence interval for β_j

$$\hat{\beta}_j \pm t_{n-(p+1),\,(1-\alpha)/2}\,\mathrm{SE}(\hat{\beta}_j)$$

Example 2.2

A 95% CI for β_5 (clients coefficient) from the audit model is

$$\begin{aligned}
&\hat{\beta}_5 \pm t_{18,\,0.975}\,\mathrm{SE}(\hat{\beta}_5)\\
&= 38.52 \pm 2.101(33.03)\\
&= 38.52 \pm 69.40\\
&= (-30.88, 107.92)
\end{aligned}$$

Summary (Hypothesis test for slope parameter β_j)

1. Present the null and alternative hypotheses H_0: $\beta_j = 0$
2. Calculate the value of the test statistic $\dfrac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-(p+1)}$, under H_0, $\beta_j = 0$
3. Obtain p-value $= \mathrm{P}(|T| \geqslant |t|) = 2\,\mathrm{P}(T \geqslant |t|)$, where $T \sim t_{22}$
4. Conclusion in context of study

   If p-value $\leqslant 0.05$, reject H_0.

   If p-value > 0.05, do not reject H_0.

   We call .05 the **significance level** of the test.

2.8 Multicollinearity

Consider the matrix of pairwise scatterplots for the audit dataset (using `> plot(audit)` in R):

Take a minute to examine the relationships between the response and expanatory variates (top row) and the relationships among the explanatory variates in the remaining rows. Note especially the relationships among size, employees, and clients.

The scatterplot reveals strong linear assocations (correlations) among some of the explanatory variables - particularly between employees and clients ($r > .99$)

Definition 2.8 When strong (linear) relationships are present among two or more explanatory variables, we say these variables exhibits **multicollinearity**.

Multicollinearity leads to inflated (i.e. increased) variances of the associated parameter estimators, and correspondingly, inflated standard errors. This in turn leads to wide (imprecise) confidence intervals and inaccurate conclusions from hypothesis tests, due to inflated p-values.

To assess the degree of multicollinearity associated with an explanatory variable, x_j:

1. Regress onto all other explanatory variables. That is, we consider x_j to be the response variable, and fit the model with all other explanatory variables.
2. Calculate the **variance inflation factor** (VIF) associated with x_j:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

   where R_j^2 is the **coefficient of determination** (Multiple R-squared in R) of the model fit with x_j as the response variable.

VIF can be interpreted as the factor by which the variance of $\hat{\beta}_j$ is increased, through the multicollinearity among x_j and the other explanatory variables, relative to the case in which all explanatory variables are uncorrelated.

One simple solution is to remove x_j from the model if $\mathrm{VIF}_j > 10$. Note that this corresponds to $R_j^2 > .90$. This is a general rule of thumb - some references consider $\mathrm{VIF}_j > 5$ to be cause for concern, depending on the context.

2.8.1 Confidence Intervals for a Response Mean and Prediction Intervals for a Response

Once we have fit the model to our data, we may wish to use the fitted model to estimate the mean response, or predict the value of the response, of a new unit in the population (one that was not used in the fit of the model).

For example, after fitting the audit model to the 24 offices, an auditor may wish to use the model on other offices in the population to assess the consistency of their claimed overhead with the office attributes (size, age, ...). This, in fact, would likely be the main objective for fitting a regression model to the audit data.

Consider the following questions an auditor might wish to address about a new office from the population that is 1000 ft, 12 years old, with 1300 clients and a cost of living index of 1.02:

1. What is the estimated mean overhead for all offices with these attributes in the population?
2. What is the predicted overhead of this office?

Definition 2.9

The distribution of $\hat{\mu}_{\text{new}}$ gives us a $(1-\alpha)100\%$ confidence interval for μ_{new}

$$\mu_{\text{new}} \pm t_{n-(p+1),\,1-\alpha/2}\,\hat{\sigma}\sqrt{\mathbf{x}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{\text{new}}}$$

Example 2.3

Provide a confidence interval for the mean overhead for offices in the population that are 1000 , 12 years old, with 1300 clients and a cost of living index of 1.02.

```
> new_x=data.frame(size=1000,age=12,col=1.02,clients=1300)
> predict(audit2.lm,new_x,interval='confidence',level=.95)
     fit      lwr      upr
104831.2 97460.07 112202.3
```

Interpretation: We can be 95% confident that the mean overhead for offices in the population with these characteristics is between \$97,460 and \$112,202.

Definition 2.10

A $(1-\alpha)100\%$ prediction interval for y_{new} of the form

$$\hat{y}_{\text{new}} \pm t_{n-(p+1),\,1-\alpha/2}\,\hat{\sigma}\sqrt{1+\mathbf{x}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{\text{new}}}$$

Example 2.4

Provide a prediction interval for an office in the population that is 1000 , 12 years old, with 1300 clients and a cost of living index of 1.02.

```
> predict(audit2.lm,new_x,interval='prediction',level=.95)
     fit      lwr      upr
104831.2 73946.72 135715.7
```

Interpretation: We predict with 95% confidence that the overhead for this office is between \$73,946 and \$135,715.

Note that for the SLR model, $\hat{\sigma}\sqrt{\mathbf{x}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{\text{new}}}$ and $\hat{\sigma}\sqrt{1+\mathbf{x}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{\text{new}}}$ reduce to

$$\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_{\text{new}}-\bar{x})^2}{S_{xx}}} \text{ and } \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{\text{new}}-\bar{x})^2}{S_{xx}}}$$

2.9 Modelling Categorical Explanatory Variables

2.10 Inference for parameters associated with indicator variables

Example 2.5

Is there a difference in mean sales between stores that used promotion A and stores that used no promotion?

Recall the output from the model fit to the promotion data:

```
    Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.870 1.665 -0.523 0.60552
PromoA 8.350 2.354 3.547 0.00145
PromoB 2.970 2.354 1.261 0.21792
```

We can answer this question using a confidence interval or hypothesis test for in the same manner as learned previously:

H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$, $t = 3.547$, p-value $= 0.00145$.

Since p-value < 0.05, we reject H_0, and conclude that stores using promotion A had significantly higher sales than stores using no promotion.

We could also calculate a 95% confidence interval for β_1 to reach the same conclusion. Similarly, we conclude that there was no significant difference in mean sales between stores using promotion B and stores using no promotion (p-value $= 0.21792$).

Example 2.6

Is there a difference in mean sales between promotion A and promotion B stores?

Since no difference in mean sales between these two promotion types $\beta_1 - \beta_2 = 0$, we can test for a difference with the null hypothesis H_0: $\beta_1 - \beta_2 = 0$.

2.11 Orthogonal X matrix designs

In certain situations, we may wish to model categorical variables in a way that creates an orthogonal $\mathbf{X}$ matrix, thereby producing independent parameter estimators.

Consider the data below from a factorial experiment designed to investigate the effect of high and low levels of three factors - mold temperature, holding pressure, and screw speed - on the shrinkage (%) of parts produced in a molding process.

We can illustrate these concepts by considering the fitted shrinkage model:

```
> Shrinkage<-c(19.7,19.1,20.0,19.5,15.9,15.3,25.5,24.9)
> Temp<-c(rep(c(-1,1),4)) #yields Temp = ( -1 1 -1 1 -1 1 -1 1)
> Pres<-c(rep(c(-1,-1,1,1),2))#yields Pres = ( -1 -1 1 1 -1 -1 1 1)
> Speed<-c(rep(-1,4),rep(1,4))#yields Speed = ( -1 -1 -1 -1 1 1 1 1)
> shrink.lm.out<-lm(Shrinkage~Temp+Pres+Speed)
> summary(shrink.lm.out)
```

Note also:

- All estimates share the same standard error, since $\text{SE}(\hat{\beta}_j) = \hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}} = \hat{\sigma}\sqrt{1/8}$
- For factors coded in this way, $\hat{\beta}_0$ no longer has a relevant interpretation.

Note that since the estimators are independent, removing Temp and Pres have no effect on the remaining estimates.

Note also the decrease in the residual standard error, and the associated decrease in the standard errors and p-values of the remaining parameter estimates.

This is a result of the increase in the degrees of freedom obtained by removing variables from the model that do not account for a significant amount of variation in the response. We will discuss this concept in more detail in a future lesson.

2.12 Analysis of Variance (ANOVA)

Recall the (sample) variance of a set of observations, $\{y_1, y_2, \ldots, y_n\}$, given by

$$s^2 = \frac{\sum(y_i - \overline{y})^2}{n-1}$$

where we can think of the **total sum of squares**, $\text{SS}(Tot) = \sum(y_i = \overline{y})^2$, as representing the total variation in y.

Definition 2.11

In regression modelling, we partition this total variation in the response into two components – the variation in the response explained by the model variables, and the variation left unexplained.

We can express this partition algebraically as

$$\sum_{i=1}^{n}(y - \overline{y})^2 = \sum(\hat{y}_i - \overline{y})^2 + \sum(y_i - \hat{\mu}_i)^2$$
$$\text{SS}(Tot) = \text{SS}(Reg) + \text{SS}(Res)$$

where the **regression sum of squares**, $\text{SS}(Reg)$, is the variation explained by the model, and the **residual sum of squares**, $\text{SS}(Res)$, is the variation in the response left unexplained (i.e., not accounted for by the model variables).

In **ANOVA** (**AN**alysis **O**f **VA**riance) methods of inferenece, we draw conclusions about the relative fit of a model or models by comparing these two sources of variation. The greater the variation explained by the model relative to the variation unexplained, the better the fit of the model.

2.12.1 Coefficient of Determination

By partitioning the total variation in the response into its two component sources of variation, as described by the relationship $\mathrm{SS}(Tot) = \mathrm{SS}(Reg) + \mathrm{SS}(Res)$, wee see that the ratio $\dfrac{\mathrm{SS}(Reg)}{\mathrm{SS}(Tot)}$, or, equivalently, $1 - \dfrac{\mathrm{SS}(Res)}{\mathrm{SS}(Tot)}$, measures the proportion of the variation in the response explained by the model.

Definition 2.12

We call this measure the **coefficient of determination**, or more simply, the (multiple) R- squared, and denote it by

$$R^2 = 1 - \frac{\mathrm{SS}(Res)}{\mathrm{SS}(Tot)}$$

Example 2.7

For the fitted audit model below (fit now without # of employees due to multicollinearity issues).

An R-squared value of 0.9549 tells us that over 95% of the variation in overhead is explained by the office's size, age, col and #of clients.

2.12.2 F-test and the ANOVA table

Example 2.8

Is there a relationship between overhead and at least one of size, age, col, or number of clients?

We can test the hypothesis:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_a: at least one of $\beta_j \neq 0$, $j = 1, 2, 3, \ldots, p$

with a test statistic that compares the relative magnitudes of the variation explained by the model, $\mathrm{SS}(Reg)$, and the variation left unexplained, $\mathrm{SS}(Res)$.

Definition 2.13

This test statistic takes the form

$$F = \frac{\mathrm{SS}(Reg)/p}{\mathrm{SS}(Res)/(n-(p+1))} = \frac{\mathrm{MS}(Reg)}{\mathrm{MS}(Res)}$$

where, under H_0, F has an F **distribution** on p, with $n - (p+1)$ degrees of freedom. $\mathrm{MS}(Reg)$ and $\mathrm{MS}(Res)$ are **mean squared** values, obtained by dividing the sum of squares by their respective degrees of freedom.

Example 2.9

Is there a (linear) relationship between overhead and at least one of size, age, col, or # of clients?

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_a: at least one of $\beta_j \neq 0$, $j = 1, 2, 3, \dots, p$

$$F = 100.5$$
$$p\text{-value} = \text{P}(F_{4,19} > 100.5) = 1.661 \times 10^{-12}$$

Reject H_0. At least one of size, age, col, # of clients is significantly related to overhead.

The test of H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ is often summarized in an **ANOVA table**, that shows not only the F test statistic and p-value, but also shows the breakdown of the sum and squares (and mean squares) of the two sources of variation.

Source	df	SS	MS	F	p-value
Regression	p	$\text{SS}(Reg)$	$\frac{\text{SS}(Reg)}{p}$	$\frac{\text{SS}(Reg)}{\text{SS}(Res)}$	$\text{P}(F_{p,n-(p+1)} > F)$
Residual	$n-(p+1)$	$\text{SS}(Res)$	$\text{SS}(Res)/(n-(p+1))$		
Total	$n-1$	$\text{SS}(Tot)$			

Example 2.10

For the audit model, confirm the value of the test statistic, $F = 100.5$, and complete the ANOVA table from values in the summary output.

Solution: We can obtain $\text{SS}(Res)$ from the residual standard error

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-(p+1)}} = \sqrt{\frac{\text{SS}(Res)}{n-(p+1)}}$$
$$\text{SS}(Res) = \hat{\sigma}^2(n-(p+1)) = 14300^2(19) = 3901629100$$

$$R^2 = 1 - \frac{\text{SS}(Res)}{\text{SS}(Tot)} = .9549 \Rightarrow \text{SS}(Tot) = \frac{\text{SS}(Res)}{1-R^2} = \frac{3901629100}{1-.9549} = 86510623060$$

Source	**df**	**SS**	**MS**	**F**	**p-value**
Regression	4	82608992960	20652248490	100.6	1.66×10^{-12}
Residual	19	3901629100	205348900		
Total	23	86510623060			

2.13 Additional Sum of Squares

Definition 2.14

To determine the better model, we assess the difference in the variation explained by the full and reduced models, expressed as $\text{SS}(Res)_{\text{full}} - \text{SS}(Reg)_{\text{red}}$, or equivalently, by

$$\text{SS}(Res)_{\text{red}} - \text{SS}(Reg)_{\text{full}}$$

We call this difference in variation between the two models the **additional sum of squares**.

We test H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ with the F statistic:

$$F = \frac{(\mathrm{SS}(Res)_{\mathrm{red}} - \mathrm{SS}(Res)_{\mathrm{full}})/(df_{\mathrm{red}} - df_{\mathrm{full}})}{\mathrm{SS}(Res)_{\mathrm{full}}/df_{\mathrm{full}}}$$

where under H_0, $F \sim F_{df_{\mathrm{red}} - df_{\mathrm{full}},\ df_{\mathrm{full}}}$

Example 2.11

After accounting for col and clients, does either size or age account for significant variation in overhead?

H_0: $\beta_1 = \beta_2 = 0$
H_a: at least one of $\beta_1, \beta_2 \neq 0$

Full model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$
Reduced model: $Y = \beta_0 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$

As an exercise, we can calculate the test statistic by obtaining the residual sum of squares of the full and reduced models from the output, in the same way we did in creating the ANOVA table.

$$\mathrm{SS}(Res)_{\mathrm{full}} = \hat{\sigma}^2_{\mathrm{full}}(df_{\mathrm{full}}) = 14330^2(19)$$
$$\mathrm{SS}(Res)_{\mathrm{red}} = \hat{\sigma}^2_{\mathrm{red}}(df_{\mathrm{red}}) = 15360^2(21)$$

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.594e+05  7.154e+04  -2.227   0.0370 *
col          1.484e+05  6.989e+04   2.124   0.0457 *
clients      8.774e+01  4.734e+00  18.532 1.71e-14 ***
Residual standard error: 15360 on 21 degrees of freedom
```

$$F = \frac{(\mathrm{SS}(Res)_{\mathrm{red}} - \mathrm{SS}(Reg)_{\mathrm{full}})/(df_{\mathrm{red}} - df_{\mathrm{full}})}{\mathrm{SS}(Res_{\mathrm{full}})/df_{\mathrm{full}}} = \frac{(15360^2(21) - 14330^2(19))/2}{14330^2} = 2.564$$

Note that $\mathrm{SS}(Res)_{\mathrm{full}} = \mathrm{SS}(Res)_{\mathrm{full}}/df_{\mathrm{full}} = \hat{\sigma}^2_{\mathrm{full}}$

From the F table, $\mathrm{P}(F_{2,19} > 3.52) = 0.05 \Rightarrow p\text{-value} = \mathrm{P}(F_{2,19} > 2.56) > .05$. Or using R

```
> 1-pf(2.564,2,19)
[1] 0.1033253
```

Since p-value > 0.05, we do not reject H_0. The reduced model is preferred. More specifically, age and size together do not account for significant additional variation in overhead after accounting for col and clients, so we do not need them in the model.

We can verify these results using the anova function in R:

```
> anova(audit.red.lm,audit.full.lm)
Analysis of Variance Table
Model 1: overhead ~ col + clients
Model 2: overhead ~ size + age + col + clients
  Res.Df        RSS Df Sum of Sq F      Pr(>F)
1     21 4954374034
2     19 3901347198  2 1.053e+09 2.5642 0.1033
```

2.13.1 Additional Sum of Squares Test and Categorical Variables

Example 2.12 Is there a difference in mean sales between promotion A and promotion B?

By considering a test of

H$_0$: $\beta_1 - \beta_2 = 0$
H$_a$: $\beta_1 - \beta_2 \neq 0$ with test statistic

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{\text{SE}(\hat{\beta}_1 - \hat{\beta}_2)}$$

Alternatively, we can test H$_0$ using the additional sum of squares test statistic

$$F = \frac{(\text{SS}(Res)_{\text{red}} - \text{SS}(Reg)_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\text{SS}(Res_{\text{full}})/df_{\text{full}}}$$

To do so, we need to fit the reduced model under H$_0$: $\beta_1 - \beta_2 = 0$ (i.e. under restriction that $\beta_1 = \beta_2 = \beta^*$), given by

$$\begin{aligned} Y &= \beta_0 + \beta^* x_1 + \beta^* x_2 + \epsilon \\ &= \beta_0 + \beta^*(x_1 + x_2) + \epsilon \\ &= \beta_0 + \beta^* x^* + \epsilon \end{aligned}$$

where

$$x^* = \begin{cases} 1 & \text{if a promotion (wither } A \text{ or } B\text{) is used} \\ 0 & \text{otherwise} \end{cases}$$

Defining x^* and fitting the reduced model in R:

```
> x_promo=x1+x2
> x_promo
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0

          Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.870      1.786  -0.487   0.6299
x_promo      5.660      2.187.  2.588.  0.0151
Residual standard error: 5.647 on 28 degrees of freedom
```

Example 2.13 Is there a difference in mean sales between promotion A and promotion B?

Testing H$_0$: $\beta_1 = \beta_2 = 0$ using additional sum of squares:

```
> anova(promo_red.lm,promo.lm)
Analysis of Variance Table
Model 1: Promo_sales ~ x_red
Model 2: Promo_sales ~ type
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     28  893.02
2     27  748.30       144.72 5.2218 0.03038
```

We reject H_0 and conclude that that mean sales associated with promotion A is significantly higher than the mean sales for promotion B.

2.13.2 Additional Sum of Squares Test and ANOVA

Example 2.14

Test of H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ Revisited

Is there a (linear) relationship between overhead and at least one of size, age, col, or # of clients?

Recall in the lesson on ANOVA we addressed this question with the hypotheses (Example 2.9) and test statistic

$$F = \frac{\mathrm{SS}(Reg)/p}{\mathrm{SS}(Res)/(n-(p+1))} = \frac{\mathrm{MS}(Reg)}{\mathrm{MS}(Res)}$$

This is just another example of an additional sum of squares test for which the reduced model is

$$Y = \beta_o + \epsilon \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

This result is consistent with out intuitive understanding of the fitted model. With no explanatory variables in the model, $\hat{\mu} = \hat{\beta}_0 = \overline{y}$. (The sample mean, $\overline{y}$, is the least squares estimate of μ).

Then the additional sum of squares test statistic reduces to:

$$\begin{aligned} F &= \frac{(\mathrm{SS}(Res)_{\mathrm{red}} - \mathrm{SS}(Reg)_{\mathrm{full}})/(df_{\mathrm{red}} - df_{\mathrm{full}})}{\mathrm{SS}(Res_{\mathrm{full}})/df_{\mathrm{full}}} \\ &= \frac{\mathrm{MS}(Reg)}{\mathrm{MS}(Res)} \end{aligned}$$

2.13.3 Additional Sum of Squares Test for Individual Parameters

Example 2.15

Test of H_0: $\beta_j = 0$ Revisited

After accounting for size, col, and clients, is age related to overhead?

Recall in an earlier lesson we addressed this question with the hypothesis H_0: $\beta_j = 0$ and the test statistic $t = \dfrac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}$, for $j-2$ (the age parameter), the value of which and associated p-value can be obtained directly from the summary output

```
Placeholder...
```

An equivalent test of H_0: $\beta_j = 0$ that yields an identical p-value can be obtained with the additional sum of squares test statistic

$$\begin{aligned} F &= \frac{(\mathrm{SS}(Res)_{\mathrm{red}} - \mathrm{SS}(Reg)_{\mathrm{full}})/(df_{\mathrm{red}} - df_{\mathrm{full}})}{\mathrm{SS}(Res_{\mathrm{full}})/df_{\mathrm{full}}} \\ &= \frac{\mathrm{SS}(Res)_{\mathrm{red}} - \mathrm{SS}(Res)_{\mathrm{full}}}{\mathrm{MS}(Res)_{\mathrm{full}}} \end{aligned}$$

Note that the degrees of freedom of the numerator, $df_{\mathrm{res}} - df_{\mathrm{full}}$, is one, since there is only one restriction imposed on the model by the null hypothesis. That is, the full model has only one more parameter than the reduced model.

To illustrate, we can perform an additional sum of squares test on the full and reduced audit models associated with H_0: $\beta_2 = 0$:

```
> anova(audit_minus_age.lm,audit.lm)
Analysis of Variance Table
Model 1: overhead ~ size + col + clients
Model 2: overhead ~ size + age + col + clients
  Res.Df        RSS  Df  Sum of Sq      F Pr(>F)
1     20 4037157770
2     19 3901347198   1  135810572 0.6614 0.4261
```

Note the equivalent p-value (0.4261) associated with the F and t test statistics.

Note also the relationship between the values of and . We see that $F = 0.661 = 0.813^2 = t^2$.

Note also that the relationship between F and t test statistics holds for any hypothesis test with a single restriction (e.g. H_0: $\beta_j = 0$, H_0: $\beta_j - \beta_k = 0$)

In general, for v degrees of freedom,

$$\mathrm{P}(|t|_v > |t|) = \mathrm{P}(F_{1,v} > t^2)$$

2.13.4 The General Linear Hypothesis

Consider the hypotheses we have tested so far in lessons using additional sum of squares:

1. H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
2. H_0: $\beta_1 = \beta_2 = 0$
3. H_0: $\beta_1 - \beta_2 = 0$
4. H_0: $\beta_2 = 0$

Definition 2.15

These hypotheses all test linear combinations of the model parameters. As such, they can all be expressed in the form of the general linear hypothesis:

$$\mathrm{H}_0 : \ \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$$

where $\mathbf{A}$ is an $l \times (p+1)$ matrix tat imposes the l linear constraints on the full model as described by H_0.

Result: Consider the full normal model given by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, and corresponding reduced model associated with the set of linear hypotheses of the form H_0: $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$. Under H_0,

$$F = \frac{(\mathrm{SS}(Res)_{\mathrm{red}} - \mathrm{SS}(Reg)_{\mathrm{full}})/(df_{\mathrm{red}} - df_{\mathrm{full}})}{\mathrm{SS}(Res)_{\mathrm{full}}/df_{\mathrm{full}}} \sim F_{df_{\mathrm{red}} - df_{\mathrm{full}},\ df_{\mathrm{full}}}$$

That is, the additional sum of squares test statistic can be used to test any set of linear hypotheses of the form H_0: $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$.

2.14 Assessing Model Adequacy (Residual Analysis)

Consider the assumptions of the normal model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2\mathbf{I})$$

- the functional form of the relationship between the response and the explanatory variables, is correctly specified by the deterministic component of the model. For the linear model, this relationship is described by $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$
- errors follow a normal distribution
- errors have a constant variance, denoted by σ^2 (this property is sometimes refereed to as **homoskedasticity**)
- the errors are independent

(Note: the last two assumptions are described byVar $\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$)

Distribution of the parameter estimators and subsequent methods of inference were derived based on these assumptions. If the assumptions do not hold, the model is inadequate, and any conclusions drawn from the fit of the model are inaccurate and meaningless.

We can assess model adequacy (i.e. the validity of the model assumptions) through examination and analysis of the fitted residuals, $e = y - \hat{\mu}$, to ensure that the behaviour of these residuals is consistent with the assumptions of the model.

The best way to examine the residuals for evidence of departures from the model assumptions is through residual plots.

Residual Plots There are many different types of residual plots that may be used to assess model adequacy. We will introduce only two of the more common ones here.

Plot of the residuals, e_i, vs the fitted values, $\hat{\mu}_i$

The most common and useful diagnostic plot for assessing model assumptions, is a plot of e_i vs $\hat{\mu}_i$. It can be shown that, if the model assumptions hold, e_i and $\hat{\mu}_i$, are uncorrelated.

Thus, if the model is adequate, we would expect to see **no observable pattern** in a plot of e_i vs $\hat{\mu}_i$. The plot should only exhibit random scatter, as illustrated in the plots below from the fit of a SLR model.

If the assumptions do not hold, we would expect to see a pattern or relationship consistent with the assumption that has been violated. Two examples are provided on the following slide.

QQ Plots QQ (Quantile-Quantile) plots are used to assess the assumption of normal errors. (This assumption is more critical for very small datasets than for relatively large ones. Why?)

In normal QQ plots, the ordered residuals ('sample Quantiles' in R plot), $e_{(i)}$, are plotted vs the expected ordered values (Theoreticall Quantiles'), $\mathrm{E}(Z_{(i)})$, where $Z_i \sim \mathrm{N}(0,1)$.

If the residuals are from a normal distribution, then $e_{(i)}$ should be proportional to $\mathrm{E}(Z_{(i)})$. Thus a straight line relationship is an indication that the assumption of normal errors has been well met.

The QQ plots below provide an example of a fitted model that meets the assumption of normal errors (left) and a model that does not (right).

(Note: we have discussed the use of residual plots to assess the assumptions of correct functional form, constant variance of the errors, and normality of the errors. We will consider plots that address the assumption of independence of the errors when we discuss time series data)

Variance Stabilizing Transformations There are several approaches available that may be used to attempt to address model inadequacies revealed in the residual plots.

Often, a transformation of the response (and possibly one or more of the explanatory variables) is sufficient to improve the adequacy of the model in terms of the model assumptions.

We call such transformations **variance stabilizing transformations**, since they address violation of the constant variance assumption, in addition to the assumptions of model misspecification and non- normality of the errors.

Example 2.16

Examples of common transformations include

- $\log y$ (natural log transformation)
- $y^{\frac{1}{2}}$ (square root transformation)
- y^{-1} (reciprocal transformation)

There transformations are particularly useful when the error variance, σ^2, is a function of the mean response, μ.

For example, it can be shown that, when the standard deviation is proportional to the mean, the log transformation is most appropriate, whereas a square root transformation is more suitable in cases where the variance is proportional to the mean.

Other possible approaches to addressing model inadequacies include:

- the addition of higher order terms (e.g. x^2) in one or more of the explanatory variables.
- the inclusion of an interaction term

2.14.1 Residual Analysis - audit model

Consider again the audit model output

The large R-squared value and relatively small p-values associated with most of the variables suggest that we have a very good fit, **providing the model assumptions are valid.**

There is nothing in the output that provides any information on the validity of the assumptions. Whenever we fit a model, we must always perform a residual analysis to ensure that the assumptions have been met and the model is adequate. If the model is not adequate, any conclusions we draw from the fit of the model are meaningless.

We will begin by examining a plot of the residuals vs the fitted values.

```
> plot(fitted(audit.lm),residuals(audit.lm),
        pch=19,xlab='fitted',ylab='residuals')
```

Note that by taking an appropriate transformation to address issues with the model assumptions, we have also arrived at a better fitting model, as evidenced by the resulting increase in the R-squared value and decrease in p-values seen in the output below:

```
> audit.sqrt.lm=lm(sqrt(overhead)~size+age+col+clients)
> summary(audit.sqrt.lm)
```

Notes:

- Remember that the response is now in units of $\sqrt{\text{dollars}}$. When taking a transformation of the response, we need to back transform estimated mean values, $\hat{\mu}$, and associated confidence and prediction intervals to the original units for interpretation purposes.
- A log transformation was also applied, but did not adequately address problems with model assumptions.

Properties of the Residuals Before we continue in our analysis of the residuals to assess model adequacy, it is important that we understand certain properties of the residuals. We will attempt to do so here.

We begin by examining the relationship between the fitted residuals and the model errors. Note that we can express the residual vector, $\mathbf{e}$, as a function of the error vector, $\boldsymbol{\epsilon}$:

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \hat{\boldsymbol{\mu}} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}\end{aligned}$$

Note that $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$. However, this assumes $(\mathbf{I} - \mathbf{H})$ in invertible and is therefore of full rank, with n linearly independent columns and rows. This is not the case. It can be shown that $\text{rank}(\mathbf{I} - \mathbf{H}) = n - (p + 1)$. $(\mathbf{I} - \mathbf{H})$, therefore is not invertible.

2.14.2 Distribution of e

Now that we have established $\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. We can derive

$$\mathrm{E}(\mathbf{e}) = \mathrm{E}\left[(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}\right] = (\mathbf{I} - \mathbf{H})\,\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$$

and

$$\begin{aligned}\mathrm{Var}(\mathbf{e}) &= (\mathbf{I} - \mathbf{H})\,\mathrm{Var}(\boldsymbol{\epsilon})(\mathbf{I} - \mathbf{H})^T \\ &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \quad (\mathbf{H} \text{ is symmetric, idempotent})\end{aligned}$$

Definition 2.16

Thus

$$\begin{aligned}\mathbf{e} &\sim \mathrm{N}(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})) \\ \rightarrow e_i &\sim \mathrm{N}(0, \sigma^2(1 - h_{ii}))\end{aligned}$$

where h_{ii} is the i^{th} diagonal element of $\mathbf{H}$, $i = 1, 2, \ldots, n$.

Note in particular:

- $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$ (Residuals have non-constant variance)
- $\mathrm{Cov}(e_j, e_k) = -\sigma^2 h_{jk} \quad j \neq k$ (Residuals are not independent. This is a consequence of the constraint, $\sum e_i = 0$, placed on the residuals in least squares estimation)

2.14.3 Studentized Residuals

Recall that we **standardized** a random variable by subtracting the mean and dividing by the standard deviation.

For example, in the case of a normal random variable X,

$$X \sim \mathrm{N}(\mu, \sigma^2) \rightarrow Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1)$$

where Z is the standardized normal random variable.

Definition 2.17

Similarly, we **studentize** a random variable by subtracting the mean and dividing by the estimate of the standard deviation.

In a previous statistics course we studentized the sample mean of a normal random variable

$$\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n}) \rightarrow t = \frac{\overline{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

In this course, we studentized the parameter estimates

$$\hat{\beta}_j \sim \mathrm{N}(\beta_j, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{ii}) \rightarrow t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} = \frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)}$$

Note that in both cases, the resulting studentized random variable follow a distribution.

In the same way, we can studentized the residuals from the fit of a normal regression model, for which we have previously shown that $e_i \sim \mathrm{N}(0, \sigma^2(1 - h_{ii}))$.

Definition 2.18

A **studentized residual** associated with the i^{th} observation and denoted by d_i, is defined as

$$d_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where distribution of d_i can be reasonably approximated by a N(0, 1) distribution for large n.

Note that by studentizing the residuals, we have insured that the resulting residuals, d_i, will have a constant (estimated) variance = 1. For this reason, the studentized residuals, d_i, are often used instead of the fitted residuals, e_i, in residual plots.

2.14.4 Extreme Values of the Response ('Outliers')

Definition 2.19

An **outlier** can be loosely defined in general as any observation which is extreme (either extremely large or extremely small) relative to the other observations.

Recall that, for large n, the distribution of can be reasonably approximated by a N(0, 1) distribution. based on our understanding of normal probability theory, we know that approx. 99% of all observations will be within ±2.5. Anything within this range is acceptable variation.

Summary

Thus, as a general rule of thumb, an observation may be considered an outlier in the response if

$$|d_i| > 2.5$$

or so. Here, the studentized residual in question is greater than 3, suggesting that the associated response is a moderate outlier.

Addressing Outliers in the Response Once an outlier is detected, the associated observation should be investigated for a possible cause. Causes of outliers may include:

- typos, misrecording of data
- values of associated potential explanatory variables not included in the model
- random variability

Deciding how to deal with outliers will depend on the cause, and should be dealt with on a case-by- case basis. It is never a good idea to remove an observation deemed to be an outlier from the fit of the model without further investigation.

2.14.5 Leverage and Influential Observations

Leverage is a measure used to identify those observations whose set of explanatory variables is extreme relative to the sets of explanatory variables of the other observations.

Definition 2.20 The leverage of the i^{th} observation in a dataset is defined as the i^{th} diagonal element of the hat matrix. denoted by h_{ii}. It is a function of the distance between the point $(x_{i1}, x_{i2}, \ldots, x_{ip})$, and the centroid, $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_p)$, of the sets of explanatory variables of the dataset.

To see how we can use leverage to identify extreme values in the sets of explanatory variables, we consider the SLR case. It can be shown that leverage can be expressed as

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum (s_i - \overline{x})^2}$$

Note that the more extreme the value of the explanatory variable, x_i, relative to the mean, $\overline{x}$, the larger the leverage.

Recall also that $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y} \Rightarrow \hat{\mu}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_i$. The leverage, h_{ii}, can be therefore be thought of as the weight of the contribution of y_i to the fitted value, $\hat{\mu}_i$. The larger the leverage relative to the other observations, the more y_i contributed to the fit of the line.

Leverage properties:

- $\frac{1}{n} \leqslant h_{ii} \leqslant 1$
- $\sum h_{ii} = tr(\mathbf{H}) = rank(\mathbf{H}) = rank(\mathbf{X}) = p + 1$

Finally, note that, as $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, **$h_{ii}$ is a function only of the explanatory variables**, not of the response.

Identifying High Leverage Cases

Unlike in assessing model assumptions, plots of the residuals, e_i, are not useful in revealing high leverage points. To see this, recall the distribution of the residuals we derived in a previous lesson, where $e_i \sim \mathrm{N}(0, \sigma^2(1 - h_{ii}))$. Note that as $h_{ii} \to 1, \mathrm{Var}(e_i) \to 0$. Consequently, the residuals of high leverage observations will tend to be close to zero.

Instead, we can plot the leverage values (‘hatvalues’ in R).

Summary

As of a rough general rule, an observation is considered to have high leverage if

$$h_{ii} > 2\bar{h} = \frac{2(p+1)}{n}$$

A plot of the hatvalues for the audit model (with square root transformation of y) is shown below.

```
> plot(hatvalues(audit.sqrt.lm),cex.lab=1.3,cex.axis=1.3,cex=1.3,pch=19)
```

Influential Observations

Definition 2.21 An observation is considered **influential** if its removal from the fit of the line changes the fitted line (i.e. changes the parameter estimates) considerably.

Only high leverage cases have the **potential** to be influential. Whereas leverage depends only on the explanatory variables, the influence of an observation also depends on the value of the response, as illustrated below.

In the above plots, the fitted regression line with the leverage point included is given by the solid black line and line fit with the leverage point omitted is given by the dotted red line.

Note that in the plot on the left, the removal of the high leverage point does not dramatically alter the fitted line, whereas in the plot on the right, omission of the leverage point alters the fitted line considerably.

Thus the high leverage observation seen in both plots is not influential in the scenario on the left, but is influential in the scenario on the right.

Identifying Influential Observations

As defined previously, an observation will be influential if its removal from the fit of the line changes the fitted line considerably. The larger the influence of an observation i, the larger the distance, $(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_{(i)})^T(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_{(i)})$, between the vector of fitted values, $\hat{\boldsymbol{\mu}}$, and the vector of fitted values with the i^{th} observation omiited, $\hat{\boldsymbol{\mu}}_{(i)}$.

Definition 2.22 **Cook’s distance** is one common measure of influence that is a function of this distance, defined as

$$D_i = \frac{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_{(i)})^T(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_{(i)})}{\hat{\sigma}^2(p+1)}$$

where $\hat{\sigma}^2$ is the estimate of the variance from the model fit with the the i^{th} observation included.

This would seem to imply that to measure the influence for each observation, we need to fit the model both without and without the i^{th} observations for all i.

However, it can be shown that Cook's distance can be expressed in the form

$$D_i = \frac{h_{ii}}{1 - h_{ii}} \cdot \frac{d_i^2}{p + 1}$$

which can be calculated from the fit of model with all observations included.

Note that to be influential, an observation must have both a relatively high leverage, h_{ii} and large (absolute) studentized residual, d_i.

As a general rule of thumb, $D_i \geqslant 1$ suggests a strongly influential observation.

```
> max(cooks.distance(audit.sqrt.lm))
[1] 0.2202189
```

2.15 Model Selection – Introduction

To understand why removing certain variables might yield a 'better' model, consider the expression for the residual standard error, given by

$$\hat{\sigma} = \sqrt{\frac{\mathrm{SS}(Res)}{n - (p + 1)}}$$

Note that removing one or more variables from the model will increase **both** the $\mathrm{SS}(Res)$ and the degrees of freedom.

If the increase in $\mathrm{SS}(Res)$ resulting from the removal of variables is small **relative to the degrees of freedom gained**, as is often the case for variables associated with large p-values, then $\hat{\sigma}$ will decrease, resulting in smaller standard errors and a more precise model.

If however, the increase in degrees of freedom obtained from removing one or more variables is not sufficient to counter-balance the associated increase in $\mathrm{SS}(Res)$, then $\hat{\sigma}$ will increase, and we will have a less precise model.

Note that we cannot simply remove all the variables that are associated with large p-values, since the removal of any one variable will change the estimates and associated p-values of all the remaining variables. A variable that is associated with a large p-value may be associated with a small p-value once a variable is removed.

2.15.1 Model Selection Methods

Definition 2.23

Iterative methods of model selection involve building a model by adding or removing variables one at a time, and refitting the model at each iteration until no more variables can be added or removed.

1. Backward elimination
   - Fit all p variables
   - Remove the variable with the largest p-value that is greater than some predetermined threshold value, α, (e.g., $\alpha = .10$)

- Refit the model the with remaining $p-1$ variables
- Continue removing one variable at each iteration of the above steps until no more variables can be removed (all p-values $< \alpha$)

2. Forward Selection
   - Fit all p single variable (i.e. SLR) models
   - Select the variable associated with the smallest p-value $< \alpha$
   - Fit the $p-1$ two-variable models that include the variable selected in the previous step
   - Continue adding one variable at each iteration, including all variables selected in the previous step, until no more variables can be added (all p-values $< \alpha$)
3. Stepwise selection
   - Begin with forward selection, and employ both forward selection and backward elimination at each step until no more variables can be added or removed

Selection From All Model Subsets

With p potential variables, there are $2^p - 1$ possible models to choose from.

Selection of reasonable models from all potential models is based on some measure of fit that takes into account both the $\text{SS}(Res)$ and the number of variables. Two such measurses are:

- Adjusted R-squared
- Mallows' Cp

Adjusted R-squared

Recall the coefficent of determination, given by

$$R^2 = 1 - \frac{\text{SS}(Res)}{\text{SS}(Tot)}$$

Note that with the addition of more variables, $\text{SS}(Res)$ will always decrease, and subsequently, R^2 will always increase when variables are added regardless of whether the variables account for a significant amount of the variation in the response. For this reason, **we cannot use R^2 as a relative measure of fit when comparing model subsets with different numbers of parameters.**

Definition 2.24 Instead, we can use the **adjusted R-squared**, given by

$$R^2_{\text{adj}} = 1 - \frac{\text{SS}(Res)/(n-(p+1))}{\text{SS}(Tot)/(n-1)}$$

Since R^2_{adj} takes into account the number of variables in the model, it will only increase if the variation accounted for by the added variable(s) **increases proportionally more than the degrees of freedom decreases** through the estimation of the additional parameters.

Note that since we can express

$$R^2_{\text{adj}} = 1 - \frac{\text{SS}(Res)/(n-(p+1))}{\text{SS}(Tot)/(n-1)} = \frac{\hat{\sigma}^2}{\text{SS}(Tot)/(n-1)}$$

as a function of the residual standard error, model selection based on a large R^2_{adj} is equivalent to selection based on a low residual standard error, $\hat{\sigma}$.

Mallows' C_p

Definition 2.25 For a k-variable model ($k = 1, 2, \ldots, p$), **Mallows' C_p** is defined as

$$C_p = \frac{\text{SS}(Res)_k}{\text{MS}(Res)_p} + 2(k+1) - n$$

Intuitively, the smaller the $\text{SS}(Res)$ for a given k, the better the model. Thus, smaller C_p values **relative to the number of variables** are associated with more suitable models.

Mallows' C_p is used to compare a k-variable model ($k < p$) with the full model (for which $k = p$.)

Summary

A k-variable model is preferred over the full model if

$$C_p \leqslant k + 1$$

Note that for the full ($k = p$) model

$$C_p = \frac{\text{SS}(Res)_k}{\text{MS}(Res)_p} + 2(k+1) - n = p + 1$$

regardless of the fit of the full model.

2.15.2 Model Selection – House Data

$\frac{1}{2}$, $\tfrac{1}{2}$