DEPARTMENT OF MATHEMATICS AND STATISTICS

Memorial University of Newfoundland **St. John's, Newfoundland**
CANADA A1C 5S7 *ph. (709) 864-8075 fax (709) 864-3010*

Alwell Julius Oyet, Phd *email: aoyet@mun.ca*

Statistics 6519 - Regression Analysis
Review Lecture Notes

1 Introduction

One of the main reasons for collecting data about a system (collection of machines, group of people or places, etc) for analysis is the desire to discover something new about the system or to answer some questions about the system. Once the data has been gathered, the data is organized and a preliminary analysis of the data is conducted to identify any special features that may be present in the data, such as the distribution, outliers, shape of the relationship between variables, etc. We then have to decide which of the many statistical techniques we should apply to produce results that will answer our questions about the system. For instance, one may ask, can the grade of students be predicted based on their age? Is there a significant relationship between age and grades that can be used to predict grades? What kind of relationship exist between the price of a home and the square footage? Any attempt at answering these questions will then begin with gathering information (data) about the appropriate variables for analysis. Regression analysis is one of the several techniques available in the literature for data analysis. It is widely used in several areas of applications. Regression methods use the relationship between two or more variables, an outcome/response/dependent variable and one or more predictor/independent/regressor variables, to develop a mathematical expression or statistical model which best describes the relationship for the purpose of predicting the outcome variable.

Example 1

Suppose we wish to develop an equation or model for predicting Stats 6519 grades given the age of students. In this example,

Grade denoted by y - outcome or response or dependent variable
Age denoted by x - regressor or predictor or independent variable

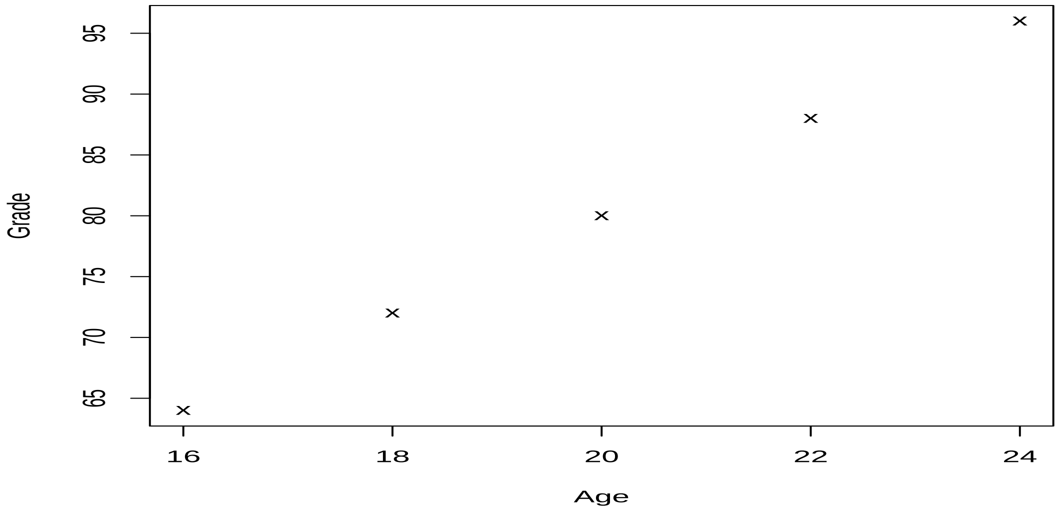


Figure 1: A scatter plot of Grades against Age of Students in Stats 6519.

In order to develop the model or equation, we first gather information or data on Age(x) and Grades(y) in Stats 6519 from past years. Suppose the data we obtained is as follows

Age(x)	16	18	20	22	24
Grade(y)	64	72	80	88	96

To help us decide on an appropriate relationship or model for the data, we begin by constructing a scatter plot of the data, using the R software, as shown in Figure 1. A visual examination of the scatter plot immediately tells us that there is a perfect straight line relationship between Age and Grade. Also, the straight line appear to pass through the origin. Thus, the intercept on the y (Grade) axis is 0. Now, from our high school math, we know that the equation of a straight line relationship between x and y is

$$y = mx + c,$$

where c is the intercept on the y axis and m is the gradient. For this data, we now know that the intercept $c = 0$. After a little bit of algebra, we find that the gradient

$$m = \frac{96 - 64}{24 - 16} = 4.$$

Therefore, the equation or model for the data is

$$Grade = 4 \times Age, \quad \text{OR} \quad y = 4x.$$

This model describes an exact relationship between Age and Grade. That is, based on this formula a 16 year old student must score a grade of 64. The older you are the better your grade becomes. It is clear that this model does not represent any realistic situation because some younger students can perform better than some older students and students of the same age may not necessarily score the same grade in Stat 6519 all the time. Therefore, a realistic model should make provision for variation in the outcome for the same value of the predictor.

Such a model which describes an exact relationship between two or more variables is referred to as a **deterministic or functional model**.

Example 2

Suppose the base price of a home (empty plot of land) in St. John's is \$50,000 and the cost of each square footage of a home is m. Then, given the square feet of a home, a real estate agent decides to predict the price of a home using the model

$$\text{Price} = \$50,000 + m \times \text{square feet.}$$

By this model, every property with the same square feet will be priced equally. Thus this is a deterministic or functional model. Unfortunately, this is not the case in real life.

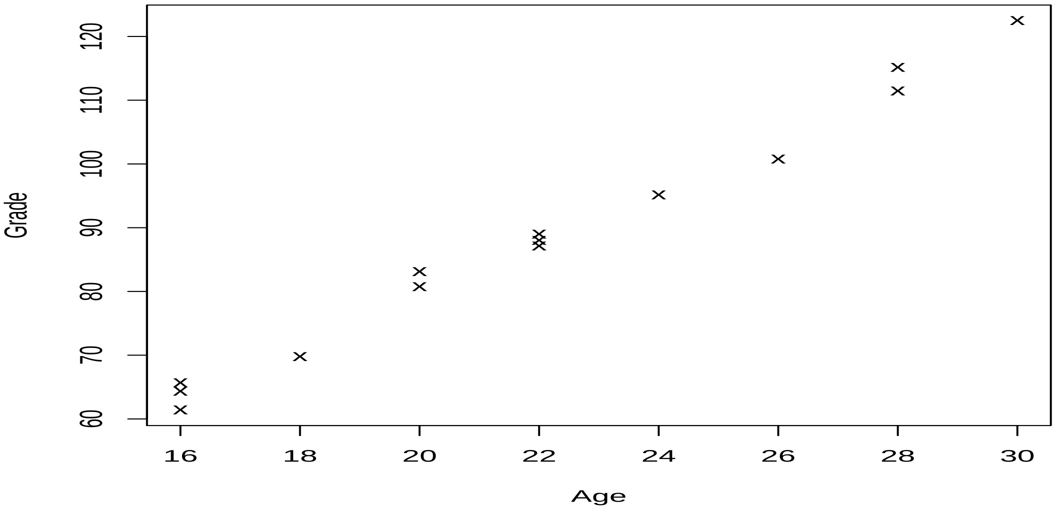


Figure 2: A scatter plot of Grades against Age of Students.

In order to account for the variation in outcomes, y for the same value of a predictor, x, statisticians add a random component to the deterministic model in order to obtain a statistical or probabilistic model. The random component is usually denoted by ϵ. Thus, a statistical or probabilistic regression model for Examples 1 and 2 will be of the form

$$Grade = 4 \times Age + \epsilon,$$

and

$$\text{Price} = \$50,000 + m \times \text{square feet} + \epsilon,$$

respectively, where ϵ is a random variable with some mean and variance. The scatter plot of data for such a model may look like the graph in Figure 2 where all the points do not fall on a perfect straight line and the outcome for the same predictor could be different. It is common to assume that the mean of ϵ is 0 and the variance is denoted by σ^2.

We know that life is not always straightforward. We sometimes encounter bumps and turns and curves in life. In the same way, the relationship between variables is not always linear or a straight line. It could be a curve sometimes, as shown in Figure 3. Furthermore, the price of a home in St. John's may not only depend on Square feet but also on Location. So, if we denote square feet by x_1 and location by x_2, then we have a model with 2 predictor variables given by

$$\text{Price} = \$50,000 + m_1 \times \text{square feet} + m_2 \times \textit{location} + \epsilon.$$

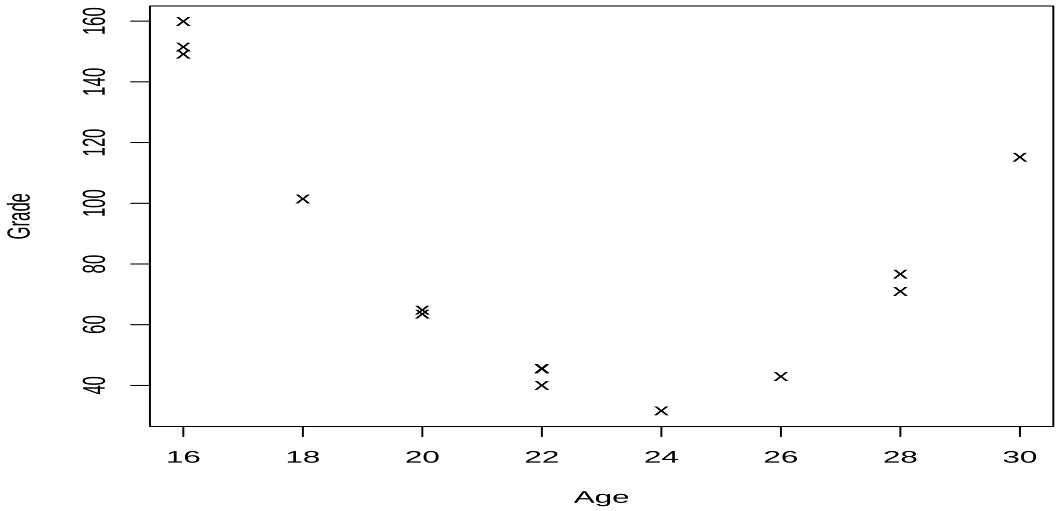


Figure 3: A curvilinear relationship between Grades and Age of Students.

These examples show that regression models can take various forms. It could be linear, nonlinear or involve more than one predictor variable. Our objective in this course is to use available data to build a suitable model for predicting an outcome variable. Our focus, will however be on linear relationships. For the purpose of prediction, we will assume that the relationship (which is our best guess based on the scatter plot and analysis) we have developed will continue into the future. Developing the model, will involve

(a) estimating the unknown coefficients of the predictor variable(s).

(b) determining whether the model we have built is suitable or adequate for predicting y. This will be accomplished through hypotheses testing and model adequacy checks.

1.1 Overview of steps in regression analysis

Step 1. Exploratory Data Analysis: Construct scatter plots, box plots and any plots needed to explore the relationships in the data, distribution of the data and presence of unusually small or large observations (possible outliers) that may affect the results of your analysis.

Step 2. Model Fitting: Build or develop tentative models for the data.

Step 3. Model Diagnostics: Examine adequacy of model. Revise models if necessary.

Step 4. Model Selection: Identify and select the most suitable model for the data.

Step 5. Inference: Conduct tests based on selected model.

Step 5. Prediction: Predict outcome, if needed.

We will begin with the simplest regression model referred to as simple linear regression (SLR).

2 Simple Linear Regression (SLR)

The simplest regression model with only 1 predictor or independent variable x assumes that on the average, there is a straight line relationship between x and the outcome variable y. The model is commonly referred to as a simple linear regression model. The SLR model comprises of a deterministic/functional component represented as, $\beta_0 + \beta_1 x$, and a random component denoted by ϵ. Suppose, a total of n outcomes were measured. Given the ith, $i = 1, 2, \ldots, n$ outcome y_i measured at the ith predictor variable x_i, the SLR model is given by,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, 3, \ldots, n. \tag{2.1}$$

In the model (2.1), the unknown parameter β_0 is the intercept which represents the average value of the outcome when the value of the predictor, x is 0. If the value of x cannot be 0, then β_0 has no practical interpretation. The unknown coefficient of the predictor β_1 is the average value by which the outcome will change (increase or decrease) if the value of the predictor increases by 1 unit. It is common to assume that,

(a) $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are uncorrelated random variables.

(b) the probability distribution of ϵ_i has mean $E(\epsilon_i) = 0$, $i = 1, \ldots, n$, and

(c) constant variance, $V(\epsilon_i) = \sigma^2$, $i = 1, \ldots, n$.

The implication of these assumptions are,

(i) $Cov(\epsilon_i, \epsilon_j) = 0$,

(ii) $y_1, y_2, \ldots, y_n$ are uncorrelated random variables. That is, $Cov(y_i, y_j) = 0$.

(*ii*) The probability distribution of y_i, has mean

$$\mu_{y_i} = E(y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i.$$

and variance

$$\sigma^2_{y_i} = V(y_i) = Var(\beta_0 + \beta_1 x_i + \epsilon_i) = \sigma^2,$$

a constant, regardless of the value of x_i.

For the purpose of inference, it is also assumed that the probability distribution of ϵ_i, $i = 1, 2, \ldots, n$, is the normal distribution.

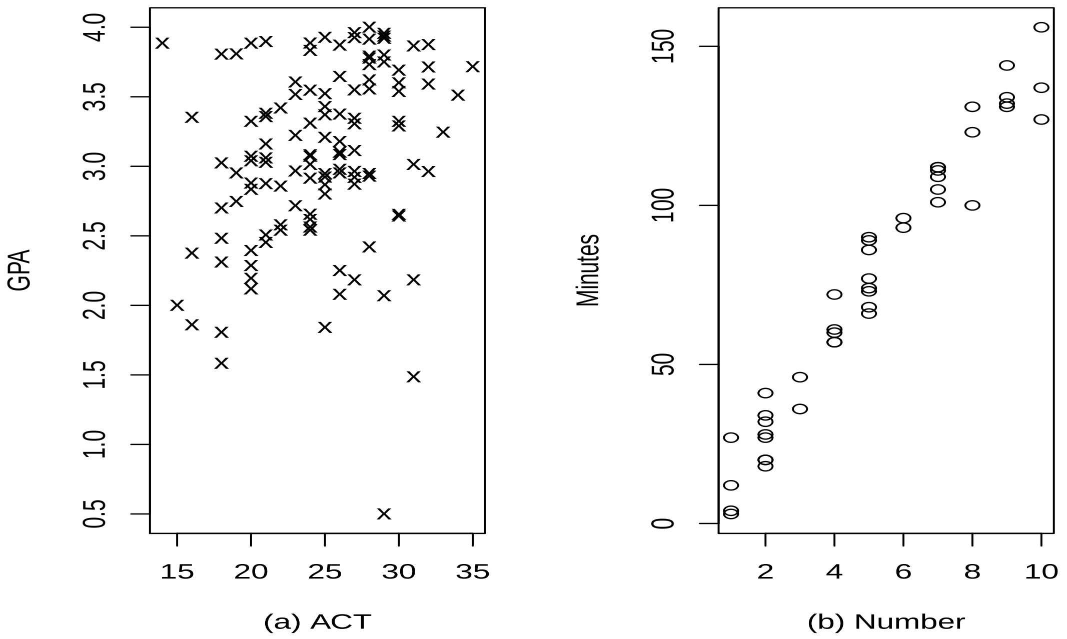


Figure 4: A scatter plot of (a) GPA against ACT Scores, (b) Minutes spent by a service person versus Number of copiers serviced.

Example 3.

Consider the scatter plot in Figure 4, of the data on Grade Point Average (GPA) and ACT scores in Problem 1.19 of Chapter 1, on Page 35 and the data on copier maintenance in Problem 1.20 of Chapter 1. For the GPA data $n = 120$ and $n = 45$ for the copier maintenance data. In our preliminary analysis of these data sets based on the scatter plot, the graph in Figure 4(a) shows that the spread of the GPA values for a given value of ACT scores varies widely from one value of ACT scores to another. Thus, one will anticipate that the assumption of constant variance, regardless of the value of x_i, may not be valid for this data. Whereas, the spread is about the same for the copier maintenance data in Figure 4(b). Also, the variation of GPA values about the average values of GPA for each

> ACT score is also large. This will make a straight line graph to be a poor fit for the GPA values. The situation is however different for the copier maintenance data. The variation about the average minutes a service person spent at a given location is not that large. So, a straight line graph may be a good fit for this data. Furthermore, the graph in Figure 4(b) does not show any point that is far away from the rest of the points, meaning there are no possible outliers in this data. On the other hand, we notice a point close to the ACT axis which appear to be far away from the rest of the points in Figure 4(a). This point may be a possible outlier or influential point. As a result of our preliminary analysis it becomes clear that the GPA data in Figure 4(a) may be a more challenging data to analyze using simple linear regression methods. The analysis of the copier maintenance data is however likely to be more straightforward.

Based on our findings from the preliminary analysis of our data, we proceed to fit a simple linear regression model to the data or seek an alternative model for our data. Here, in order to confirm our findings, we will use the two data sets to illustrate how to use regression analysis methods to decide whether a SLR model is a suitable model for a data set.

2.1 Estimation of the regression function

There are 3 unknown parameters in the regression model (2.1), namely β_0, β_1 and σ^2. These unknown parameters are usually estimated using the observed measurements (x_1, y_1), (x_2, y_2), $\ldots, (x_n, y_n)$. The method we will be using to estimate the parameters of the straight line regression function β_0 and β_1 is called the method of least squares. For the purpose of illustration, consider the data on GPA in Figure 4(a). Notice that there are several possible lines that we can fit to the points on the graph. Three of these lines are shown in Figure 5. The perpendicular distance from each point on the scatter plot to the line is the difference between the observed value and the estimated value for that point. The method of least squares uses the sum of squares of these differences for all the points as the criterion for selecting the estimated values of β_0 and β_1. Among all possible lines that one can draw to fit the points on the scatter plot, the method of least squares selects the values of β_0 and β_1 that gives the smallest sum of squares of the deviations. We shall denote these estimates by $\hat{\beta}_0$ and $\hat{\beta}_1$.

Mathematically, this means that among all possible values of β_0 and β_1, the method of least squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values that minimize the sum of squares of deviations, given by,

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2. \tag{2.2}$$

From calculus, we recall that to minimize an objective function with respect to an unknown parameter, we first find the stationary points of the function. To do this, we differentiate with respect to the parameters, equate to zero and solve the equations. In this case, we solve

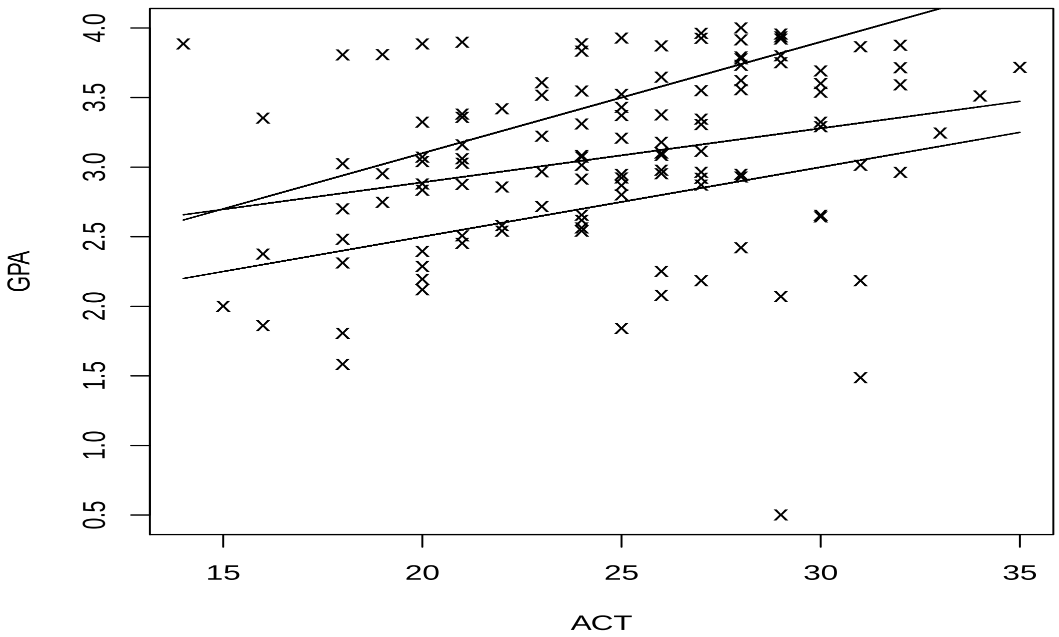


Figure 5: An Example of Some Fitted Lines To The GPA Data for Students in Stats 3521.

the the simultaneous equations given by,

$$\left.\frac{\partial Q}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 2\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]\cdot(-1) = 0 \tag{2.3}$$

$$\left.\frac{\partial Q}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 2\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]\cdot(-x_i) = 0. \tag{2.4}$$

Equations (2.3) and (2.4) are commonly referred to as the normal equations. Simplifying equations (2.3) and (2.4) leads to the equations

$$n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i \tag{2.5}$$

$$\hat{\beta}_0\sum_{i=1}^{n}x_i + \hat{\beta}_1\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_i y_i, \tag{2.6}$$

respectively. Solving equations (2.5) and (2.6) simultaneously for $\hat{\beta}_0$ and $\hat{\beta}_1$, yields the least squares estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}, \tag{2.7}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \tag{2.8}$$

where $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$ and $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$. We note that,

$$\left.\frac{\partial^2 Q}{\partial \beta_0^2}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 2n > 0 \quad \text{and} \quad \left.\frac{\partial^2 Q}{\partial \beta_1^2}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 2\sum_{i=1}^{n} x_i^2 > 0.$$

Therefore, the points $\hat{\beta}_0$ and $\hat{\beta}_1$, are minimum points. The fitted regression function is then written as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \ldots, n. \tag{2.9}$$

Using the expression for $\hat{\beta}_0$, we can rewrite $\hat{y}_i$ as,

$$\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}), \quad i = 1, 2, \ldots, n. \tag{2.10}$$

From (2.1) we see that the mean of the probability distribution of the outcome or response variable is

$$\mu_{y_i} = E(y_i) = \beta_0 + \beta_1 x_i.$$

Thus, it is clear that the fitted regression function (2.9) is actually an estimate of the mean μ_{y_i} of the probability distribution of y at the predictor x_i.

Furthermore, notice that the expression for $\hat{\beta}_1$ can also be written as

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] y_i = \sum_{i=1}^{n} k_i y_i, \tag{2.11}$$

where

$$k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.12}$$

Clearly,

$$\sum_{i=1}^{n} k_i = 0, \quad \text{and} \quad \sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.13}$$

Using the result in (2.11) it is easy to show that,

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - k_i \bar{x}\right) y_i. \tag{2.14}$$

These results show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of the observed responses $y_1, \ldots, y_n$. We encourage students to pay attention to these important properties because they will be very useful later.

Example 4 (Grade Point Average): Problem 1.19, Page 35

For the data on grade point average, the outcome variable (y) is GPA and the predictor or independent variable (x) is ACT scores. Applying the least squares method to this data, we have

$$n = 120, \quad \sum_{i=1}^{n} x_i = 2967, \quad \bar{x} = 2967/120 = 24.725, \quad \sum_{i=1}^{n} x_i^2 = 75739,$$

$$\sum_{i=1}^{n} y_i = 368.886, \quad \bar{y} = 368.886/120 = 3.07405, \quad \sum_{i=1}^{n} x_i y_i = 9213.112.$$

It follows that,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{9213.112 - 120 \cdot (24.725) \cdot (3.07405)}{75739 - 120 \cdot (24.725)^2} \approx 0.03883, \tag{2.15}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.07405 - (0.03883)(24.725) \approx 2.114. \tag{2.16}$$

That is, an average student with an ACT score of 0 is expected to have a GPA of 2.114. If the ACT Score of an average student increases by 1 unit, their GPA is expected to increase by approximately 0.0339. Thus, the estimated or fitted regression function for the GPA and ACT scores data is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 2.114 + 0.03883 x_i. \tag{2.17}$$

The fitted regression line is overlaid on the scatter plot in Figure 6(a). A visual examination of Figure 6(a) shows that the points are widely spread out around the fitted regression line. This is an indication that the proportion of variation in the data accounted for or explained by the fitted line is likely going to be very small. This also indicates that the estimate of the variance σ^2 of the probability distribution of the response y is likely going to be large in magnitude. Thus, the fitted line is likely going to be a poor fit for this data. For the GPA data, the ACT score x_5 for the 5th student was $x_5 = 21$. On the average, any stduent with a similar ACT score is expected to have an estimated GPA of,

$$\hat{y}_5 = 2.114 + 0.03883 * 21 \approx 2.9294.$$

One can use a similar approach to compute the fitted or estimated GPA for students with ACT scores that are similar to those in the data.

Example 5 (Copier Maintenance): Problem 1.20, Page 35

In the copier maintenance data, the outcome or dependent variable (y) is the total number of minutes spent by the service person and the independent or

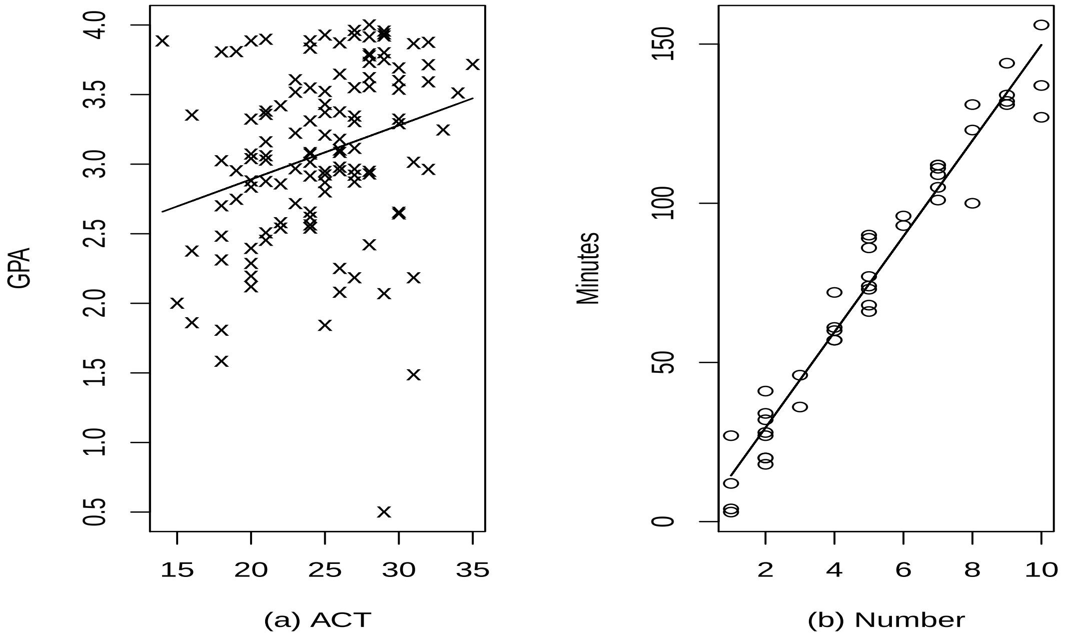


Figure 6: Fitted Regression Lines Overlaid on GPA and Copier Maintenance Data.

predictor variable or regressor (x) is the number of copiers serviced. Using the R software, we found that

$$n = 45, \quad \sum_{i=1}^{n} x_i = 230, \quad \bar{x} = 230/45 \approx 5.111, \quad \sum_{i=1}^{n} x_i^2 = 1516,$$

$$\sum_{i=1}^{n} y_i = 3432, \quad \bar{y} = 3432/45 \approx 76.267, \quad \sum_{i=1}^{n} x_i y_i = 22660.$$

Therefore, for the copier maintenance data,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{22660 - 45 \cdot (5.111) \cdot (76.267)}{1516 - 45 \cdot (5.111)^2} \approx 15.035, \tag{2.18}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 76.267 - (15.035)(5.111) \approx -0.5802. \tag{2.19}$$

Here, the number of minutes spent by a service person should be 0 if no copier was serviced. Thus, in this example $\hat{\beta}_0$ has no practical interpretation. However, the average number of minutes a service person spends on a copier is expected to increase by 15.035 minutes if the number of copiers increases by 1 unit. Based on the estimates obtained, the fitted or estimated regression function for the copier maintenance data is,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = -0.5802 + 15.035 x_i. \tag{2.20}$$

Again, we overlay the fitted regression line on the scatter plot in Figure 6(b). This graph shows that the points are very close to the fitted regression line. Thus, the fitted line will likely explain a large proportion of the variation in the data. Also, the estimate of the variance of the probability distribution of the outcome variable y is unlikely to be too large in magnitude. Therefore, the fitted line is likely going to be a good fit. These conclusions made by visually examining the graphs will be subsequently confirmed or disproved through formal statistical analysis in later sections of this course. The service person serviced $x_6 = 10$ copiers at the 6th location he visited. We can estimate that on the average, he spent about

$$\hat{y}_6 = -0.5802 + (15.035)(10) \approx 149.772,$$

minutes at this location.

2.2 Residuals

The ith residual is the perpendicular distance in the y-direction between the ith point (x_i, y_i) on the scatter plot and the ith point $(x_i, \hat{y}_i)$ on the overlaid fitted regression line. Thus, the ith residual denoted by e_i measures the deviation of the ith fitted response $\hat{y}_i$ from the observed value y_i. Mathematically, this means,

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n.$$

Using the result in (2.10), it is clear that the residuals can also be written as

$$e_i = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x}), \quad i = 1, 2, \ldots, n. \tag{2.21}$$

For instance, the residual for the 5th GPA score is

$$e_5 = y_5 - \hat{y}_5 = 3.028 - 2.9294 = 0.0986.$$

For the copier maintenance data,

$$e_6 = y_6 - \hat{y}_6 = 137 - 149.77232 = -12.77232.$$

From these examples, we can see that some of the residuals will be negative and some will be positive. This makes sense since $\hat{y}$ is an estimate of the mean of y. So, we expect $\hat{y}$ to be in the "middle" of the y values (see Figure 6). This implies that for a normally distributed response variable, 50% of the values of $\hat{y}$ will be smaller than the y values and the other half will be larger. That is, $e_i < 0$, 50% of the time and $e_i > 0$, 50% of the time. So, we expect $\sum_{i=1}^{n} e_i = 0$. We will prove this mathematically, later.

We also note that the residuals e_i are an estimate of the unknown error terms ϵ_i. We had assumed that the unknown error terms are uncorrelated normal random variables with mean zero and constant variance σ^2. The residuals, which are known, will therefore be quite useful in examining the validity of these assumptions. We mention, here, that if these assumptions do not hold for any data set, the SLR model cannot be used as a model for that data. The analyst must take remedial measures to correct any violations before the SLR model can be used for that data.

2.3 Properties of the fitted regression model

1. The first property we shall discuss is that the sum of the residuals of a fitted regression function is 0. Using equation (2.21), it is clear that

$$\sum_{i=1}^{n} e_i = 0.$$

It is easy to verify that this property holds for the GPA and copier maintenance data.

2. As a result of the least squares method, the minimum value of the objective function for the least squares method in (2.2),

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2,$$

is attained at $\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$. That is, the minimum value of $Q(\beta_0, \beta_1)$ is

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 = \sum_{i=1}^{n} e_i^2,$$

the sum of squares of the residual denoted by SSE. In our example on GPA, $SSE = 45.81761$ and for the copier maintenance data $SSE = 3416.377$. Using, the result in (2.21), we can show that

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = SS_{yy} - \hat{\beta}_1^2 SS_{xx},$$

where,

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{and} \quad SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

3. From Property 1, we have that

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0.$$

It follows that,

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$$

That is, the sum of the observed outcomes are equal to the sum of the fitted values.

4. From the normal equation (2.4), we have

$$\sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] \cdot (-x_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i) \cdot (-x_i) = \sum_{i=1}^{n} x_i e_i = 0.$$

5. Next, using Properties 1 and 4 above, we have that,

$$\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i = 0.$$

6. Using equation (2.10) it is clear that when $x_i = \bar{x}$, the fitted value becomes

$$\hat{y}_i = \bar{y} + \hat{\beta}_1(\bar{x} - \bar{x}) = \bar{y}.$$

   That is, the regression function always goes through the point $(\bar{x}, \bar{y})$.

7. The last property is an important theorem referred to as the Gauss-Markov theorem which we shall not prove in this course. It is however very important for us to understand it. It states,

   > *Under the conditions of the SLR model, the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all unbiased linear estimators.*

   That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the best linear unbiased estimators (BLUE) of β_0 and β_1, respectively.

   (a) **Linearity:** It is clear that $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of $y_1, y_2, \ldots, y_n$.

   (b) **Unbiased:** By unbiased we mean that,

   $$E(\hat{\beta}_0) = \beta_0, \quad \text{and} \quad E(\hat{\beta}_1) = \beta_1.$$

   This means that $\hat{\beta}_0$ and $\hat{\beta}_1$ will not systematically underestimate or overestimate β_0 and β_1.

   (c) **Minimum Variance (Best):** This property implies that $\hat{\beta}_0$ and $\hat{\beta}_1$ are much more precise estimators of β_0 and β_1 than any other unbiased estimators that are linear functions of $y_1, y_2, \ldots, y_n$.

2.4 Estimation of σ^2

Let Y be a random variable with mean μ_y and variance σ^2. Then, theoretically

$$\sigma^2 = E[(Y - \mu_y)^2].$$

If $y_1, y_2, \ldots, y_n$ is a set of n measurements on Y, then μ_y can be estimated by $\bar{y}$ and σ^2 estimated by the sample variance,

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}.$$

A quick explanation of why we divide by $n-1$ is that the n deviations in the numerator of s^2, $(y_i - \bar{y})$, $i = 1, \ldots, n$, are linearly dependent but $n-1$ of the deviations are linearly independent. That is why,

$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0, \quad \text{but} \quad \sum_{i=1}^{n-1}(y_i - \bar{y}) \neq 0.$$

Notice also, that μ_y was replaced by it's estimate, $\bar{y}$ in the sample version, s^2, of σ^2.

In regression analysis, we have seen that the mean of the distribution of Y, μ_y, is not estimated by $\bar{y}$ but by $\hat{y}$. So, in regression analysis, we will replace μ_y by $\hat{y}$ in the sample version of σ^2. Once μ_y is replaced by $\hat{y}$, we see that $(y_i - \hat{y}_i)$ are the residuals e_i and that

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Clearly, the n deviations $e_i = (y_i - \hat{y}_i)$ contains 2 deviations $(y_i - \bar{y})$ and $(x_i - \bar{x})$, that are each linearly dependent. So, to estimate σ^2 in simple linear regression we divide by $(n-2)$, instead of $(n-1)$. Thus, in SLR an estimate of the variance of the distribution of y_i is

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}. \tag{2.22}$$

We will see later that the estimate s^2 of σ^2 is the mean squared error (MSE). We will refer to the term, $(n-2)$ as the error degrees of freedom. Also, it can be shown that s^2, the MSE is an unbiased estimator of σ^2.

Example 6 (Grade Point Average)

Continuing with the data on GPA of 120 students with ACT scores, we find that

$$SSE = \sum_{i=1}^{n} e_i^2 = 45.81761.$$

Therefore, the variance estimate for this data becomes

$$MSE = s^2 = \hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{45.81761}{118} \approx 0.3883.$$

Example 7 (Copier Maintenance)

For the 45 observed responses in the copier maintenance data we have that

$$SSE = \sum_{i=1}^{n} e_i^2 = 3416.377.$$

It follows that

$$MSE = s^2 = \hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{3416.377}{43} \approx 79.4506.$$

We end this section by noting that no distributional assumptions where required or used for the least squares estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$. In the next section, we will discuss interval estimation of some of the model parameters and hypotheses testing. These will require us to use the normality assumption of the error terms we made earlier.

Statistics 6519 - Regression Analysis
Lecture Notes (Contd)

3 Inferences in SLR

In Section 2, we assumed that a linear relationship exist between x and y based on a visual examination of the scatter plot of y against x. We then fitted a simple linear regression model for predicting y given a new value of x, to the data. In this section, we will develop formal statistical methods for testing whether the assumed linear relationship exist and whether it is significant enough for the model to be useful for prediction purposes.

3.1 Inferences concerning β_1

We observe that if $\beta_1 = 0$ in the SLR model (2.1), then the model will have no predictor or independent variable and reduces to $y_i = \beta_0 + \epsilon$. That means, the predictor has no influence on the outcome y and there is no linear relationship between x and y. Note that in practice, β may not be exactly zero due to the random component of the data. Thus, we examine if there is a statistically significant linear relationship between x and y by testing the null H_0 and alternative H_a hypotheses,

$$H_0 : \ \beta_1 = 0, \quad vs \quad H_a : \ \beta_1 \neq 0. \tag{3.1}$$

Before developing the test statistic for the hypotheses we will derive the sampling distribution of the estimator $\hat{\beta}_1$.

<u>**Sampling distribution of $\hat{\beta}_1$**</u>

Recall that any value computed from a sample, is a sample statistic. That means, $\hat{\beta}_1$ is a random variable and a statistic. Therefore, it has a probability distribution with some mean and some variance. The probability distribution of a sample statistic is commonly referred to as the sampling distribution of the statistic. In Section 2, we assumed that $\epsilon_1, \dots, \epsilon_n$ follow a normal distribution with mean 0 and variance σ^2, we will write as, $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$. It follows from the model (2.1) that $y_i \sim N(\mu_{y_i}, \sigma^2)$, where $\mu_{y_i} = \beta_0 + \beta_1 x_i$. Also, in Section 2, equations (2.11) and (2.12), we showed that

$$\hat{\beta}_1 = \sum_{i=1}^{n} k_i y_i, \quad k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Since $\hat{\beta}_1 = \sum_{i=1}^n k_i y_i$ is a linear combination of normally distributed random variables, it follows that $\hat{\beta}_1$ will also be normally distributed with some mean $\mu_{\hat{\beta}_1}$ and variance $\sigma^2_{\hat{\beta}_1}$. Now,

$$\mu_{\hat{\beta}_1} = E(\hat{\beta}_1) = \sum_{i=1}^n k_i E(y_i) = \sum_{i=1}^n k_i \mu_{y_i} = \sum_{i=1}^n k_i(\beta_0 + \beta_1 x_i). \tag{3.2}$$

Using the result $\sum_{i=1}^n k_i = 0$ in (2.13) and the fact that $\sum_{i=1}^n k_i x_i = 1$, in (3.2), we have that

$$\mu_{\hat{\beta}_1} = \beta_1. \tag{3.3}$$

That is, $\hat{\beta}_1$ is an unbiased estimator of β_1. Furthermore, the variance of the probability distribution of $\hat{\beta}_1$ is,

$$\sigma^2_{\hat{\beta}_1} = \mathrm{V}(\hat{\beta}_1) = \sum_{i=1}^n k_i^2 V(y_i) = \sum_{i=1}^n k_i^2 \sigma^2.$$

From (2.13), we have that

$$\sum_{i=1}^n k_i^2 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Therefore,

$$\sigma^2_{\hat{\beta}_1} = \mathrm{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{3.4}$$

In summary, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where $\sigma^2_{\hat{\beta}_1}$ is given by (3.4). It is clear that $\sigma^2_{\hat{\beta}_1}$ cannot be computed because σ^2 is unknown. The variance of the probability distribution of $\hat{\beta}_1$ is therefore estimated by replacing σ^2 with $\hat{\sigma}^2 = MSE$. That is,

$$s^2_{\hat{\beta}_1} = \hat{\sigma}^2_{\hat{\beta}_1} = \frac{MSE}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{3.5}$$

The square root of $s^2_{\hat{\beta}_1}$, denoted by $s_{\hat{\beta}_1} = \sqrt{s^2_{\hat{\beta}_1}}$ is commonly referred to as the standard error of $\hat{\beta}_1$.

Once the sampling distribution of $\hat{\beta}_1$ has been derived the test statistic for the hypotheses (3.1) can be constructed by first standardizing the probability distribution of $\hat{\beta}_1$ to obtain the standard normal random variable,

$$z_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2_{\hat{\beta}_1}}} \sim N(0,1). \tag{3.6}$$

After, some algebra, we find that, under the null distribution, a test statistic for the hypotheses (3.1) is a random variable t_0 which follows a student-t distribution with $n-2$ degrees of freedom given by

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{s^2_{\hat{\beta}_1}}} \sim t(n-2), \tag{3.7}$$

with $\beta_1 = 0$ under H_0. The corresponding p-value can be computed as

$$p - \text{value} = P(|T_{n-2}| > |t_0|) = 2[1 - P(T_{n-2} < |t_0|)]. \tag{3.8}$$

Then, for a fixed significance level α, we reject H_0 if $|t_0| > t(1-\alpha/2, n-2)$, where $t(1-\alpha/2, n-2)$ is the $(1-\alpha/2)100$ percentile of the student-t distribution with $n-2$ degrees of freedom. Note that due to symmetry of the t distribution, $t(1-\alpha/2, n-2) = -t(\alpha/2, n-2)$. Alternatively, we reject H_0 if p-value $< \alpha$.

Example 8 (Grade Point Average)

Suppose that we wish to test whether there is a significant linear relationship between ACT scores of students and their GPA based on the data available, at a significance level of 0.05. First, we estimate the variance of $\hat{\beta}_1$. Now,

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

From the data, we have

$$n = 120, \quad \sum_{i=1}^{n} x_i^2 = 75739 \ \text{ and } \bar{x} = 24.725.$$

It follows that,

$$SS_{xx} = 75739 - 120(24.725^2) = 2379.925.$$

From Example 6, $MSE = \hat{\sigma}^2 = 0.3883$. Therefore,

$$s^2_{\hat{\beta}_1} = \hat{\sigma}^2_{\hat{\beta}_1} = \frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{0.3883}{2379.925} \approx 0.000163,$$

and the standard error of $\hat{\beta}_1$ is $s_{\hat{\beta}_1} = \sqrt{0.000163} \approx 0.0128$. Then, we proceed as follows to test the required hypotheses.

$H_0: \ \beta_1 = 0$
$H_a: \ \beta_1 \neq 0.$

Significance level: $\alpha = 0.05$.

Test Statistic Value:

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{s^2_{\hat{\beta}_1}}} = \frac{0.03882713}{\sqrt{0.000163}} \approx 3.0398.$$

p-value: p-value $= 2[1 - P(T_{118} < 3.0398)] \approx 0.00292$.

Critical value: $t(0.975, 118) = 1.9803$.

Conclusion: We reject H_0 since $3.0398 > 1.9803$ (alternatively, $0.00292 < 0.05$). That means, evidence in the data supports the claim that there is a significant linear relationship between ACT scores and GPA.

Example 9 (Copier Maintenance)

For this data,

$$n = 45, \quad \sum_{i=1}^{n} x_i^2 = 1516 \ \text{ and } \bar{x} \approx 5.111.$$

Therefore,

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 1516 - 45(5.111^2) = 340.4444.$$

From Example 7, $MSE = \hat{\sigma}^2 \approx 79.4506$. So,

$$s^2_{\hat{\beta}_1} = \hat{\sigma}^2_{\hat{\beta}_1} = \frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{79.4506}{340.4444} \approx 0.2334,$$

and $s_{\hat{\beta}_1} \approx 0.4831$. We then proceed as in Example 8 to test the required hypotheses.

$H_0: \ \beta_1 = 0$
$H_a: \ \beta_1 \neq 0.$

Significance level: $\alpha = 0.05$.

Test Statistic Value:

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{s^2_{\hat{\beta}_1}}} = \frac{15.03525}{\sqrt{0.2333732}} \approx 31.123.$$

p-value: p-value $= 2[1 - P(T_{43} < 31.123)] = 0$.

Critical value: $t(0.975, 43) \approx 2.0167$.

Conclusion: We reject H_0 since $31.123 > 2.0167$ (alternatively, $0 < 0.05$). That means, there is very strong evidence in the data in support of the claim that there is a significant linear relationship between the number of copiers serviced and the time it takes the service person to complete the job.

Confidence Interval for β_1

We have seen that the random variable,

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{s^2_{\hat{\beta}_1}}},$$

follows a student-t distribution with $n-2$ degrees of freedom. For fixed α, we are then able to make the probability statement

$$P\left[t(\alpha/2, n-2) \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{s^2_{\hat{\beta}_1}}} \leq t(1-\alpha/2, n-2)\right] = 1 - \alpha.$$

After some rearrangement, we find that a $(1-\alpha)100\%$ confidence interval for β_1 are,

$$\hat{\beta}_1 \pm t(1-\alpha/2, n-2)\sqrt{s^2_{\hat{\beta}_1}}. \tag{3.9}$$

Example 10 (Grade Point Average)

From Examples 4 and 8 we have that $n = 120$, $\hat{\beta}_1 \approx 0.03883$ and $s_{\hat{\beta}_1} \approx 0.0128$. At $\alpha = 0.05$ we have that $t(1-\alpha/2, n-2) = t(0.975, 118) = 1.9803$. Therefore, a 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t(1-\alpha/2, n-2)\sqrt{s^2_{\hat{\beta}_1}} = 0.03883 \pm (1.9803)(0.0128).$$

Thus, the required interval is $0.0135 \leq \beta_1 \leq 0.0642$ or $(0.0135, 0.0642)$.

Example 11 (Copier Maintenance)

For this data, $n = 45$. Results from Examples 5 and 9 show that $\hat{\beta}_1 \approx 15.035$ and $s_{\hat{\beta}_1} \approx 0.4831$. At $\alpha = 0.01$ we have that $t(1-\alpha/2, n-2) = t(0.995, 43) = 2.6951$. Thus, a 99% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t(1-\alpha/2, n-2)\sqrt{s^2_{\hat{\beta}_1}} = 15.035 \pm (2.6951)(0.4831).$$

The required interval can then be written as, $13.733 \leq \beta_1 \leq 16.337$ or $(13.733, 16.337)$.

Analysis of variance (ANOVA) approach and F test for $\hat{\beta}_1$

The ANOVA approach to regression analysis is an alternative approach to the t test for the significance of a linear relationship between the response or outcome y and the predictor variable x. In this approach, we begin by partitioning the total variation in the response variable y into its components. This variation is usually measured by $y_i - \bar{y}$. In Figure 6, notice that the fitted regression lines $\hat{y}$ passes through some of the points on the scatterplots while some points are below or above the lines. That means, the fitted line only captures or explains or accounts for some of the variation in the response but does not capture or explain or account for the variation in the points above or below it. Thus, the total variation in y is made up of two components and we write,

Total variation = Explained Variation + Unexplained Variation.

Now, the total variation in y is measured by the overall deviation of the observed responses from the mean of the responses. That is,

$$\text{Total Variation (SST)} = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Next, we note that the deviation, $y_i - \bar{y}$, in the total variation above can be partitioned as follows,

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}.$$

If we square both sides and sum over the n points, we obtain

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

Students can use Properties 3 and 5 we discussed in Section 2.3 to show that

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0.$$

It follows that,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2. \tag{3.10}$$

In (3.10),

$$\text{Explained Variation} = \text{Regression sum of squares (SSR)} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

and,

$$\text{Unexplained Variation} = \text{Error sum of squares (SSE)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Students can use equation (2.10) to show that the regression sum of squares (SSR) can also be written as

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.11}$$

Students will find that equation (3.11) is an easier expression to use for hand calculations. Furthermore, since only $n-1$ of the n deviations, $y_i - \bar{y}$ in the total variation are linearly independent, the degrees of freedom associated with the total variation (SST) commonly referred to as the total degrees of freedom (df$_T$), is $n-1$. Similar to the partitioning of the total variation into two components, we also partition (df$_T$) into two components, regression degrees of freedom (df$_R$) that is associated with SSR and error degrees of freedom (df$_E$), associated with SSE. So, we have,

$$n-1\ (\mathrm{df}_T) = 1\ (\mathrm{df}_R) + n-2\ (\mathrm{df}_E).$$

In Section 2.4, we saw that $MSE = SSE/(n-2) = SSE/df_E$. In the same way, we compute $MSR = SSR/df_R$. It can be shown that the MSE is an unbiased estimator of the error variance σ^2. That is,

$$E(MSE) = \sigma^2.$$

We can also show that

$$E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Thus, when $H_0:\ \beta_1 = 0$ is true, both the MSE and MSR are unbiased estimators of the error variance. However, when H_0 is false, the MSR will be much larger than the MSE depending on the magnitude of β_1. Thus, another test statistic used in testing whether H_0 is true or false is the ratio,

$$F_0 = \frac{MSR}{MSE}.$$

When H_0 is true, it can be shown that F_0 follows an F-distribution with numerator degrees of freedom, 1 and denominator degrees of freedom $n-2$, we shall write in the form $F(1, n-2)$. For a fixed value of α we will reject H_0 if $F_0 > F(\alpha, 1, n-2)$, where $F(\alpha, 1, n-2)$ is the $(1-\alpha)100$ percentile of the F-distribution with degrees of freedoms, 1 and $n-2$. The associated p-value is computed as

$$p - \text{value} = 1 - P(F_{1,n-2} \leq F_0).$$

These results are usually summarized in a table, commonly referred to as an ANOVA table as shown below.

Table 1. Analysis of variance table for regression analysis

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F	p-value
Regression	SSR	1	MSR	$F_0 = \frac{MSR}{MSE}$	
Error	SSE	$n-2$	MSE		
Total	SST	$n-1$			

Example 12 (Grade Point Average)

Continuing with the GPA data with $n = 120$, we see that,

$$\mathrm{df}_R = 1,\ \ \mathrm{df}_T = n-1 = 119 \ \text{ and } \ \mathrm{df}_E = n-2 = 118.$$

Furthermore,

$$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 1183.379 - 120(3.07405^2) = 49.40545,$$

$$\begin{aligned}\text{SSR} &= \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}_1^2 \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = (0.03883^2)[75739 - 120(24.725^2)] \\ &\approx 3.58838.\end{aligned}$$

The error sum of squares (SSE) can then be computed by subtraction to obtain

$$SSE = SST - SSR = 49.40545 - 3.58838 = 45.81708.$$

These results show that a very large proportion, $45.81708/49.40545 \approx 92.7\%$ of the total variation in GPA is not accounted for by the fitted regression line. Though the result of the hypothesis test in Example 8 shows that there is a siginificant linear relationship between GPA and ACT scores, this result in Example 12 is a clear indication that the fitted regression line is not a good fit for this data. The line explains only 7.3% of the total variation in GPA scores.

Next, we compute the mean squares,

$$\text{MSR} = \frac{SSR}{df_R} = 3.58838, \quad \text{MSE} = \frac{SSE}{df_E} = \frac{45.81708}{118} \approx 0.38828,$$

and the F_0 value,

$$F_0 = \frac{MSR}{MSE} = \frac{3.58838}{0.38828} \approx 9.2417,$$

with corresponding p-value,

$$p - value = 1 - P(F_{1,118} \leq 9.2417) \approx 0.002914.$$

The ANOVA table for the GPA data is shown in Table 2.

Table 2. Analysis of variance table for GPA data

Source of	Sum of	Degrees of	Mean		
Variation	Squares	freedom	Square	F	p-value
Regression	3.5883	1	3.5883	9.2417	0.002914
Error	45.8171	118	0.3883		
Total	49.4054	119			

Based on the p-value in the ANOVA table, it is clear that evidence in the data does not support the null hypothesis, $H_0 : \ \beta_1 = 0$. Therefore, we conclude that there is evidence of a linear relationship between GPA and ACT scores. We will like to draw the attention of students to the fact that in Example 8, $t_0 = 3.0398$ and $t_0^2 \approx 9.241 \approx F_0$. It can be shown, theoretically, that in general $t_0^2 = F_0$.

Example 13 (Copier Maintenance)

For this data with $n = 45$, we see that,

$$\mathrm{df}_R = 1, \;\; \mathrm{df}_T = n - 1 = 44 \;\text{ and }\; \mathrm{df}_E = n - 2 = 43.$$

Furthermore,

$$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 342124 - 45(76.26667^2) = 80376.8,$$

$$\begin{aligned}\mathrm{SSR} &= \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}_1^2 \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = (15.035^2)[1516 - 45(5.111^2)] \\ &\approx 76957.88.\end{aligned}$$

The error sum of squares (SSE) can then be computed by subtraction to obtain

$$SSE = SST - SSR = 80376.8 - 76957.88 = 3418.916.$$

Contrary to the previous example, only $3418.916/80376.8 \approx 4.25\%$ of the total variation in the minutes spent on servicing the copiers was not accounted for by the fitted regression line. The fitted regression line captured $76957.88/80376.8 \approx 95.75\%$ of the total variation in the minutes spent. This is a clear indication that the fitted regression line is a good fit for this data.

Next, we compute the mean squares,

$$\mathrm{MSR} = \frac{SSR}{df_R} = 76957.88, \quad \mathrm{MSE} = \frac{SSE}{df_E} = \frac{3418.916}{43} \approx 79.50968,$$

and the F_0 value,

$$F_0 = \frac{MSR}{MSE} = \frac{76957.88}{79.50968} \approx 967.9058,$$

with corresponding p-value,

$$p - value = 1 - P(F_{1,118} \leq 967.9058) = 0.$$

The ANOVA table for the copier maintenance data is shown in Table 3.

Table 3. Analysis of variance table for copier maintenance data

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F	p-value
Regression	76957.9	1	76957.9	967.91	0
Error	3418.92	43	79.51		
Total	80376.8	44			

Based on the p-value in the ANOVA table, it is clear that there is very strong evidence in the data against the null hypothesis, $H_0: \ \beta_1 = 0$. Therefore, we conclude that there is very strong evidence of a linear relationship between the minutes a service person spends at a location and the number of copiers serviced. Again, we see that in Example 9, $t_0 = 31.123$ and $t_0^2 \approx 968.64 \approx F_0$. In general, $t_0^2 = F_0$.

3.2 Inferences concerning β_0

We noted earlier that when the predictor x cannot take the value 0, the intercept parameter β_0 has no practical interpretation. In this case, it is not necessary to conduct any test of hypothesis on β_0. In some cases, it may be necessary to test for the significance of the intercept. We begin with a discussion of the sampling distribution of $\hat{\beta}_0$.

Sampling distribution of $\hat{\beta}_0$

In Section 2.1, equation (2.14) we saw that $\hat{\beta}_0$ can be written as a linear combination of $y_1, \ldots, y_n$. Since y_i is normally distributed, it follows that $\hat{\beta}_0$ is also normally distributed with some mean $\mu_{\hat{\beta}_0}$ and variance $\sigma^2_{\hat{\beta}_0}$. Now, from (2.14) we have that

$$\mu_{\hat{\beta}_0} = E[\hat{\beta}_0] = \sum_{i=1}^{n} \left(\frac{1}{n} - k_i \bar{x} \right) E(y_i),$$

where $\mu_{y_i} = E(y_i) = \beta_0 + \beta_1 x_i$. Again, using the result, $\sum_{i=1}^{n} k_i = 0$ in (2.13) and the fact that $\sum_{i=1}^{n} k_i x_i = 1$, in (3.2), it can be shown that

$$\mu_{\hat{\beta}_0} = \beta_0. \tag{3.12}$$

This implies that $\hat{\beta}_0$ is also an unbiased estimator of β_0. Next, we obtain the variance, $\sigma^2_{\hat{\beta}_0}$ of the distribution of $\hat{\beta}_0$. To do this, we take the variance of (2.14) to obtain

$$\sigma^2_{\hat{\beta}_0} = V[\hat{\beta}_0] = \sum_{i=1}^{n} \left(\frac{1}{n} - k_i \bar{x} \right)^2 \sigma^2.$$

If we simplify the right hand side of the equation above and use $\sum_{i=1}^{n} k_i = 0$ in (2.13) and $\sum_{i=1}^{n} k_i x_i = 1$, in (3.2), we obtain

$$\sigma^2_{\hat{\beta}_0} = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \sigma^2. \tag{3.13}$$

Thus, the sampling distribution of $\hat{\beta}_0$ is the normal distribution with mean and variance given by (3.12) and (3.13), respectively. Clearly, $\sigma^2_{\hat{\beta}_0}$ in (3.13) cannot be computed because σ^2 is unknown. So, to estimate $\sigma^2_{\hat{\beta}_0}$, we replace σ^2 with $\hat{\sigma}^2 = MSE$ to obtain the variance estimate given by

$$s^2_{\hat{\beta}_0} = MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]. \tag{3.14}$$

The standard error of $\hat{\beta}_0$ is then $s_{\hat{\beta}_0} = \sqrt{s^2_{\hat{\beta}_0}}$.

Using the same arguments as before, the student can easily show that a test statistic for testing the hypotheses, $H_0:\ \beta_0 = 0$ against $H_a:\ \beta_0 \neq 0$ is

$$t_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{s^2_{\hat{\beta}_0}}} \sim t(n-2),$$

with $\beta_0 = 0$ under H_0. For a fixed value of α we reject H_0 if $|t_0| > t(1-\alpha/2, n-2)$ or if the $p - value < \alpha$, where

$$p - value = 2[1 - P(T_{n-2} < |t_0|)].$$

Confidence Interval for β_0

Construction of confidence intervals for β_0 follow directly from the results of the previous section on sampling distribution and tests for $\hat{\beta}_0$. A $(1-\alpha)100\%$ confidence interval for β_0 are

$$\hat{\beta}_0 \pm t(1-\alpha/2, n-2)s_{\hat{\beta}_0}.$$

Example 14 (Grade Point Average)

For the data on GPA we found that $n = 120$, $\bar{x} = 24.725$, $\hat{\beta}_0 \approx 2.114$ and $\hat{\beta}_1 \approx 0.03883$ (see Example 4). In Examples 6 and 8, we obtained $MSE \approx 0.3883$ and $SS_{xx} = \sum_{i=1}^{120}(x_i - \bar{x})^2 = 2379.925$, respectively. It follows that the estimate of the variance of the sampling distribution of $\hat{\beta}_0$ is,

$$s^2_{\hat{\beta}_0} = MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = (0.3883)\left[\frac{1}{120} + \frac{24.725^2}{2379.925}\right] \approx 0.10297.$$

Therefore, the standard error of $\hat{\beta}_0$ is, $s_{\hat{\beta}_0} = 0.320893$.

We now test the hypotheses on β_0 as follows.

$H_0:\ \beta_0 = 0$
$H_a:\ \beta_0 \neq 0.$

Significance level: $\alpha = 0.05$.

Test Statistic Value:

$$t_0 = \frac{\hat{\beta}_0}{\sqrt{s^2_{\hat{\beta}_0}}} = \frac{2.114}{0.320893} \approx 6.5879.$$

p-value: p-value $= 2[1 - P(T_{118} < 6.5879)] \approx 1.305e - 09$.

Critical value: $t(0.975, 118) = 1.9803$.

Conclusion: We reject H_0 since $6.5879 > 1.9803$ (alternatively, $1.305e - 09 < 0.05$). That means, evidence in the data supports the claim against $H_0 : \ \beta_0 = 0$.

Confidence Interval: A 95% confidence interval for β_0 are

$$\hat{\beta}_0 \pm t(0.975, 118) s_{\hat{\beta}_0} = 2.114 \pm (1.9803)(0.320893).$$

That is, the required interval is, $1.4785 \leq \beta_0 \leq 2.7495$.

Example 15 (Copier Maintenance)

For the copier maintenance data $\bar{x} = 5.111$, $\hat{\beta}_0 \approx -0.5802$ and $\hat{\beta}_1 \approx 15.035$ (see Example 5). In Examples 7 and 9, we obtained $MSE \approx 79.4506$ and $SS_{xx} = \sum_{i=1}^{120}(x_i - \bar{x})^2 = 340.4444$, respectively. It follows that the estimate of the variance of the sampling distribution of $\hat{\beta}_0$ is,

$$s^2_{\hat{\beta}_0} = MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = (79.4506)\left[\frac{1}{45} + \frac{5.111^2}{340.4444}\right] \approx 7.8677.$$

Therefore, the standard error of $\hat{\beta}_0$ is, $s_{\hat{\beta}_0} = 2.8049$.

We now test the hypotheses on β_0 as follows.

$H_0 : \ \beta_0 = 0$
$H_a : \ \beta_0 \neq 0$.

Significance level: $\alpha = 0.05$.

Test Statistic Value:

$$t_0 = \frac{\hat{\beta}_0}{\sqrt{s^2_{\hat{\beta}_0}}} = \frac{-0.5802}{2.8049} \approx -0.20685.$$

p-value: p-value $= 2[1 - P(T_{43} < 0.20685)] \approx 0.837$.

Critical value: $t(0.975, 43) = 2.0167$.

Conclusion: We do not reject H_0 since $0.20685 < 2.0167$ (alternatively, $0.837 > 0.05$). That means, evidence in the data supports the claim that $H_0 : \ \hat{\beta}_0 = 0$.

Confidence Interval: A 95% confidence interval for β_0 are

$$\hat{\beta}_0 \pm t(0.975, 43) s_{\hat{\beta}_0} = -0.5802 \pm (2.0167)(2.8049).$$

That is, the required interval is, $-6.236 \leq \beta_0 \leq 5.076$.

3.3 Interval estimation of mean response μ_y

Recall that the fitted regression line, $\hat{y}_h = \hat{\beta}_0 + \hat{\beta} x_h$ is an estimate of the mean of y, μ_{y_h} at a given value, $x = x_h$ of the predictor variable. For instance, it may be of interest to estimate the average GPA of students with low, medium and high ACT scores to determine whether students with high ACT scores outperform students with medium or low ACT scores, in general. An important and useful estimate is an interval estimate of μ_{y_h} which takes into account the spread of the mean over all values of the predictor. Proceeding as before, we first derive the sampling distribution of the point estimate, $\hat{\mu}_{y_h}$ of μ_{y_h}.

Sampling distribution of $\hat{\mu}_{y_h}$

In Section 2, equations (2.11) and (2.14) we saw that $\hat{\beta}_0$ and $\hat{\beta}_1$ can be expressed as linear combinations of $y_1, \ldots, y_n$ which are normally distributed random variables. Therefore, we concluded that $\hat{\beta}_0$ and $\hat{\beta}_1$ were also normally distributed with some mean and some variance. Now,

$$\hat{\mu}_{y_h} = \hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h, \tag{3.15}$$

is a linear function of $\hat{\beta}_0$ and $\hat{\beta}_1$ which are normally distributed random variables. Therefore, $\hat{\mu}_{y_h}$ is also a normally distributed random variable with some mean $\mu_{\hat{\mu}_{y_h}}$ and some variance $\sigma^2_{\hat{\mu}_{y_h}}$. In Sections 3.1 and 3.2 we saw that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively. That is,

$$E[\hat{\beta}_0] = \beta_0 \quad \text{and} \quad E[\hat{\beta}_1] = \beta_1.$$

It then follows from (3.15), that the mean of the sampling distribution of μ_{y_h}, is given by

$$\mu_{\hat{\mu}_{y_h}} = E[\hat{\mu}_{y_h}] = E[\hat{\beta}_0] + E[\hat{\beta}_1] x_h = \beta_0 + \beta_1 x_h = \mu_{y_h}.$$

Therefore, $\hat{\mu}_{y_h}$ is also an unbiased estimator of μ_{y_h}.

Next, we find the variance $\sigma^2_{\hat{\mu}_{y_h}}$. To do this, we first use equations (2.11) and (2.14) to rewrite $\hat{\mu}_{y_h}$ as,

$$\hat{\mu}_{y_h} = \sum_{i=1}^{n} \left[\frac{1}{n} + k_i (x_h - \bar{x}) \right] y_i. \tag{3.16}$$

Then, take the variance of (3.16) and simplify to obtain,

$$\sigma^2_{\hat{\mu}_{y_h}} = \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \sigma^2. \tag{3.17}$$

Thus, the sampling distribution of $\hat{\mu}_{y_h}$ is the normal distribution with mean, μ_{y_h} and variance given by (3.17). It is clear that to estimate the variance we replace σ^2 in (3.17) with the MSE to obtain,

$$s^2_{\hat{\mu}_{y_h}} = \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] MSE. \tag{3.18}$$

Standardizing $\hat{\mu}_{y_h}$ and proceeding as before, we can show that the statistic

$$t_0 = \frac{\hat{\mu}_{y_h} - \mu_{y_h}}{s_{\hat{\mu}_{y_h}}},$$

follows the student t-distribution with $n-2$ degrees of freedom. Therefore, for fixed α, a $(1-\alpha)100\%$ confidence interval for μ_{y_h} are,

$$\hat{\mu}_{y_h} \pm t(1-\alpha/2, n-2) s_{\hat{\mu}_{y_h}}. \tag{3.19}$$

Example 16 (Grade Point Average)

From Examples 4, 6 and 8 we have, $n = 120$, $\bar{x} = 24.725$, $\hat{\beta}_0 \approx 2.114$, $\hat{\beta}_1 \approx 0.03883$, $MSE \approx 0.3883$ and $SS_{xx} = \sum_{i=1}^{120}(x_i - \bar{x})^2 = 2379.925$. Suppose, we wish to estimate the average GPA of students with an ACT score of $x_h = 23$. The estimate is,

$$\hat{\mu}_{y_h} = \hat{\beta}_0 + \hat{\beta}_1 x_h = 2.114 + (0.03883)(23) \approx 3.007.$$

Also, the estimate of the variance of the sampling distribution of $\hat{\mu}_{y_h}$ is,

$$s^2_{\hat{\mu}_{y_h}} = \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] MSE = (0.3883)\left[\frac{1}{120} + \frac{(23 - 24.725)^2}{2379.925}\right] \approx 0.00372.$$

Thus, the standard error of $\hat{\mu}_{y_h}$ is $s_{\hat{\mu}_{y_h}} = \sqrt{0.00372} \approx 0.061$. At $\alpha = 0.05$, $t(0.975, 118) = 1.9803$. Then, a 95% confidence interval for μ_{y_h} are,

$$\hat{\mu}_{y_h} \pm t(1-\alpha/2, n-2) s_{\hat{\mu}_{y_h}} = 3.007 \pm (1.9803)(0.061).$$

That is, $2.886 \leq \mu_{y_h} \leq 3.128$.

Example 17 (Copier Maintenance)

For the copier maintenance data, $n = 45$, $\bar{x} = 5.111$, $\hat{\beta}_0 \approx -0.5802$, $\hat{\beta}_1 \approx 15.035$, $MSE \approx 79.4506$ and $SS_{xx} = \sum_{i=1}^{120}(x_i - \bar{x})^2 = 340.4444$ (see Examples 5, 7 and 9). Suppose, we wish to estimate the average minutes a service technician spends at a location with $x_h = 9$ copiers. The estimate is,

$$\hat{\mu}_{y_h} = \hat{\beta}_0 + \hat{\beta}_1 x_h = -0.5802 + (15.035)(9) \approx 134.735.$$

Also, the estimate of the variance of the sampling distribution of $\hat{\mu}_{y_h}$ is,

$$s^2_{\hat{\mu}_{y_h}} = \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] MSE = (79.4506)\left[\frac{1}{45} + \frac{(9 - 5.111)^2}{340.444}\right] \approx 5.295.$$

Therefore, the standard error of $\hat{\mu}_{y_h}$ is $s_{\hat{\mu}_{y_h}} = \sqrt{5.295} \approx 2.301$. At $\alpha = 0.05$, $t(0.975, 43) = 2.0167$. Then, a 95% confidence interval for μ_{y_h} are,

$$\hat{\mu}_{y_h} \pm t(1-\alpha/2, n-2) s_{\hat{\mu}_{y_h}} = 134.735 \pm (2.0167)(2.301).$$

That is, $130.094 \leq \mu_{y_h} \leq 139.375$.

3.4 Simultaneous estimation of mean responses

We mentioned earlier that it may be of interest to estimate the average GPA of students with low, medium and high ACT scores to determine whether students with high ACT scores outperform students with medium or low ACT scores, in general. When the mean response at a single ACT score is estimated with an interval estimate for a fixed significance level α, we can be certain that the confidence level is $(1-\alpha)100\%$. This is however not the case when we estimate mean responses simultaneously at two or more ACT scores with interval estimates. We can use the laws of probability to show that the combined or family error rate, say α^* of all the intervals will be at least α. That is, $\alpha^* \geq \alpha$. This means that the actual or true confidence level of all the interval estimates $(1-\alpha^*)100\%$ will be smaller than the desired confidence level $(1-\alpha)100\%$. Some approaches have been developed in the literature in order to ensure that the confidence level of simultaneous confidence intervals is at least equal to the desired level. You will notice that the structure of the intervals remain the same as in (3.19),

$$\hat{\mu}_{y_h} \pm t(1-\alpha/2, n-2)s_{\hat{\mu}_{y_h}},$$

except that the critical value $t(1-\alpha/2, n-2)$ was replaced. Two of the methods are as follows.

Working-Hotelling Method: This method recommends to compute the simultaneous confidence intervals by replacing $t(1-\alpha/2, n-2)$ with $\sqrt{2F(1-\alpha, 2, n-2)}$, for fixed α. Thus, the Working-Hotelling simultaneous confidence interval for two or more mean responses is given by

$$\hat{\mu}_{y_h} \pm s_{\hat{\mu}_{y_h}}\sqrt{2F(1-\alpha, 2, n-2)}, \tag{3.20}$$

where $s^2_{\hat{\mu}_{y_h}}$ is given by (3.18) and $F(1-\alpha, 2, n-2)$ is the $(1-\alpha)100$ percentile of the F-distribution with 2 and $n-2$ degrees of freedom.

Bonferroni Method: Let the number of intervals to be computed be r. This is a popular approach which replaces $t(1-\alpha/2, n-2)$ with $t(1-\alpha/2r, n-2)$. So that the Bonferroni's intervals are given by

$$\hat{\mu}_{y_h} \pm t(1-\alpha/2r, n-2)s_{\hat{\mu}_{y_h}}, \tag{3.21}$$

where $s^2_{\hat{\mu}_{y_h}}$ is given by (3.18) and $t(1-\alpha/2r, n-2)$ is the $(1-\alpha/2r)100$ percentile of the t-distribution with $n-2$ degrees of freedom.

Example 18 (Grade Point Average)

Suppose we wish to compute 95% simultaneous interval estimates of the mean GPA at $x_{h1} = 16$, $x_{h2} = 24$ and $x_{h3} = 33$. Here, $r = 3$,

$$\begin{aligned}
\hat{\mu}_{y_{h1}} &= \hat{\beta}_0 + \hat{\beta}_1 x_{h1} = 2.114 + (0.03883)(16) \approx 2.7353, \\
\hat{\mu}_{y_{h2}} &= \hat{\beta}_0 + \hat{\beta}_1 x_{h2} = 2.114 + (0.03883)(24) \approx 3.0459, \\
\hat{\mu}_{y_{h3}} &= \hat{\beta}_0 + \hat{\beta}_1 x_{h3} = 2.114 + (0.03883)(33) \approx 3.3954.
\end{aligned}$$

The respective estimate of the variance of the sampling distribution of $\hat{\mu}_{y_h}$ are,

$$\begin{aligned}
s^2_{\hat{\mu}_{y_{h1}}} &= \left[\frac{1}{n} + \frac{(x_{h1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right] MSE = (0.3883)\left[\frac{1}{120} + \frac{(16-24.725)^2}{2379.925}\right] \approx 0.01565,\\
s^2_{\hat{\mu}_{y_{h2}}} &= \left[\frac{1}{n} + \frac{(x_{h2}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right] MSE = (0.3883)\left[\frac{1}{120} + \frac{(24-24.725)^2}{2379.925}\right] \approx 0.00332,\\
s^2_{\hat{\mu}_{y_{h3}}} &= \left[\frac{1}{n} + \frac{(x_{h3}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right] MSE = (0.3883)\left[\frac{1}{120} + \frac{(33-24.725)^2}{2379.925}\right] \approx 0.01441.
\end{aligned}$$

At $\alpha = 0.05$, $F(0.95, 2, 118) = 3.073$ and $t(0.9917, 118) = 2.4286$. Therefore, The Working-Hotelling intervals are given by

$$\begin{aligned}
2.7353 &\pm \sqrt{2(3.073)}\sqrt{0.01565}, \quad 3.0459 \pm \sqrt{2(3.073)}\sqrt{0.00332} \quad \text{and}\\
3.3954 &\pm \sqrt{2(3.073)}\sqrt{0.01441}.
\end{aligned}$$

That is, $(2.4251, 3.0455)$, $(2.9030, 3.1888)$ and $(3.0978, 3.6930)$ are the required intervals for $x_{h1} = 16$, $x_{h2} = 24$ and $x_{h3} = 33$, respectively. The Bonferroni intervals are,

$$\begin{aligned}
2.7353 &\pm 2.4286\sqrt{0.01565}, \quad 3.0459 \pm 2.4286\sqrt{0.00332} \quad \text{and}\\
3.3954 &\pm 2.4286\sqrt{0.01441},
\end{aligned}$$

yielding, $(2.4314, 3.0392)$, $(2.9060, 3.1859)$ and $(3.1039, 3.6869)$ for $x_{h1} = 16$, $x_{h1} = 24$ and $x_{h1} = 33$, respectively.

3.5 Prediction of new observation

Here, we consider the problem of predicting a new observation or outcome y that corresponds to a new level of the predictor that has not been observed. For instance, the GPA data does not include GPA scores for students with ACT score of 17 or over 35. The model which has been fitted based on the data available can be used to predict the average GPA scores for these students. For the prediction to be meaningful, we will assume that the model we have fitted continue to be valid for the new level of the predictor and the new observation. Let y_{new} be the new observation we wish to predict at a new level of x denoted by x_{new}. Then, using the fitted model, the predicted value is given by,

$$\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new},$$

with prediction error,

$$e_{new} = y_{new} - \hat{y}_{new}.$$

Note that $\hat{y}_{new}$ was computed using the data from previous measurements or experiments. So, y_{new} and $\hat{y}_{new}$ are independent. This implies that the variance of the prediction error (see equation (3.17)), is

$$\sigma^2_{pred} = V(e_{new}) = V(y_{new}) + V(\hat{y}_{new}) = \left[1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]\sigma^2, \tag{3.22}$$

with the estimate given by

$$s^2_{pred} = MSE\left[1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]. \tag{3.23}$$

To obtain a prediction interval it can be shown that the probability distribution of the statistic

$$t_0 = \frac{e_{new}}{s_{pred}} = \frac{y_{new} - \hat{y}_{new}}{s_{pred}},$$

is the student t-distribution with $n - 2$ degrees of freedom. Then, a $(1 - \alpha)100\%$ confidence interval for y_{new} is

$$\hat{y}_{new} \pm t(1 - \alpha/2, n - 2)s_{pred}.$$

Example 19 (Copier Maintenance)

Suppose that we wish to predict the number of minutes it will take a service person to service 12 copiers at a new location. Here, $x_{new} = 12$. Using the estimated regression function for the copier maintenance data in Example 5, we have that the predicted number of minutes will be

$$\hat{y}_{new} = -0.5802 + 15.035(12) = 179.8398.$$

Since the number of minutes will vary from location to location and depend on the amount of work required, it will be best to obtain a prediction interval. Now, the variance estimate of the prediction error is (see Examples 7 and 9),

$$\begin{aligned} s^2_{pred} &= MSE\left[1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = 79.4506\left[1 + \frac{1}{45} + \frac{(12 - 5.111)^2}{340.4444}\right] \\ &\approx 92.292. \end{aligned}$$

That is, the standard error of the prediction error is $s_{pred} = \sqrt{92.292} = 9.6069$. At, $\alpha = 0.02$, $t(1 - \alpha/2, n - 2) = t(0.99, 43) = 2.416$. Then, a 98% prediction interval for y_{new} is,

$$\hat{y}_{new} \pm t(1 - \alpha/2, n - 2)s_{pred} = 179.8398 \pm (2.416)(9.6069).$$

That is, $156.627 \leq y_{new} \leq 203.052$.

3.6 Simultaneous prediction intervals for new observations

Simultaneous prediction intervals for, say r new observations are similar in structure to the simultaneous intervals for mean responses we have discussed in Section 3.4 with minor differences. To be specific, the intervals are defined below.

Working-Hotelling Method: The Working-Hotelling simultaneous prediction interval for r new observations is given by

$$\hat{y}_{new} \pm s_{pred}\sqrt{rF(1-\alpha, r, n-2)}, \tag{3.24}$$

where s_{pred}^2 is given by (3.23) and $F(1-\alpha, r, n-2)$ is the $(1-\alpha)100$ percentile of the F-distribution with r and $n-2$ degrees of freedom.

Bonferroni Method: The Bonferroni simultaneous prediction interval for r new observations are given by

$$\hat{y}_{new} \pm t(1-\alpha/2r, n-2)s_{pred}, \tag{3.25}$$

where s_{pred}^2 is given by (3.23) and $t(1-\alpha/2r, n-2)$ is the $(1-\alpha/2r)100$ percentile of the t-distribution with $n-2$ degrees of freedom.

Example 20 (Copier Maintenance)

Suppose that we wish to predict the number of minutes it will take a service person to service 12, 15 and 20 copiers at a new location. Then, $x_{new1} = 12$, $x_{new2} = 15$ and $x_{new3} = 20$. Using the estimated regression function for the copier maintenance data in Example 5, we have that the predicted number of minutes for the three new observations will be

$$\begin{aligned}
\hat{y}_{new1} &= -0.5802 + 15.035(12) = 179.8398, \\
\hat{y}_{new2} &= -0.5802 + 15.035(15) = 224.9448, \\
\hat{y}_{new3} &= -0.5802 + 15.035(20) = 300.1198.
\end{aligned}$$

The corresponding variance estimates of the prediction error for the 3 new observations are respectively (see Examples 7 and 9),

$$\begin{aligned}
s_{pred1}^2 &= MSE\left[1 + \frac{1}{n} + \frac{(x_{new1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = 79.4506\left[1 + \frac{1}{45} + \frac{(12 - 5.111)^2}{340.4444}\right] \\
&\approx 92.292, \\
s_{pred2}^2 &= MSE\left[1 + \frac{1}{n} + \frac{(x_{new2} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = 79.4506\left[1 + \frac{1}{45} + \frac{(15 - 5.111)^2}{340.4444}\right] \\
&\approx 104.0383, \\
s_{pred3}^2 &= MSE\left[1 + \frac{1}{n} + \frac{(x_{new3} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = 79.4506\left[1 + \frac{1}{45} + \frac{(20 - 5.111)^2}{340.4444}\right] \\
&\approx 132.9509.
\end{aligned}$$

That is, the standard error of the prediction errors are $s_{pred1} = \sqrt{92.292} = 9.6069$, $s_{pred2} = \sqrt{104.0383} = 10.19992$ and $s_{pred3} = \sqrt{132.9509} = 11.53043$. At, $\alpha = 0.02$, $t(1-\alpha/6, n-2) = t(0.9967, 43) = 2.8511$ and $F(1-\alpha, r, n-2) = F(0.98, 3, 43) = 3.6386$.

Then, a 98% Working-Hotelling simultaneous prediction interval for y_{new1}, y_{new2} and y_{new3} are,

$$\begin{aligned}
\hat{y}_{new1} \pm s_{pred1}\sqrt{rF(1-\alpha,r,n-2)} &= 179.8398 \pm (9.6069)\sqrt{(3)(3.6386)},\\
\hat{y}_{new2} \pm s_{pred2}\sqrt{rF(1-\alpha,r,n-2)} &= 224.9448 \pm (10.19992)\sqrt{(3)(3.6386)},\\
\hat{y}_{new3} \pm s_{pred3}\sqrt{rF(1-\alpha,r,n-2)} &= 300.1198 \pm (11.53043)\sqrt{(3)(3.6386)}.
\end{aligned}$$

That is, $148.0998 \leq y_{new1} \leq 211.5798$, $191.2454 \leq y_{new2} \leq 258.6442$ and $262.0245 \leq y_{new3} \leq 338.2151$.

A 98% Bonferroni simultaneous prediction interval for y_{new1}, y_{new2} and y_{new3} are,

$$\begin{aligned}
\hat{y}_{new1} \pm s_{pred1}t(1-\alpha/2r,n-2) &= 179.8398 \pm (9.6069)(2,8511),\\
\hat{y}_{new2} \pm s_{pred2}t(1-\alpha/2r,n-2) &= 224.9448 \pm (10.19992)(2,8511),\\
\hat{y}_{new3} \pm s_{pred3}t(1-\alpha/2r,n-2) &= 300.1198 \pm (11.53043)(2.8511).
\end{aligned}$$

That is, $152.4494 \leq y_{new1} \leq 207.2302$, $195.8635 \leq y_{new2} \leq 254.0261$ and $267.2451 \leq y_{new3} \leq 332.9945$.

3.7 Descriptive measures of linear association between x and y

In Section 3.1 we discussed the F test for β_1 and partitioned the total variation (SST) in the observations into the variation explained, captured or accounted for by the fitted regression line (SSR) and the unexplained variation (SSE). Thus, SSR measures the reduction in the uncertainty in predicting y which is usually expressed as a proportion of the total variation as,

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \geq 0.$$

Since, $0 \leq SSR \leq SST$, it follows that $0 \leq R^2 \leq 1$. The measure, R^2 is referred to as the **coefficient of determination**. It is the proportion of the total variation explained by fitting a regression model to y with x as the predictor. In general, the closer the value of R^2 is to 1, the better the fit. We wish to note that if the relationship between x and y is nonlinear, it is possible for the value of R^2 to be close to 1 even when a linear model is not the appropriate model for the data. Thus, no single descriptive measure of linear association can capture the essential information as to whether a given regression model is useful in any particular application. This decision has to be made based on a combination of tests, diagnostics and other measures.

Now, from equation (3.10) and (3.11), we have that

$$R^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1^2\sum_{i=1}^n(x_i-\bar{x})^2}{\sum_{i=1}^n(y_i-\bar{y})^2}.$$

Replacing $\hat{\beta}_1$ with the expression in equation (2.7) and simplifying, we find that R^2 can be computed directly as

$$R^2 = \left[\frac{\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^n(y_i-\bar{y})^2}}\right]^2 \tag{3.26}$$

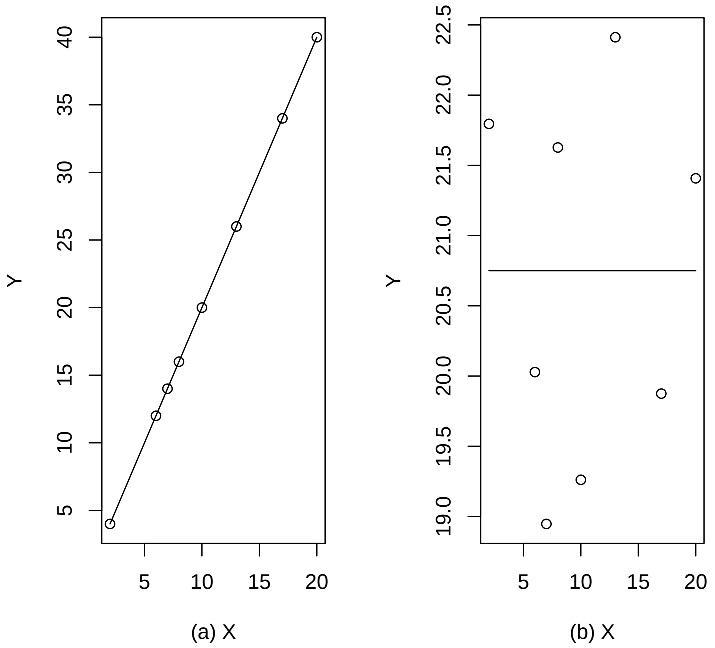


Figure 7: Example of Regression plots when (a) $R^2 = 1$ and (b) $R^2 = 0$

Remarks

1. When the regression line passes through all the points on the scatter plot (perfect straight line relationship between x and y), the unexplained variation is zero. That is, $SSE = 0$ and $SST = SSR$. In this case, $R^2 = 1$. See Figure 7(a).

2. When the regression line is parallel to the x-axis, it means $\hat{y}_i = \bar{y}$ and $SSR = 0$. In this case, $R^2 = 0$. See Figure 7(b).

3. When both x and y are random, the square root of R^2, denoted by $r = \pm\sqrt{R^2}$, is referred to as **the correlation coefficient**. From (3.26), we see that,

$$r = \pm\sqrt{R^2} = \pm\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{3.27}$$

The correlation coefficient is also a measure of the degree of linear association between x and y. If y increases as x increases the slope of the regression line will be positive and $r > 0$. If y decreases as x increases the slope of the regression line will be negative and $r < 0$.

Example 21 (Grade Point Average)

In Example 12, we constructed the ANOVA table for the data on grade point average shown below.

Table 4. Analysis of variance table for GPA data

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F	p-value
Regression	3.5883	1	3.5883	9.2417	0.002914
Error	45.8171	118	0.3883		
Total	49.4054	119			

From the ANOVA table we have that, $SSR = 3.5883$, $SST = 49.4054$. Therefore, based on this data

$$R^2 = \frac{3.5883}{49.4054} = 0.0726$$

That is, only 7.26% of the total variation in GPA was accounted for by the model with ACT scores as predictor. Thus, the degree of linear association between ACT scores and GPA is very weak and based on this measure, ACT score will be a poor predictor of student's GPA.

Example 22 (Copier Maintenance)

For the copier maintenance data, the ANOVA table below shows that $SSR = 76957.9$ and $SST = 80376.8$.

Table 5. Analysis of variance table for copier maintenance data

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F	p-value
Regression	76957.9	1	76957.9	967.91	0
Error	3418.92	43	79.51		
Total	80376.8	44			

Therefore,

$$R^2 = \frac{76957.9}{80376.8} = 0.9575.$$

That is, 95.75% of the total variation in minutes spent by a service person was captured by the model with number of copiers serviced as the predictor. This is an indication that there is a very strong degree of linear association between minutes spent by a service person at a location and the number of copiers serviced. Thus, based on this measure, number of copiers serviced is a good predictor of minutes spent by a service person at a location.

Statistics 6519 - Regression Analysis
Lecture Notes (Contd)

4 Model Adequacy Checking and Remedial Measures

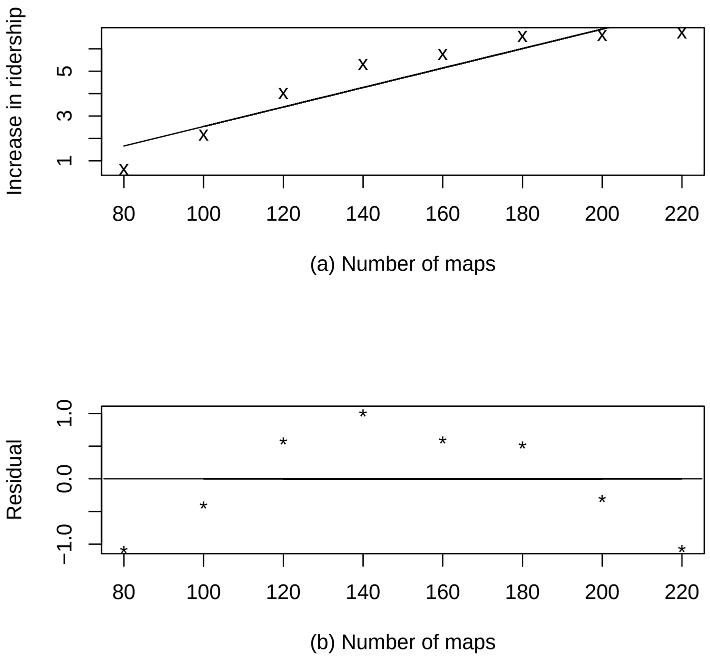


Figure 8: Graph illustrating residual plot in (b) is more effective than scatter plot in (a) for checking appropriateness of linear model

In this section we will be discussing methods for verifying the validity of the assumptions underlying the linear regression model of Section 2. The assumptions we will consider are as follows.

1. Linearity of regression function.

2. The assumption of constant error variance.

3. Independence of error terms.

4. The error terms are normally distributed.

We will also briefly discuss how to verify whether

1. the data contains unusually small or large observations with the potential to influence the model. These are possible outliers or influential observations.

2. one or several important predictor variables have been omitted from the model.

The residuals we discussed in Section 2.2 will be used throughout this section. We will be using graphical and formal statistical tests to examine the model assumptions listed above. The residual plots we will be using include

1. Plot of residuals against predictor variable.

2. Plot of absolute or squared residuals against predictor variable.

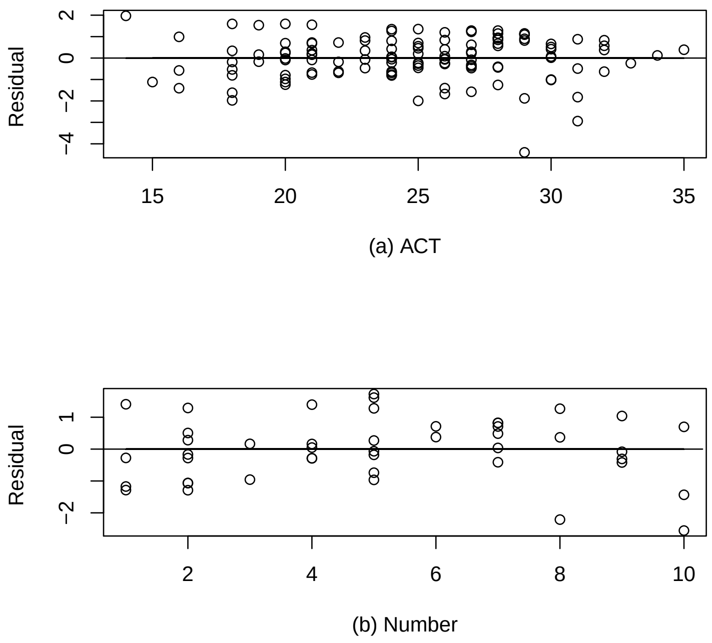
3. Plot of residuals against fitted values.

4. Plot of residuals against time or otl1er sequence.

5. Plots of residuals against omitted predictor variables.

6. Box plot of residuals.

7. Normal probability plot of residuals.



(a) ACT

(b) Number

Figure 9: A plot of residual against predictor variable for (a) GPA data, (b) Copier maintenance data

Departures from the assumptions are sometimes more pronounced and easier to identify if the standardized residuals, $e_i/\sqrt{Var(e_i)}$, are used instead of the residuals, e_i. We will show later that $Var(e_i)$ is not a constant and has a slightly complicated structure. For the purpose of residual analysis, the variance of e_i can be approximated by $\hat{\sigma}^2 = MSE$ since the variance of the error terms ϵ_i is σ^2. The approximately standardized residuals

$$e_i^* = \frac{e_i}{\sqrt{MSE}},$$

are then referred to as the *semistudentized residuals.*

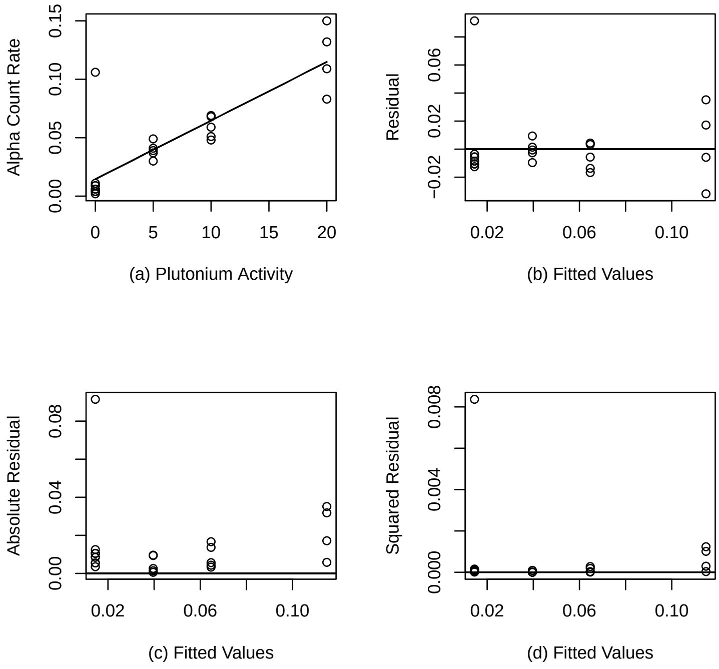


Figure 10: Residual and scatter plots plutonium measurements at a restoration site for diagnosing nonconstant variance

Linearity of Regression Function

Graphical method: In Section 1, we mentioned that a scatter plot of the data can help us decide whether a straight line or linear model is suitable for the data under study. The scatter plot of the data is sometimes not as effective as a residual plot. A very useful graph for checking whether a linear or straight line function is appropriate a given data is a plot of the residuals e_i against the predictor variable x_i or a plot of the residuals e_i against the fitted values $\hat{y}_i$. For the purpose of illustration, let x = number of bus transit maps distributed free to residents of a city at the beginning of a test period, and y = increase during the test period in average daily bus ridership during nonpeak hours. A scatter plot of the data is shown in Figure 8(a). Figure 8(b) is the plot of the residuals from the fitted linear regression model against x. In this example, the fact that the fitted linear regression model is not appropriate is more pronounced in the residual plot in Figure 8(b). The points on the residual plot in Figure 8(b) alternate from negative values to positive values and then back to negative values clearly indicating a curvilinear relationship between x and y. The nonlinearity in the relationship between x and y is however not so clear in the scatter plot shown in Figure 8(a). A plot of the residuals against the fitted values would produce a similar pattern which can also be used for the same purpose. In Figure 9, we show a plot of the residuals versus the predictors for the fitted linear regression models in Examples 4 and 5 in Section 2 for the GPA and copier maintenance data. The residuals in Figure 9 fall within a horizontal band centered around the 0 line, displaying no systematic pattern of positive and negative values. This shows that the assumption of linearity is valid for these data sets.

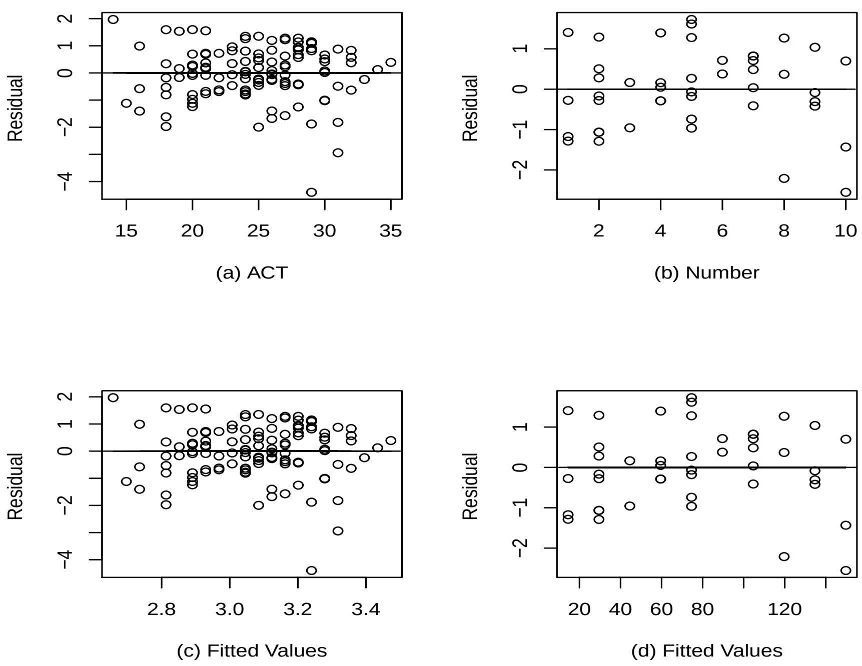


Figure 11: Semistudentized residual plots for diagnosing nonconstant variance in models for GPA and copier maintenance data

NonConstant Variance

Graphical method: Plots of the residuals against the predictor variable or against the fitted values are also helpful in determining whether the assumption of constant variance of the error terms is valid. One can also use plots of the absolute values of the residuals or of the squared residuals against the predictor variable x or against the fitted values $\hat{y}$ for diagnosing nonconstancy of the error variance since the signs of the residuals are not meaningful for examining the constancy of the error variance. These plots are especially useful when there are not many cases in the data set because plotting of either the absolute or squared residuals places all of the information on changing magnitudes of the residuals above the horizontal zero line so that one can more readily see whether the magnitude of the residuals (irrespective of sign) is changing with the level of x or $\hat{y}$. In Figure 10, we display these plots for data on plutonium measurements at a restoration site. The predictor variable x is the plutonium activity measured in pCi/g and the response or outcome variable is the alpha count rate measured in #/sec. The data can be found in Table 3.10 of the course text. A simple linear regression model, shown in Figure 10(a) was fitted to the data and the residuals and fitted values were computed. Figure 10(b), (c), (d) are graphs of the residual, absolute value of residual and square of the residual against the fitted values, respectively. The graphs show that the residuals become more spread out for locations with higher plutonium activity. Since the relationship between x and y is positive, this suggests that the error variance is larger for locations with higher plutonium activity than for locations with lower plutonium activity. Thus, for this data the residual plots indicate that the assumption of constant variance is not valid.

On the contrary, the points in the residual plots in Figure 11 for the GPA and copier maintenance models fall within a horizontal band centered around the 0 line, displaying no systematic pattern. This is an indication that the assumption of constant variance is valid for these data sets.

Hypothesis Testing: Students may have noticed that the interpretation of residual plots is sometimes challenging and in some cases vary from person to person. When residual plots do not lead to clear and precise interpretations, it may be necessary to conduct a formal statistical test for nonconstant variance. There are a few tests for nonconstant variance in the literature, we will however consider only the Breusch-Pagan test. The assumptions of the test are that the error terms are, (a) independent, (b) normally distributed, and (c) have variances σ_i^2 that increases exponentially as the predictors x_i increases. That is, x_i is related to σ_i^2 as follows,

$$ln\sigma_i^2 = a + bx_i.$$

We note that if the variances are constant, then $b = 0$. Thus, for constant variance, we test the hypotheses, $H_0 : b = 0$ against $H_a : b \neq 0$. The test statistic is given by

$$\chi_0 = \frac{n^2}{2} \cdot \frac{SSR^*}{SSE^2}. \tag{4.1}$$

In (4.1), SSR^* is the regression sum of squares for the model

$$e_i^2 = \alpha_0 + \alpha_1 x_i + u_i,$$

where u_i are the error terms, e_i are the residuals from fitting (2.1), and SSE is the residual sum of squares for the linear regression model (2.1). If n is sufficiently large, χ_0 follows approximately a chi-square distribution with 1 degree of freedom. H_0 is rejected if for fixed α, $\chi_0 > \chi^2(1-\alpha, 1)$ where $\chi^2(1-\alpha, 1)$ is the $(1-\alpha)100$ percentile of the chi-square distribution with 1 degree of freedom.

Example 23 (Copier Maintenance)

Using the R software we fit a simple linear regression model to the copier maintenance data. A summary of the output from the fitted model from the R software is shown below.

```
Call:
lm(formula = Minutes   Number)

 Residuals:
       Min        1Q   Median       3Q       Max
  -22.7723   -3.7371   0.3334   6.3334   15.4039

 Coefficients:
              Estimate   Std. Error   t value   Pr(>—t—)
 (Intercept)   -0.5802       2.8039    -0.207   0.837
 Number        15.0352       0.4831    31.123   <2e-16 ***
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.914 on 43 degrees of freedom
Multiple R-squared: 0.9575, Adjusted R-squared: 0.9565
F-statistic: 968.7 on 1 and 43 DF, p-value: < 2.2e-16

Analysis of Variance Table


Response: Minutes
             Df   Sum Sq   Mean Sq   F value   Pr(>F)
 Number       1    76960     76960    968.66   < 2.2e-16 ***
 Residuals   43     3416        79
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

From the ANOVA table we have that $SSE = 3416$. Using the squared residuals from this fit we use the R software again to fit a SLR model to the squared residuals with Number of copiers serviced x, as predictor. The regression sum of squares from this fit is $SSR^{*} = 15155$. It follows that,

$$\chi_0 = \frac{n^2}{2} \cdot \frac{SSR^{*}}{SSE^2} = \frac{45^2}{2} \cdot \frac{15155}{3416^2} \approx 1.315.$$

At $\alpha = 0.05$, we have $\chi^2(0.95, 1) \approx 3.841$ with $p - value \approx 0.251$. It follows that evidence in the data is strongly in favour of H_0. That is, the error terms satisfy the constant variance assumption.

Example 24 (Plutonium Activity - Figure 10)

We saw earlier that the residual plots in Figure 10 for the plutonium activity data indicates that the constant variance assumption was not satisfied. We now use the Breusch-Pagan test to examine whether the conclusion will be the same. Using the plutonium activity data with $n = 24$, we repeat the same process as in Example 23 to obtain $SSE = 0.012371$ and $SSR^* = 3.99 \times 10^{-7}$. Therefore, for this data

$$\chi_0 = \frac{n^2}{2} \cdot \frac{SSR^*}{SSE^2} = \frac{24^2}{2} \cdot \frac{3.99 \times 10^{-7}}{0.012371^2} \approx 0.751.$$

At $\alpha = 0.05$, we have $\chi^2(0.95, 1) \approx 3.841$ with $p - value \approx 0.386$. That is we cannot reject H_0 suggesting that the constant variance assumption was not violated by the error terms. We note that the Breusch-Pagan test is a large sample test. Here, we have only 24 observations. Thus, the conclusions from this test may not be reliable since the sample size is small. This may account for the contradictory conclusions between the graphical method and the test.

Outlier Detection

Graphical method: There are many sophisticated procedures that have been developed for identification of outliers (unsually small or large observations when compared to the rest of the data). Since this is a beginners course on regression, we will only consider elementary procedures that depend only on residual plots. Identification of outliers is particularly important because they have the potential to influence the values of estimated parameters and hence the fitted regression line. In Figure 10, we notice a point in the graph that is particularly far away from the rest of the points in the y direction. This point can be viewed as an outlier in the y direction. It is about five (5) standard deviations away from zero. In Figure 10(a) we see that despite the presence of this point, the regression line appear to fit the bulk of the observations well. The outlier did not pull the line away from the bulk of the observations. So, this outlier may not be a highly influential outlier. It probably influenced the values of the parameter estimates slightly but not by much. Of great concern are those outliers that are highly influential. Highly influential outliers tend to pull the fitted regression line towards the outlying point and away from the rest of the points. Thus, it is common for analyst to discard or remove such points from the data during analysis. Sometimes, outliers contain important information about the process that produced the data. It may be an indication that the process was out of control. Therefore, it is not good practice to simply discard or remove outliers from a data set. A safe rule that has been suggested is to discard an outlier only if there is direct evidence that it represents an error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance.

There exist several methods in the literature, commonly referred to as **robust methods** for estimation when outliers are present in data. These methods are beyond the scope of this course. In this class, to determine if an outlier is influential we will fit the line with the outlier present in the data and refit the model after removing the outlier. We then determine the effect of the outlier by comparing the values of the estimates. A highly influential outlier will cause the values of the estimates to be altered drastically whereas an outlier that is not influential will lead to minimal or no change in the values of the estimates. Plotting of semistudentized residuals is particularly helpful for distinguishing outlying observations. This plot makes it easy to identify residuals that lie many standard deviations from zero.

A rough rule of thumb when the number of cases is large is to consider semistudentized residuals with absolute value of four or more to be outliers. A boxplot of the residuals is also useful for outlier detection.

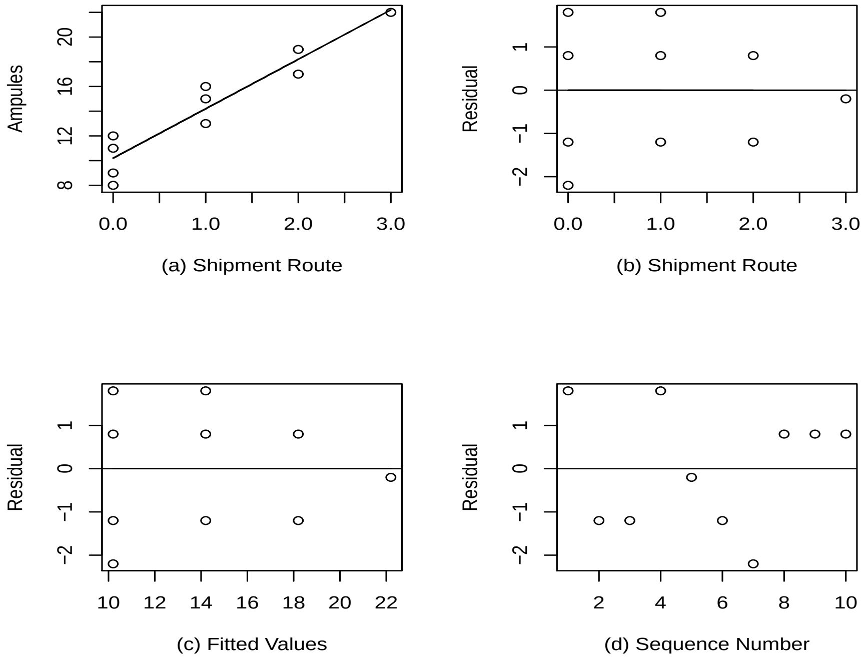


Figure 12: Example of residual plots for diagnosing nonconstant variance and nonindependence

Independence of Error Terms

Graphical method: If the sequence in which the observations were measured is known, one can construct a plot of the residuals against the sequence number, in order to examine the assumption of independence of the error terms. If the sequence is unknown, interpretation of such a plot will be meaningless and unreliable. The fact that the observations were measured sequentially may cause adjacent observations to be dependent which may in turn result in serial dependence and correlation between the observations, in particular, if the measurements were taken sequentially in time. The dependence between observations will cause a systematic pattern in the sequence plot such as an alternating pattern between negative and positive values of residuals as shown in Figure 12(d). This is sometimes referred to as cyclical nonindependence. The dependence between observations can also result in a linear trend in the residuals where all the residuals fall on a straight line on the sequence plot.

Normality of Error Terms

Two useful residual plots for examining the distribution of the error terms are the boxplot of residuals and the normal probability or quantile-quantile (Q-Q) plot of residuals. The data

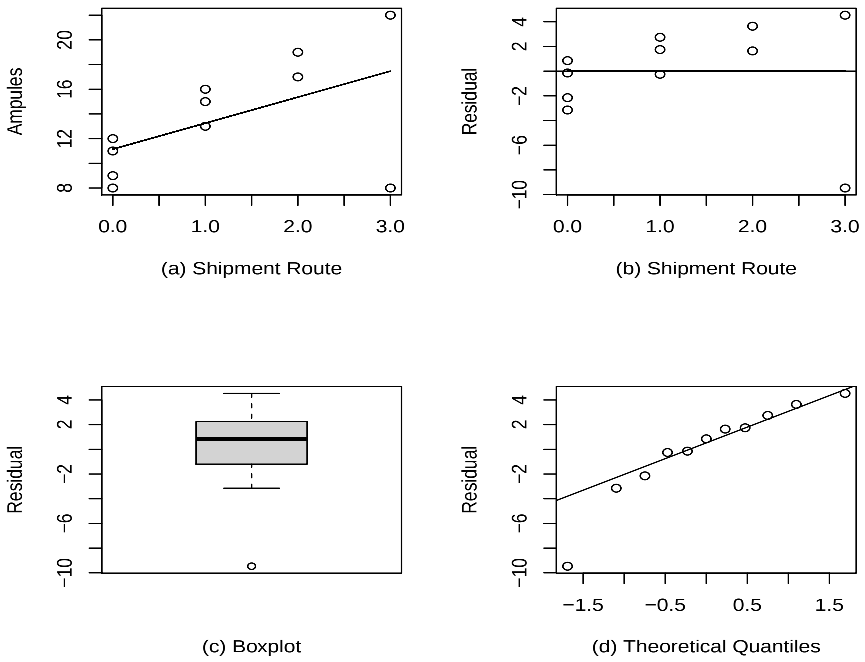


Figure 13: Example of residual plots for diagnosing normality of error terms

used in producing the plots in Figure 13 is exactly the same data used in Figure 12 except that one influential outlier in the x direction was added to the data. An outlier in the x direction is said to be a leverage point. The effect of this leverage point on the analysis is evident in the plots in Figure 13 when compared to the plots in Figure 12. In Figure 13(a), the leverage point pulls the fitted regression line downwards towards the outlying observation and causes the line to be a poor fit for the bulk of the data. The leverage point also alters the structure of the systematic pattern in the plot of the residuals versus the shipment route in Figure 13(b) when compared to the pattern in Figure 12(b). There is however still an indication of nonconstant variance. The boxplot in Figure 13(c) clearly shows the presence of an outlier in the data. The boxplot appear to indicate that the distribution of the observation is approximately symmetric with a median close to zero. In terms of normality, the points on the normal probability plot in Figure 13(d) fall approximaetely on a straight line aside from the outlying point. This is an indication that the error terms are approximately normally distributed.

4.1 Lack of Fit Test

So far, we have assumed that a simple linear regression model is an appropriate model for the relationship between an outcome variable y and a predictor variable x. That means we are assuming that the structure of the mean of y_i at the predictor value x_i is

$$\mu_i = \beta_0 + \beta_1 x_i.$$

Note that the assumption of a linear relationship may be wrong in which case $\mu_i \neq \beta_0+\beta_1 x_i$. In general, the true structure of μ_i is unknown. So, in this section we will say that the unspecified μ_i is the full model and simply write,

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots, n.$$

That is, we are declaring that we do not know the true structure of μ_i. Now, to examine whether the linear model is a good fit for the data, the null hypothesis for the test becomes $H_0 : \mu_i = \beta_0 + \beta_1 x_i$. This is the test that is referred to as the Lack of Fit test.

In order to conduct this test, repeated observations at some values of x are required. These repeated observations will be called replicates. For instance, 3 observations y_{11}, y_{21} and y_{31} may be measured at x_1; 2 observations y_{13} and y_{23} may be measured at x_3 and so on. So, following the notations in the course text, in general, n_j observations, $y_{1j}, y_{2j}, \ldots, y_{n_j,j}$ are measured at x_j, $j = 1, 2, \ldots, c$, where c is the number of distinct x values in the data. To fix ideas, consider the data on solution concentration shown in Table 6. In this example, x = time and y = solution concentration.

Table 6. Repeated Measurements on Solution Concentration

	Solution Concentration				
	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
Replicate	$x_1=1$	$x_2=3$	$x_3=5$	$x_4=7$	$x_5=9$
$i=1$	2.84	1.22	0.49	0.16	0.07
$i=2$	2.57	1.15	0.58	0.17	0.09
$i=3$	3.10	1.07	0.53	0.21	0.08
Total	8.51	3.44	1.6	0.54	0.24
Mean ($\bar{y}_j$)	2.837	1.147	0.533	0.180	0.080

In Table 6, $c = 5$, $x_1 = 1$, $x_2 = 3$, $x_3 = 5$, $x_4 = 7$ and $x_5 = 9$ with $n_1 = n_2 = n_3 = n_4 = n_5 = 3$. It follows that the total number of observations is,

$$n = \sum_{j=1}^{c} n_j = \sum_{j=1}^{5} n_j = 15.$$

Also, as an example $y_{11} = 2.84$ and $y_{34} = 0.21$. In general, we denote the observed responses as y_{ij}, $i = 1, 2, \ldots, n_j$.

Full Model: To construct the test statistic, we first use the method of least squares to fit the full model

$$y_{ij} = \mu_j + \epsilon_{ij}, \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, c.$$

It can be shown that under the full model,

$$\hat{\mu}_j = \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}.$$

Thus, the residual or error sum of squares under the full model is

$$SSE(Full) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} e_{ij}^2 = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2. \tag{4.2}$$

The error sum of squares for the full model is usually called the Pure Error Sum of Squares denoted by $SSPE$. That is $SSPE = SSE(Full)$. Note that under the full model the model parameters we are estimating are $\mu_1, \mu_2, \ldots, \mu_c$. That is, we are estimating c parameters. Thus, the degrees of freedom associated with $SSE(Full)$ or $SSPE$ is

$$df(Full) = n - c.$$

As an example, using the data in Table 6,

$$SSPE = (2.84 - 2.837)^2 + (2.57 - 2.837)^2 + \cdots + (0.09 - 0.08)^2 + (0.08 - 0.08)^2 = 0.1574$$

with correspoding degrees of freedom, $df(Full) = 15 - 5 = 10$. Using the R software the ANOVA table for the full model for the solution concentration data in Table 6 is shown below with the value of $SSPE$ in boldface.

```
                    Analysis of Variance Table

Response: Conc
                   Df   Sum Sq  Mean Sq  F value    Pr(>F)
as.factor(Time)     5  29.0543   5.8109   369.18  5.292e-11 ***
Residuals          10   0.1574   0.0157
```

Reduced Model: The model resulting from assuming that the null hypothesis is true is usually referred to as the reduced model. Recall that when the null hypothesis is true, $\mu_j = \beta_0 + \beta_1 x_j$. That means, we replace μ_j in the full model with $\mu_j = \beta_0 + \beta_1 x_j$ to obtain the reduced model as

$$y_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij}, \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, c.$$

Using the results from Section 2, we obtain

$$SSE(Reduced) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} e_{ij}^2 = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \hat{y}_{ij})^2. \tag{4.3}$$

Since we are estimating only 2 parameters, the degrees of freedom associated with the error sum of squares for the reduced model is

$$df(Reduced) = n - 2.$$

Using the R software, the ANOVA table for the full model for the solution concentration data in Table 6 is shown below.

```
                    Analysis of Variance Table

Response: Conc
                   Df   Sum Sq  Mean Sq  F value    Pr(>F)
Time                1  12.5971   12.597   55.994  4.611e-06 ***
Residuals          13   2.9247    0.225
```

From the ANOVA table above, we see that $SSE(\textit{Reduced}) = 2.9247$. Next, the test statistics for lack of fit is constructed by applying the general linear test approach. The test statistic is given by

$$F_0 = \frac{\frac{SSE(\textit{Reduced}) - SSE(Full)}{df(\textit{Reduced}) - df(Full)}}{\frac{SSE(Full)}{df(Full)}}. \tag{4.4}$$

From the expression for F_0 we deduce that sum of squares for lack of fit (SSLF) is given by

$$SSLF = SSE(\textit{Reduced}) - SSE(Full),$$

with degrees of freedom (df(LOF)),

$$df(LOF) = df(\textit{Reduced}) - df(Full).$$

When H_0 is true, F_0 follows the F-distribution with numerator degrees of freedom $df(\textit{Reduced}) - df(Full) = (n-2)-(n-c) = c-2$ and denominator degrees of freedom $df(Full) = n-c$. The null hypothesis is rejected if, for fixed α, $F_0 > F(1-\alpha, c-2, n-c)$, where $F(1-\alpha, c-2, n-c)$ is the $(1-\alpha)100$ percentile of the F-distribution with $c-2$ and $n-c$ degrees of freedom. The p-value for the test can be computed as,

$$p - \text{value} = 1 - P(F_{c-2,n-c} \leq F_0).$$

Using the results from the ANOVA tables for the full and reduced models for the data in Table 6, we find that

$$F_0 = \frac{\frac{SSE(\textit{Reduced}) - SSE(Full)}{df(\textit{Reduced}) - df(Full)}}{\frac{SSE(Full)}{df(Full)}} = \frac{\frac{2.9247-0.1574}{13-10}}{\frac{0.1574}{10}} = 58.6044,$$

with corresponding p-value,

$$p - \text{value} = 1 - P(F_{3,10} \leq 58.6044) = 1.194 \times 10^{-6}.$$

Based on this result, there is very strong evidence that the simple linear model is not a good fit for the data on solution concentration in Table 6. One can also use the R software to obtain the ANOVA table for the lack of fit test. The table is shown below.

Analysis of Variance Table for lack of fit test

```
Model 1: Conc ˜ Time
 Model 2: Conc ˜ 0 + as.factor(Time)

  Res.Df     RSS  Df  Sum of Sq        F     Pr(>F)
1     13  2.9247
2     10  0.1574   3     2.7673   58.603  1.194e-06 ***
```

The ANOVA table (last row) above shows that $SSLF = 2.7673$, $df(LOF) = 3$, $F_0 = 58.603$ and p-value = 1.194e-06.

4.2 Remedial Measures

So far, we have discussed estimation, test of hypotheses, confidence intervals and model diagnostics for simple linear regression models. It is recommended to conduct model adequacy checking once the model paramters have been estimated and before testing hypothesis. In this way, if the model is found not to be appropriate a decision can be made either to abandon the model and consider alternative models or apply a suitable transformation to the data so that a simple linear model will be an appropriate model for the transformed data. The application of transformations to the data in order to make the SLR model (2.1) to be appropriate is what is referred to as remedial measure.

Nonlinearity of relationship between x and y

If the model diagnostics indicates that the error terms are approximately normal and have constant variance but the regression relationship is not linear, one can consider transforming the predictor x values (not y values). The reason is that transforming the y values will also transform the error terms which may cause the assumptions of normality and constant variance that are already valid to be violated. Let x^* be the transformed x values. Some transformations that can be applied are as follows.

1. Log base 10: $x^* = log_{10}x$.
2. Square root: $x^* = \sqrt{x}$.
3. Exponential: $x^* = exp(x)$ or $x^* = exp(-x)$.
4. Squared: $x^* = x^2$.
5. Inverse: $x^* = \frac{1}{x}$.

It is usually a challenge to determine which of the transformations should be applied in any given situation. Thus, it is common to use a trial and error approach until we find a transformation that works. Once a transformation has been applied to the data, the analysis is repeated using the transformed data (x^*, y). Model diagnostics is then applied to verify whether all assumptions are now valid for the transformed data.

Constant Variance and Normality Assumptions

It has been found that transformations that correct violation of the assumption of normality also correct any violation of the constant variance assumption. These are assumptions that are made on the distribution of the error terms which is the same as the distribution of the outcome variable y. So, violation of these two assumptions is usually corrected by transforming the observed outcomes or responses y only. Students will find that in many cases, departures from normality or unequal variances occur due to increasing skewness and increasing variability of the distributions of the error terms as the mean response or predictor increases. See for example Figures 10 and 12. Some transformations that may help correct violations of these assumptions in these and other cases is the family of power transformations $y^* = y^\lambda$, also commonly referred to as Box-Cox transformations, where λ is a parameter to be determined by the data. Some commonly used transformations are as follows.

1. Natural logarithm: $y^* = log_e y$.
2. Square root: $y^* = \sqrt{y}$.
3. Squared: $y^* = y^2$.
4. Inverse: $x^* = \frac{1}{x}$.
5. Inverse square root: $y^* = \frac{1}{\sqrt{y}}$.

We mentioned earlier that it is common to use a trial and error approach until we find a transformation that works. Once a transformation has been applied to the data, the analysis is repeated using the transformed data (x^*, y^*). Model diagnostics is then applied to verify whether all assumptions are now valid for the transformed data.

Statistics 6519 - Regression Analysis
Lecture Notes (Contd)

5 Matrix Approach To SLR

In this section we will use matrix algebra to discuss the results we have obtained in the previous sections. So, no new regression concepts will be introduced here. The application of matrix algebra is useful because the mathematical expressions are more compact. It is also a more efficient way to derive the results for multiple regression analysis. We will assume that students are familiar with matrix operations such as addition, multiplication, transpose and inversion of matrices. We will also assume that students have applied matrix approach to solving systems of equations in linear algebra.

The matrix equivalent of the simple linear regression model (2.1) comes from writing the model for the response variable y at each value of the predictor variable x in the following way.

$$\begin{array}{lcl} y_1 & = & \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 & = & \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \cdots & = & \cdots\cdots\cdots\cdots \\ \cdots & = & \cdots\cdots\cdots\cdots \\ y_{n-1} & = & \beta_0 + \beta_1 x_{n-1} + \epsilon_{n-1} \\ y_n & = & \beta_0 + \beta_1 x_n + \epsilon_n \end{array} \tag{5.1}$$

The coefficients of β_0 and β_1 then form the column of the matrix $\boldsymbol{X}$ of regressors while the responses on the left hand side and the random errors on the right hand side of the equations in (5.1) are used to construct vectors of responses $\boldsymbol{Y}$ and of error terms $\boldsymbol{\epsilon}$ as shown below. The parameters β_0 and β_1 also form a vector denoted by $\boldsymbol{\beta}$.

$$\boldsymbol{X}_{n\times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ \cdots & \cdots \\ 1 & x_{n-1} \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ \cdots \\ y_{n-1} \\ y_n \end{bmatrix}, \quad \boldsymbol{\epsilon}_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \cdots \\ \epsilon_{n-1} \\ \epsilon_n \end{bmatrix}, \quad \boldsymbol{\beta}_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}. \tag{5.2}$$

The model (2.1) in matrix notations becomes

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5.3}$$

where the random vector $\boldsymbol{\epsilon}$ of error terms is assumed to follow a multivariate normal distribution with a mean vector, $E(\boldsymbol{\epsilon}) = \mathbf{0}$ of zeros and variance-covariance matrix,

$$\boldsymbol{\Sigma}_{n\times n} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2 \boldsymbol{I}_n. \tag{5.4}$$

This implies that the mean of the vector of outcomes $\boldsymbol{Y}$ is,

$$\boldsymbol{\mu_Y} = E(\boldsymbol{Y}) = \begin{bmatrix} E(y_1) \\ E(y_2) \\ \cdots \\ \cdots \\ E(y_{n-1}) \\ E(y_n) \end{bmatrix} = \begin{bmatrix} \mu_{y_1} \\ \mu_{y_2} \\ \cdots \\ \cdots \\ \mu_{y_{n-1}} \\ \mu_{y_n} \end{bmatrix} = \boldsymbol{X\beta}. \tag{5.5}$$

5.1 Least Squares Estimation

In matrix notations, the objective function (2.2), is written as,

$$Q(\beta_0, \beta_1) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}).$$

Differentiating the objective function and simplifying leads to the normal equations (2.5) and (2.6), which we now write as,

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}, \tag{5.6}$$

where,

$$\hat{\boldsymbol{\beta}}_{2\times 1} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}.$$

Solving the normal equations in (5.6), we obtain the least squares estimates,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{5.7}$$

5.2 Estimated Regression Function and Residuals

Once the model parameter vector $\boldsymbol{\beta}$ has been estimated, we can write the fitted values or the estimated regression function as,

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

By substituting the expression for $\hat{\boldsymbol{\beta}}$ from equation (refLSEM) into the above expression for fitted values, we obtain,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{HY},$$

where,

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T,, \tag{5.8}$$

is called the hat matrix. The hat matrix plays a very important and useful role in identification of outliers and has some important and useful properties. It is quite easy and straightforward for students to show that,

$$\boldsymbol{H}^T = \boldsymbol{H}, \quad \text{and} \quad \boldsymbol{H}^2 = \boldsymbol{H}\cdot\boldsymbol{H} = \boldsymbol{H}.$$

That means, the hat matrix is symmetric and idempotent. The residual vector $\boldsymbol{e}$, can then be computed as,

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \tag{5.9}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix. It can also be shown that the matrix $\boldsymbol{I} - \boldsymbol{H}$, is also symmetric and idempotent. In Section 4, we approximated the estimate of the residual variance with the MSE when computing the semistudentized residuals because the residuals are not uncorrelated (as the error terms) and the variances of the residuals are not constant but has a slightly complicated structure. From equation (5.9), it is clear that the variance-covariance matrix of the residual vector is

$$Var(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T). \tag{5.10}$$

We note that the diagonal elements of the matrix (5.10) are the variances of the residuals, while the off-diagonal elements are the covariances. It can be shown that the ith, $i = 1, 2, \ldots, n$, diagonal element of (5.10), which is the variance of e_i, can be written as

$$Var(e_i) = \sigma^2(1 - \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i) = \sigma^2(1 - h_{ii}), \tag{5.11}$$

where $\boldsymbol{x}_i$ is the ith row of the matrix $\boldsymbol{X}$ and $h_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$ is the ith diagonal element of the hat matrix $\boldsymbol{H}$.

Example 25 (Copier Maintenance)

Using the copier maintenance data we illustrate the matrix approach to least squares estimation of the regression parameters we have just discussed. For this data, we have

$$\boldsymbol{X}_{n\times 2} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 2 \\ 1 & 1 \\ \cdots & \cdots \\ \cdots & \cdots \\ 1 & 4 \\ 1 & 5 \end{bmatrix}, \quad \boldsymbol{Y}_{n\times 1} = \begin{bmatrix} 20 \\ 60 \\ 46 \\ 41 \\ 12 \\ \cdots \\ \cdots \\ 61 \\ 77 \end{bmatrix}.$$

Using the R software, we obtain

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 2 & 4 & 3 & 2 & \cdots & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 2 \\ 1 & 1 \\ \cdots & \cdots \\ \cdots & \cdots \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 45 & 230 \\ 230 & 1516 \end{bmatrix},$$

with

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} 0.09895 & -0.01501 \\ -0.01501 & 0.00293 \end{bmatrix}.$$

Also, we find that,

$$\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 2 & 4 & 3 & 2 & \cdots & 4 & 5 \end{bmatrix} \begin{bmatrix} 20 \\ 60 \\ 46 \\ 41 \\ 12 \\ \cdots \\ \cdots \\ 61 \\ 77 \end{bmatrix} = \begin{bmatrix} 3432 \\ 22660 \end{bmatrix}.$$

Then, the least squares estimate of the vector of model parameters becomes,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &= \begin{bmatrix} 0.09895 & -0.01501 \\ -0.01501 & 0.00293 \end{bmatrix} \begin{bmatrix} 3432 \\ 22660 \end{bmatrix} \\ &= \begin{bmatrix} -0.58016 \\ 15.03524 \end{bmatrix}. \end{aligned}$$

That is, $\hat{\beta}_0 \approx -0.5802$ and $\hat{\beta}_1 \approx 15.035$, as before in Example 5. The fitted values and residuals are then computed as,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ \cdots & \cdots \\ \cdots & \cdots \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} -0.58016 \\ 15.03524 \end{bmatrix} = \begin{bmatrix} 29.49034 \\ 59.56084 \\ 44.52559 \\ 29.49034 \\ \cdots \\ \cdots \\ 59.56084 \\ 74.59608 \end{bmatrix},$$

and,

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{Y} = \begin{bmatrix} 20 \\ 60 \\ 46 \\ 41 \\ \cdots \\ \cdots \\ 61 \\ 77 \end{bmatrix} - \begin{bmatrix} 29.49034 \\ 59.56084 \\ 44.52559 \\ 29.49034 \\ \cdots \\ \cdots \\ 59.56084 \\ 74.59608 \end{bmatrix} = \begin{bmatrix} -9.4903394 \\ 0.4391645 \\ 1.4744125 \\ 11.5096606 \\ \cdots \\ \cdots \\ 1.4391645 \\ 2.4039164 \end{bmatrix},$$

respectively.

5.3 Inference for $\boldsymbol{\beta}$

Earlier, we discussed two approaches namely, t-test and ANOVA, for testing for β_1. The test statistic for the t-test was

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}.$$

We now know how to compute $\hat{\beta}_1$ using the matrix approach. In what follows, we will discuss the matrix approach for computing the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ that is required for computing t_0. Recall that, the variance-covariance matrix of $\boldsymbol{Y}$ and the estimate of the parameter vector $\boldsymbol{\beta}$ (see (5.4) and (5.7)) were,

$$\Sigma = \sigma^2 \boldsymbol{I} \quad \text{and} \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$

respectively. So,

$$\boldsymbol{\sigma}^2(\hat{\boldsymbol{\beta}}) = Var(\hat{\boldsymbol{\beta}}) = [(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T][\sigma^2\boldsymbol{I}][(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

This implies that,

$$\boldsymbol{\sigma}^2(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\hat{\beta}_1) \end{bmatrix} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Now, using matrix algebra we can show that

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ x_1 & x_2 & \cdots & x_{n-1} & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix},$$

with

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} & -\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ -\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{bmatrix}.$$

It follows that the covariance matrix of the statistic $\hat{\boldsymbol{\beta}}$ is given by

$$\boldsymbol{\sigma}^2(\hat{\boldsymbol{\beta}}) = \left[\begin{array}{cc} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\hat{\beta}_0) \end{array} \right] = \sigma^2 \left[\begin{array}{cc} \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} & -\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ -\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{array} \right]. \qquad (5.12)$$

By equating elements of the matrices in (5.12), we see that,

$$\begin{aligned} \sigma^2_{\hat{\beta}_0} &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad \sigma(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \sigma^2_{\hat{\beta}_1} &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \end{aligned}$$

as before. To estimate the variances and covariance, we replace σ^2 with MSE.

Example 26 (Copier Maintenance)

Using the results from Examples 13 and 25, we have,

$$\hat{\boldsymbol{\sigma}}^2(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1} = 79.51 \left[\begin{array}{rr} 0.09895 & -0.01501 \\ -0.01501 & 0.00293 \end{array} \right] = \left[\begin{array}{rr} 7.86208 & -1.19279 \\ -1.19279 & 0.23337 \end{array} \right]$$

That is, the estimates of the variances and covariance are,

$$\hat{\sigma}^2_{\hat{\beta}_0} \approx 7.862, \ \hat{\sigma}(\hat{\beta}_0, \hat{\beta}_1) \approx -1.1928, \ \text{ and } \ \hat{\sigma}^2_{\hat{\beta}_1} \approx 0.2334.$$

5.4 Mean response and Prediction Intervals

We note that the mean response in (3.15), can be written as

$$\hat{\mu}_{y_h} = \hat{\beta}_0 + \hat{\beta}_1 x_h = (1 \ \ x_h) \left[\begin{array}{c} \hat{\beta}_0 \\ \hat{\beta}_1 \end{array} \right] = \boldsymbol{x}_h^T \hat{\boldsymbol{\beta}},$$

where, $\boldsymbol{x}_h^T = (1 \ \ x_h)$. The variance of the sampling distribution of $\hat{\mu}_{y_h}$ in (3.17), can then be written as

$$\sigma^2_{\hat{\mu}_{y_h}} = \boldsymbol{x}_h^T \sigma^2(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_h = \sigma^2 \boldsymbol{x}_h^T (\boldsymbol{X}^T\boldsymbol{X})^{-1} \boldsymbol{x}_h. \qquad (5.13)$$

Concerning the variance of the probability distribution of the prediction error in (3.22), we write

$$\sigma^2_{pred} = V(e_{new}) = V(y_{new}) + V(\hat{y}_{new}) = \sigma^2 + \sigma^2 \boldsymbol{x}_h^T (\boldsymbol{X}^T\boldsymbol{X})^{-1} \boldsymbol{x}_h. \qquad (5.14)$$

5.5 Analysis of variance

In Section 3, we decomposed the total variation into two sources of variations. Here, we express those components in matrix notations. First, we recall that the total variation,

$$SST = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2.$$

Now, let $\mathbf{1}^T = (1, 1, \ldots, 1)$ be the unit vector with all elements 1. Then, in matrix notations we can write

$$\sum_{i=1}^{n} y_i^2 = \boldsymbol{Y}^T\boldsymbol{Y}, \quad \text{and} \quad \bar{y} = \frac{1}{n}\mathbf{1}^T\boldsymbol{Y}.$$

Therefore,

$$SST = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \boldsymbol{Y}^T\boldsymbol{Y} - n\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{Y}\right)^T\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{Y}\right).$$

Define the $n \times n$ matrix, $\boldsymbol{J} = \mathbf{1}\mathbf{1}^T$, with elements 1 as,

$$\boldsymbol{J} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}. \tag{5.15}$$

Then, after some matrix algebra we find that the total sum of squares can be written as,

$$SST = \boldsymbol{Y}^T\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}. \tag{5.16}$$

Next, we recall that the error or residual sum of squares is defined as,

$$SSE = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^T\boldsymbol{e}.$$

Using the result in (5.9) and the fact that the marix, $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ is an idempotent matrix, we are then able to write the SSE as,

$$SSE = [(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}]^T[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}] = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}. \tag{5.17}$$

We can also show that the SSE can be written as,

$$SSE = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}. \tag{5.18}$$

Finally, the matrix expression for the regression sum of squares can be obtained by subtraction as,

$$SSR = SST - SSE = \boldsymbol{Y}^T\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} - \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^T\left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}. \tag{5.19}$$

An alternative expression for SSR is given by,

$$\begin{aligned} SSR &= SST - SSE = \boldsymbol{Y}^T\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} \\ &= \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} - \left(\frac{1}{n}\right)\boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y}. \end{aligned} \tag{5.20}$$

Example 27 (Copier Maintenance)

Continuing with the copier maintenance data we will compute the SST, SSR and SSE by the matrix approach. Now,

$$\begin{aligned} \boldsymbol{Y}^T\boldsymbol{Y} &= \begin{bmatrix} 20 & 60 & 46 & \cdots & 61 & 77 \end{bmatrix}\begin{bmatrix} 20 \\ 60 \\ 46 \\ \vdots \\ 61 \\ 77 \end{bmatrix} = 342124, \\ \boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y} &= \begin{bmatrix} 20 & 60 & 46 & \cdots & 61 & 77 \end{bmatrix}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} 20 \\ 60 \\ 46 \\ \vdots \\ 61 \\ 77 \end{bmatrix} = 11778624 \\ \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} &= \begin{bmatrix} -0.5802 & 15.035 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 2 & 4 & 3 & 2 & \cdots & 4 & 5 \end{bmatrix}\begin{bmatrix} 20 \\ 60 \\ 46 \\ 41 \\ 12 \\ \vdots \\ 61 \\ 77 \end{bmatrix} = 338707.6. \end{aligned}$$

It follows that, for the copier maintenance data

$$\begin{aligned} SST &= \boldsymbol{Y}^T\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} = 342124 - \left(\frac{1}{45}\right)11778624 = 80376.8, \\ SSR &= \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} - \left(\frac{1}{n}\right)\boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y} = 338707.6 - \left(\frac{1}{45}\right)11778624 = 76960.42, \\ SSE &= \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} = 342124 - 338707.6 = 3416.4. \end{aligned}$$

Statistics 6519 - Regression Analysis
Lecture Notes (Contd)

6 Multiple Linear Regression (MLR)

In some situations, the response or outcome of an experiment or the dependent variable y may be influenced by more than a single predictor variable. Suppose, y depends on say, $p-1$ predictor variables denoted by, $x_1, x_2, \ldots, x_{p-1}$. Suppose further that n values of the response y, denoted by $y_1, \ldots, y_n$ are measured at n values of the predictor variables. Then, for $i = 1, \ldots, n$, the regression model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \tag{6.1}$$

is said to be a multiple linear regression model or a first-order model with $p-1$ predictor variables. One special case of model (6.1) is the polynomial regression model with $x_1 = x,\ x_2 = x^2, \ldots, x_{p-1} = x^{p-1}$. In this case, only a single predictor variable x with higher orders of x are involved in the model. The model then becomes,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1} + \epsilon_i.$$

The polynomial regression model is also a multiple linear regression model. This example illustrates the fact that the word, "linear", in multiple linear regression does not necessarily mean that there is a linear relationship between y and x_1, x_2, etc. It simply means that the model is linear in the model parameters. In terms of interpretation of the model parameters, β_0 is the intercept on the surface of the regression function and represents the mean value of the regression function when $x_1 = 0, x_2 = 0, \ldots, x_{p-1} = 0$. Again, we emphasize that β_0 has no practical interpretation if any one of the predictor variables $x_1, \ldots, x_{p-1}$ cannot take the value 0. The coefficient of x_k, which is β_k measures the effect of x_k on y when all the other variables are held constant. It also measures the change in the mean response $E(y)$, for every unit increase in x_k when all the other variables are held constant.

The analysis of a multiple linear regression model is handled efficiently when the model is expressed in matrix notations as in (5.3). Following our approach in Section 4, we first rewrite the model (6.1) in the form,

$$\begin{array}{rcl} y_1 & = & \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_{p-1} x_{1,p-1} + \epsilon_1 \\ y_2 & = & \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_{p-1} x_{2,p-1} + \epsilon_i \\ \cdots & = & \cdots\cdots\cdots\cdots\cdots\cdots \\ y_n & = & \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{p-1} x_{n,p-1} + \epsilon_i. \end{array} \tag{6.2}$$

Then, we define the coefficient matrix $\boldsymbol{X}$ and the parameter vector $\boldsymbol{\beta}$ as in (5.2),

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} \quad \boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}. \tag{6.3}$$

Clearly, the observation vector $\boldsymbol{Y}$ and the error vector $\boldsymbol{\epsilon}$ remain the same as in (5.2). The model (6.1) can then be expressed as (5.3),

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where the random vector $\boldsymbol{\epsilon}$ of error terms is assumed to follow a multivariate normal distribution with a mean vector, $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ of zeros and variance-covariance matrix, $\sigma^2\boldsymbol{I}_n$. It then follows that all the results discussed in Section 5 hold for the multiple regression model (6.1), with the appropriate coefficient matrix and parameter vector defined by (6.3). We now use an example to illustrate the application of the results in Section 5 to multiple linear regression problems.

Example 28 (Patient Satisfaction, Problem 6.15, Page 250)

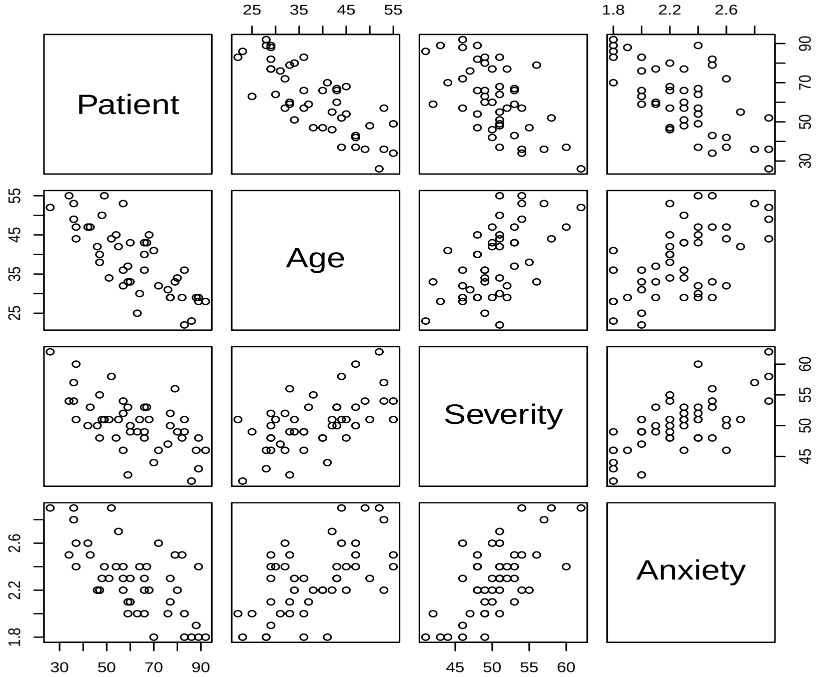


Figure 14: A matrix plot of patient satisfaction data

In order to illustrate the estimation method in MLR, we consider the data on patient satisfaction (Y), patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3 , an index). For this data, we have

$$\boldsymbol{Y}_{n\times 1} = \begin{bmatrix} 48 \\ 57 \\ 66 \\ \vdots \\ 68 \\ 59 \\ 92 \end{bmatrix}, \quad \boldsymbol{X}_{n\times p} = \begin{bmatrix} 1 & 50 & 51 & 2.3 \\ 1 & 36 & 46 & 2.3 \\ 1 & 40 & 48 & 2.2 \\ \vdots & \vdots & \vdots & \\ 1 & 37 & 53 & 2.1 \\ 1 & 28 & 46 & 1.8 \end{bmatrix} \quad \boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

First, we explore the data by constructing the matrix plot shown in Figure 14. The plot shows that as patient's age increases their satisfaction, severity of illness and anxiety level generally decreases indicating a linear relationship between each of the predictor variables and patient satisfaction with negative slope. Thus, one will expect the coefficients of the predictors in the fitted model for this data to be negative. No possible outliers appear to be present in the data as well. Applying the matrix approach to this data, we find that

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 46 & 1766 & 2320 & 105 \\ 1766 & 71378 & 90051 & 4107.2 \\ 2320 & 90051 & 117846 & 5344.7 \\ 105.2 & 4107.2 & 5344.7 & 244.62 \end{bmatrix}, \quad \boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 2832 \\ 103282 \\ 140814 \\ 6327 \end{bmatrix},$$

with

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} 3.2477 & 0.0092 & -0.0679 & -0.0673 \\ 0.0092 & 0.00046 & -0.00032 & -0.00466 \\ -0.0679 & 0.00032 & 0.0239 & -0.01771 \\ -0.673 & -0.00466 & -0.01771 & 0.4982 \end{bmatrix}.$$

Thus, the least squares estimate of the vector of model parameters becomes,

$$\hat{\boldsymbol{\beta}} \;=\; \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 158.4912 \\ -1.1416 \\ -0.4420 \\ -13.4702 \end{bmatrix}.$$

That is, $\hat{\beta}_0 \approx 158.4912$, $\hat{\beta}_1 \approx -1.1416$, $\hat{\beta}_2 \approx -0.4420$ and $\hat{\beta}_3 \approx -13.4702$. For $i = 1, \ldots, 46$, the fitted multiple linear regression model can then be written as,

$$\hat{y}_i = 158.49 - 1.1416x_{i1} - 0.442x_{i2} - 13.4702x_3.$$

The fitted values and residuals are computed as,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 50 & 51 & 2.3 \\ 1 & 36 & 46 & 2.3 \\ 1 & 40 & 48 & 2.2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 37 & 53 & 2.1 \\ 1 & 28 & 46 & 1.8 \end{bmatrix} \begin{bmatrix} 158.4912 \\ -1.1416 \\ -0.4420 \\ -13.4702 \end{bmatrix} = \begin{bmatrix} 47.88707 \\ 66.07965 \\ 61.97621 \\ 67.99068 \\ 83.27364 \\ \vdots \\ 64.53804 \\ 81.94763 \end{bmatrix},$$

and,

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{Y} = \begin{bmatrix} 48 \\ 57 \\ 66 \\ \vdots \\ 68 \\ 59 \\ 92 \end{bmatrix} - \begin{bmatrix} 47.88707 \\ 66.07965 \\ 61.97621 \\ 67.99068 \\ 83.27364 \\ \vdots \\ 64.53804 \\ 81.94763 \end{bmatrix} = \begin{bmatrix} 0.1129334 \\ -9.0796538 \\ 4.0237858 \\ 2.0093153 \\ \vdots \\ -5.5380448 \\ 10.0523698 \end{bmatrix},$$

respectively.

6.1 Estimation of σ^2

Similar to the approach in Section 2.4, the error variance is estimated by the MSE, given by

$$MSE = \frac{SSE}{df_E},$$

where df_E is the error degrees of freedom. In a multiple linear regression model with p predictor variables we estimate p unknown parameters. Thus, the degrees of freedom for error is $n - p$. Furthermore, in Section 5.5 we saw that the residual/error sum of squares can be written as

$$SSE = \boldsymbol{e}^T\boldsymbol{e} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$

It follows that,

$$\hat{\sigma}^2 = MSE = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p} = \frac{\boldsymbol{Y}^{\boldsymbol{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}}{n-p}. \tag{6.4}$$

Example 29 (Patient Satisfaction, Problem 6.15, Page 250)

Using the results from Example 28, we can see that

$$SSE = \boldsymbol{e}^T\boldsymbol{e} = \begin{bmatrix} 0.1129 & -9.0796 & 4.0237 & \cdots & 10.0524 \end{bmatrix} \begin{bmatrix} 0.1129 \\ -9.0796 \\ 4.0238 \\ 2.0093 \\ \vdots \\ 10.0524 \end{bmatrix} = 4248.841.$$

Therefore, for the patient satisfaction data,

$$\hat{\sigma}^2 = MSE = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p} = \frac{4248.841}{46-4} = 101.1629.$$

Next, we construct the ANOVA table for MLR models.

6.2 F Test for Significant Relationship between the response and predictor variables

In MLR analysis the null and alternative hypotheses for testing for a significant relationship between the response and the predictors are stated as,

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \\ H_a &: \text{Some } \beta_k,\ k = 1, \ldots, p-1, \text{ are not zero.} \end{aligned}$$

The test is conducted using the F_0 statistic in the ANOVA table for regression. The components of the ANOVA table are constructed following the expressions in Section 5.5. We now use the patient satisfaction data to illustrate the construction of the ANOVA table.

Example 30 (Patient Satisfaction, Problem 6.15, Page 250)

First, we use (5.15) to compute the components of equations (5.16) and (5.20). That is,

$$\boldsymbol{Y}^T\boldsymbol{Y} = \begin{bmatrix} 48 & 57 & 66 & \cdots & 59 & 92 \end{bmatrix} \begin{bmatrix} 48 \\ 57 \\ 66 \\ \vdots \\ 59 \\ 92 \end{bmatrix} = 187722$$

$$\boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y} = \begin{bmatrix} 48 & 57 & 66 & \cdots & 59 & 92 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 48 \\ 57 \\ 66 \\ \vdots \\ 59 \\ 92 \end{bmatrix} = 8020224$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} &= \begin{bmatrix} 158.49 & -1.14 & -0.44 & -13.47 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 50 & 36 & \cdots & 28 \\ 51 & 46 & \cdots & 46 \\ 2.3 & 2.3 & \cdots & 1.8 \end{bmatrix} \begin{bmatrix} 48 \\ 57 \\ 66 \\ \vdots \\ 59 \\ 92 \end{bmatrix} \\ &= 183473.2. \end{aligned}$$

It follows that, for the copier maintenance data

$$\begin{aligned} SST &= \boldsymbol{Y}^T\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} = 187722 - \left(\frac{1}{46}\right)8020224 = 13369.3, \\ SSR &= \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} - \left(\frac{1}{n}\right)\boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y} = 183473.2 - \left(\frac{1}{46}\right)8020224 = 9120.464, \\ SSE &= \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} = 187722 - 183473.2 = 4248.8. \end{aligned}$$

We now summarize this information in an ANOVA table as shown below. Note that MSR, MSE and F_0 are computed as before. Also, since in the null hypothesis we are considering only $p-1$ parameters, the degrees of freedom for SSR is $p-1$.

Table 5. Analysis of variance table for patient satisfaction data

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F	p-value
Regression	9120.46	3	3040.16	30.05	1.542e-10
Error	4248.8	42	101.2		
Total	13369.3	45			

To test for a significant relationship between patient satisfaction and the predictors, patients' age, severity of illness and anxiety level, we then proceed as follows.

$H_0: \ \beta_1 = \beta_2 = \beta_3 = 0$
H_a : Some $\beta_k, \ k = 1, 2, 3,$ are not zero.

Significance level: $\alpha = 0.05$.

Test Statistic Value:

$$F_0 = \frac{MSR}{MSE} = \frac{3040.16}{101.2} \approx 30.05.$$

p-value: p-value $= 1 - P(F_{3,42} < 30.05)] = 1.541973e - 10$.

Critical value: $F(0.95, 3, 42) \approx 2.827$.

Conclusion: We reject H_0 since $30.05 > 2.827$ (alternatively, $1.541973e - 10 < 0.05$). That means, there is very strong evidence in the data in support of the claim that there is a significant relationship between patient satisfaction and the predictors, patients' age, severity of illness and anxiety level.

Recall that SSR is the amount of variation in the response captured or explained by the fitted regression model or the reduction in the uncertainty in predicting y when the predictors are included in the model. So, when $p > 1$ the **coefficient of multiple determination**,

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST},$$

serves as a measure of association between y and the predictors. As explained earlier, $0 \leq R^2 \leq 1$. We mention that as p increases, the reduction in uncertainty in predicting y becomes larger (*i.e* SSE becomes smaller) due to the increase in the number of predictors. Hence, SSR increases and R^2 becomes larger. This, however does not necessarily mean that a model with more predictor variables is better than a model with fewer predictor variables. Therefore, for the purpose of comparing two models with unequal number of predictor variables in order to determine which model is better the R^2 is adjusted to account for the number of predictor variables in a model. The adjusted R^2, denoted by $R^2_{adjusted}$ is obtained by penalizing the SSE and SST by dividing by their degrees of freedoms as shown in the expression given by,

$$R^2_{adjusted} = 1 - \frac{\frac{SSE}{df_E}}{\frac{SST}{df_T}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}. \tag{6.5}$$

In this way, adding more variables to a model will not automatically increase the value of the adjusted coefficient of multiple determination unless the additional predictor variable(s) contribute significantly to the prediction of the response y.

As an example, the $R^2_{adjusted}$ for the fitted model for the patient satisfaction data is

$$R^2_{adjusted} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST} = 1 - \left(\frac{45}{42}\right)\frac{4248.8}{13369.3} \approx 0.6595,$$

whereas, the unadjusted R^2 value is $R^2 \approx 0.6822$.

6.3 Inferences about model parameters

In Section 5, equation (5.7), we found that

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

Using the fact that, $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$ it can be verified that the mean of the sampling distribution of the estimator $\hat{\boldsymbol{\beta}}$ is,

$$E(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X\beta} = \boldsymbol{\beta}.$$

Again, using matrix notation we show that $\hat{\boldsymbol{\beta}}$ is unbiased. For multiple linear regression models the variance-covariance matrix of the sampling distribution of $\hat{\boldsymbol{\beta}}$ is similar to the expression in Section 5.3. It is a $p \times p$ matrix given by,

$$\sigma^2(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & \cdots & Cov(\hat{\beta}_0, \hat{\beta}_{p-1}) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\hat{\beta}_1) & \cdots & Cov(\hat{\beta}_1, \hat{\beta}_{p-1}) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\hat{\beta}_{p-1}, \hat{\beta}_0) & Cov(\hat{\beta}_{p-1}, \hat{\beta}_1) & \cdots & Var(\hat{\beta}_{p-1}) \end{bmatrix}. \tag{6.6}$$

The covariance matrix (6.6) is estimated by,

$$s^2(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{6.7}$$

Example 31 (Patient Satisfaction, Problem 6.15, Page 250)

For the patient satisfaction data, one can use the R software to verify that the estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ in Example 28, Page 61, is,

$$s^2(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 328.5478 & 0.9328 & -6.8721 & -6.8081 \\ 0.9328 & 0.0461 & -0.0322 & -0.4716 \\ -6.8721 & -0.0322 & 0.2420 & -1.7916 \\ -6.8081 & -0.4716 & -1.7916 & 50.4052 \end{bmatrix}.$$

This implies that the estimated variances of the model parameters are, $s^2_{\hat{\beta}_0} \approx 328.5478$, $s^2_{\hat{\beta}_1} \approx 0.0461$, $s^2_{\hat{\beta}_2} \approx 0.2420$ and $s^2_{\hat{\beta}_3} \approx 50.4052$. Also, the covariances are $s(\hat{\beta}_0, \hat{\beta}_1) \approx 0.9328$, $s(\hat{\beta}_0, \hat{\beta}_2) \approx -6.8721$, $s(\hat{\beta}_0, \hat{\beta}_3) \approx -6.8081$, $s(\hat{\beta}_1, \hat{\beta}_2) \approx -0.0322$, $s(\hat{\beta}_1, \hat{\beta}_3) \approx -0.4716$ and $s(\hat{\beta}_2, \hat{\beta}_3) \approx -1.7916$.

Using the information from the estimated covariance matrix one may wish to examine the significance of the effect of each predictor variable on the response. To test whether the kth, $k = 1, 2, \ldots, p-1$, predictor variable contributes significantly towards the prediction of y, we use the t_0 test statistic,

$$t_0 = \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2_{\hat{\beta}_k}}},$$

to test the null hypothesis $H_0 : \beta_k = 0$, against a two-sided alternative. The null distribution of the test statistic t_0 is the student t-distribution with $n-p$ degrees of freedom. Thus, for fixed α, we reject H_0 if $t_0 > t(1-\alpha/2, n-p)$ or if the p-value given by,

$$p - \text{value} = 2[1 - P(T_{n-p} \leq |t_0|)],$$

is less than α. As before, the sampling distribution of $\hat{\beta}_k$ can then be used to construct confidence intervals for each estimated parameter given by,

$$\hat{\beta}_k \pm t(1-\alpha/2, n-p)\sqrt{s^2_{\hat{\beta}_k}}, \quad k = 1, 2, \ldots, p-1.$$

The procedures for conducting these tests have been illustrated in Examples 8, 9, 10, 11, 14 and 15. Therefore, an example will not be repeated here.

6.4 Interval estimates of mean response and prediction intervals

Suppose, it is required to estimate the mean response at a given value of each of the $p-1$ predictor variables, say, $x_1 = x_{h1}, x_2 = x_{h2}, \ldots, x_{p-1} = x_{h,p-1}$. Following the approach in Section 5.4, we rewrite the mean response at a single set of values of the predictor variables as,

$$\mu_{y_h} = E(y_h) = \begin{bmatrix} 1 & x_{h1} & x_{h2} & \cdots & x_{h,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} = \boldsymbol{x}_h^T \boldsymbol{\beta}, \tag{6.8}$$

where $\boldsymbol{x}_h = (1, x_{h1}, x_{h2}, \cdots, x_{h,p-1})^T$. The variance of the sampling distribution of the estimate of the mean response,

$$\hat{\mu}_{y_h} = \boldsymbol{x}_h^T \hat{\boldsymbol{\beta}},$$

is then given by equation (5.13) of Section 5.4, with the matrix of regressors given by (6.3) and $\boldsymbol{x}_h = (1, x_{h1}, x_{h2}, \cdots, x_{h,p-1})^T$. That is,

$$\sigma^2_{\hat{\mu}_{y_h}} = \boldsymbol{x}_h^T \sigma^2(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_h = \sigma^2 \boldsymbol{x}_h^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_h. \tag{6.9}$$

The estimate of $\sigma^2_{\hat{\mu}_{y_h}}$ is

$$s^2_{\hat{\mu}_{y_h}} = MSE \boldsymbol{x}_h^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_h, \tag{6.10}$$

where the MSE is computed using (6.4). For a fixed value of α, the $(1-\alpha)100\%$ confidence interval for the mean response at $\boldsymbol{x}_h$, μ_{y_h}, can then be computed following previous methods, as

$$\hat{\mu}_{y_h} \pm t(1-\alpha/2, n-p) \sqrt{s^2_{\hat{\mu}_{y_h}}}. \tag{6.11}$$

In the same way, we can show that the variance of the probability distribution of the prediction error in (3.22) is given by,

$$\sigma^2_{pred} = V(e_{new}) = V(y_{new}) + V(\hat{y}_{new}) = \sigma^2 + \sigma^2 \boldsymbol{x}_{new}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_{new}, \tag{6.12}$$

where, again the matrix of regressors is given by (6.3) and $\boldsymbol{x}_{new} = (1, x_{new1}, x_{new2}, \cdots, x_{new,p-1})^T$. The variance of the probability distribution of the prediction error (6.12) is estimated by,

$$s^2_{pred} = MSE \left[1 + \boldsymbol{x}_{new}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_{new} \right], \tag{6.13}$$

with the MSE computed using (6.4). For a fixed value of α, the $(1-\alpha)100\%$ prediction interval for the mean response at a new value $\boldsymbol{x}_{new}$, $\hat{y}_{new}$, can then be computed following previous methods, as

$$\hat{y}_{new} \pm t(1-\alpha/2, n-p) \sqrt{s^2_{pred}}. \tag{6.14}$$

Example 32 (Patient Satisfaction, Problem 6.15, Page 250)

Suppose, we wish to obtain a 90% interval estimate of the mean satisfaction when $x_{h1} = 35, x_{h2} = 45$, and $x_{h3} = 2.2$. In this case,

$$\boldsymbol{x}_h = (1, 35, 45, 2.2)^T.$$

Then, we compute

$$\hat{\mu}_{y_h} = \boldsymbol{x}_h^T \hat{\boldsymbol{\beta}} = \left[\begin{array}{cccc} 1 & 35 & 45 & 2.2 \end{array} \right] \left[\begin{array}{c} 158.4912 \\ -1.1416 \\ -0.4420 \\ -13.4702 \end{array} \right] \approx 69.01.$$

Next, using results from Examples 28 and 31, we compute the variance estimate of the estimated mean response as,

$$\begin{aligned} s^2_{\hat{\mu}_{y_h}} &= MSE\boldsymbol{x}_h^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_h \\ &= 101.2 * \left[\begin{array}{c} 1\ 35\ 45\ 2.2 \end{array}\right] \left[\begin{array}{rrrr} 3.2477 & .0092 & -.0679 & -.0673 \\ .0092 & .0005 & -.0003 & -.0047 \\ -.0679 & .0003 & .0239 & -.0177 \\ -.673 & -.0047 & -.0177 & .4982 \end{array}\right] \left[\begin{array}{r} 1 \\ 35 \\ 45 \\ 2.2 \end{array}\right] \\ &= 7.100156. \end{aligned}$$

Now, at $\alpha = 0.1$ we have that, $t(1-\alpha/2, n-p) = t(0.95, 42) = 1.681952$. It follows that a 90% confidence interval for the mean response at $\boldsymbol{x}_h = (1, 35, 45, 2.2)^T$ is given by

$$\hat{\mu}_{y_h} \pm t(1-\alpha/2, n-p)\sqrt{s^2_{\hat{\mu}_{y_h}}} = 69.01 \pm 1.682 * \sqrt{7.1}.$$

That is, $64.52854 \leq \mu_{y_h} \leq 73.49204$).

If $\boldsymbol{x}_{new} = (1, 35, 45, 2.2)^T$, the prediction interval can be computed in a similar way with

$$\begin{aligned} s^2_{pred} &= MSE + MSE\boldsymbol{x}_h^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_h \\ &= 101.2 \left\{ 1 + \left[\begin{array}{c} 1\ 35\ 45\ 2.2 \end{array}\right] \left[\begin{array}{rrrr} 3.248 & .0092 & -.0679 & -.0673 \\ .0092 & .0005 & -.0003 & -.0047 \\ -.068 & .0003 & .0239 & -.0177 \\ -.673 & -.0047 & -.0177 & .4982 \end{array}\right] \left[\begin{array}{r} 1 \\ 35 \\ 45 \\ 2.2 \end{array}\right] \right\} \\ &= 108.263. \end{aligned}$$

Using this result to compute a 90% prediction interval for y_{new} at $\boldsymbol{x}_{new} = (1, 35, 45, 2.2)^T$, we obtain

$$\hat{y}_{new} \pm t(1-\alpha/2, n-p)\sqrt{s^2_{pred}} = 69.01 \pm 1.682 * \sqrt{108.263}, \qquad (6.15)$$

which leads to, $51.50965 \leq y_{new} \leq 86.51092$.

Statistics 6519 - Regression Analysis
Lecture Notes (Contd)

7 Multiple Linear Regression: Quantitative and Qualitative Predictor Variables

There are many real life situations where variables which cannot be measured numerically can be useful predictors of response variables. Such predictor variables can often be classified into two or several categories. As an example, gender (male or female) cannot be measured numerically but is a powerful predictor variable of many responses. It is well known that, in general, women purchase certain products more than men and there are products more men purchase than women. So, volume of sales of many products can be influenced by gender. Such variables which can only be measured in terms of categories are called qualitative or categorical variables. Since categorical variables can be useful predictors, it is important to know how these variables can be used as predictors in a regression model.

A commonly used approach is to define one numerical variable for each category of the categorical variable. These numerical variables assign the value 0 or 1 to each category depending on whether that category has been observed or not. For instance, for gender we may define numerical variables with the following assignments of 0 and 1, say

$$x_2 = \begin{cases} 1, & \text{if observed gender is female,} \\ 0, & \text{otherwise,} \end{cases} \quad \text{and,} \quad x_3 = \begin{cases} 1, & \text{if observed gender is male,} \\ 0, & \text{otherwise.} \end{cases}$$

The numerical variables x_2 and x_3 defined above are called **indicator** or **dummy** variables. Using the definition above, we notice that if $x_2 = 1$, then the observed gender was female. Also, if $x_2 = 0$, it means a male was observed. It is then clear that the second variable x_3 is not needed in order to determine if a male was observed. As a result, only one indicator variable x_2 should have been defined when the variable has only two categories. This idea can be generalized. In general, if a categorical variable exist at, say m categories, only $m-1$ indicator or dummy variables are needed to completely characterize the categorical variable. Now, suppose we wish to write a regression model for sales volume y with predictors $x_1 =$ number of households and gender. The model can now be written using the dummy variables we defined in the following way.

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

Then, if we wish to write a sales volume model for females ($x_2 = 1$) only, we substitute $x_2 = 1$ into the model to obtain,

$$y_i = (\beta_0 + \beta_2) + \beta_1 x_{i1} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

We can also write a model for males as,

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Based on the models for males and females, we can examine whether sales volume depends on gender by testing $H_0 : \beta_2 = 0$. Clearly, β_2 measures the differential effect of gender on sales volume. This test can be conducted using the general linear test approach we discussed earlier. If sales volume does not depend on gender we will not reject H_0, otherwise we will reject H_0.

Are there other ways that numerical codes can be assigned to categorical variables? The answer is yes. However, dummy variables make the interpretation of the regression model parameters associated with the numerically coded variables clearer and easier to understand. In another example, suppose we wish to examine the effect of region (Western, Eastern and Atlantic Canada) on the sale volume of masks y given the population density x. We can chose to define the dummy variables for region as

$$\begin{aligned} D_{i1} &= \begin{cases} 1, & \text{if } i\text{th sale volume is from the west,} \\ 0, & \text{otherwise,} \end{cases} \\ D_{i2} &= \begin{cases} 1, & i\text{th sale volume is from the east,} \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Clearly, if $D_{i1} = 1$ and $D_{i2} = 0$ the ith observed sale volume is from the west. If $D_{i1} = 0$ and $D_{i2} = 1$ the ith observed sale volume is from the east. Furthermore, if $D_{i1} = 0$ and $D_{i2} = 0$ the ith observed sale volume is from the atlantic region. The model then becomes,

$$y_i = \beta_0 + \beta_1 x_i + \beta_1 D_{i1} + \beta_1 D_{i2} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

We can then write down regression models for each of the regions and conduct various tests using the statistical tools we have studied to examine the effect of regions on the sale of masks. We now use an example to illustrate the use of categorical variables to improve the predictive power of regression models.

Example 33 (Grade Point Average, Problem 8.16, Page 337)

In Example 21, we saw that the fitted simple linear regression model for the GPA data only accounted for 7.26% of the total variation in GPA when ACT Scores was used as the predictor variable. This implied that the uncertainty in predicting GPA with ACT score (x_1) as the predictor variable, is very large. In order to further reduce the uncertainty and improve the predictive power of the model, an assistant to the director of admissions suggested adding information on whether the student had chosen a major field of concentration at the time the application was submitted. To add this information to the model, we first define,

$$x_{i2} = \begin{cases} 1, & \text{if major field of concentration indicated} \\ & \text{at time of application by } i\text{th student,} \\ 0, & \text{otherwise.} \end{cases}$$

The updated model for the GPA data, then becomes,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, 2, \ldots, 120.$$

Using the R software, we obtain the results shown below

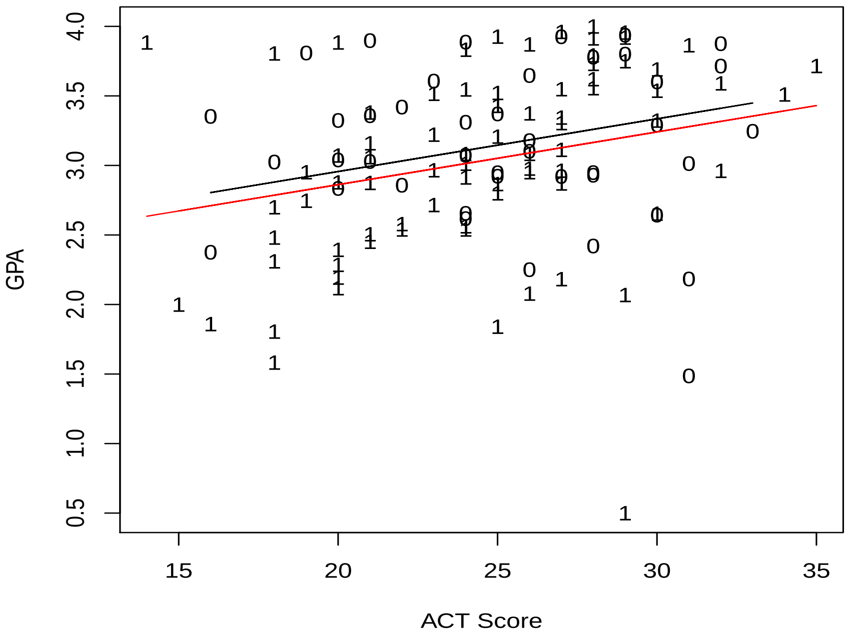


Figure 15: A scatterplot of GPA data with fitted regression line of students who indicated their major at time of application in red and those who did not in black.

Coefficients:

	Estimate	Std. Error	t value	Pr(> $|t|$)
(Intercept)	2.19842	0.33886	6.488	2.18e-09 ***
ACT	0.03789	0.01285	2.949	0.00385 **
Maj	-0.09430	0.11997	-0.786	0.43341

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6241 on 117 degrees of freedom Multiple R-squared: 0.07749, Adjusted R-squared: 0.06172 F-statistic: 4.914 on 2 and 117 DF, p-value: 0.008928

From the output shown above, the fitted model is

$$\hat{y}_i = 2.198 + 0.03789 x_{i1} - 0.0943 x_{i2}, \quad i = 1, 2, \ldots, 120.$$

The fitted regression function for students who did not indicate their major of concentration at time of application is

$$\hat{y}_i = 2.198 + 0.03789x_{i1}, \quad i = 1, 2, \ldots, 42.$$

The fitted line for this group of students is the solid black line in Figure 15. The fitted regression function for the students who indicated their major of concentration at time of application is the solid red line in Figure 15. The model for this group is given by,

$$\hat{y}_i = 2.104 + 0.03789x_{i1}, \quad i = 1, 2, \ldots, 78.$$

It is clear that for this data the extra information added to the model did not lead to an improvement in the predictive capability of the model because this additional information only contributed an extra, 7.749-7.26 = 0.489% reduction in the uncertainty in predicting GPA when ACT score is in the model. Also, the models indicate that on the average, the GPA of students who provided the extra information was marginally lower than that of the students who did not since $\hat{\beta}_2 = -0.0943$ is negative. Note that β_2 measures the difference in the GPA of the two groups of students.

Now, a test can be carried out to determine if there is a significant difference between the GPA of students who provided the extra information and those who did not by testing $H_0 : \beta_2 = 0$ against a two-sided alternative. From the R output, the p-value for this test is 0.43341. Thus, evidence in the data strongly supports the argument that there is no difference in the GPA of the two groups of students.

Statistics 6519 - Regression Analysis
Lecture Notes (Contd)

8 Model Building: Model Selection and Validation

8.1 Extra Sum of Squares

The concept of extra sum of squares is quite useful in model building because it measures the marginal reduction in the error sum of squares when one or several predictor variables are added to a regression model, given that other predictor variables are already in the model. Alternatively, an extra sum of squares is a measure of the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model. Note that any increase in SSR also means a decrease in the amount of variation in y that is not captured by the fitted model. The question we seek to answer is whether such an increase in SSR (or decrease in SSE) is statistically significant. If no, it means the predictors that were added does not contribute significantly in reducing the uncertainty in predicting y. In that case, it makes sense to drop them from the model.

We will now use an example to introduce new notations for SSE, SSR and the extra sum of squares. Consider the data in Table 7.1 of the course text on amount of body fat y, triceps skinfold thickness (x_1), thigh circumference (x_2), and midarm circumference (x_3). Suppose we wish to build the best model for predicting y using these three predictors with SSE, SSR, R^2 and $R^2_{adjusted}$ as the criteria for choosing the best model. We proceed as follows.

Step 1. We begin by fitting models with only one of the predictors to the data. Using the R software, we obtain the following model for x_1,

$$\hat{y} = -1.4961 + 0.8572x_1.$$

For the purpose of clarity, the SSE and SSR for this model will be denoted by $SSE(x_1) = 143.12$ and $SSR(x_1) = 352.27$ with $R^2 = 0.7111$ and $R^2_{adjusted} = 0.695$. Similarly, for x_2 we obtain

$$\hat{y} = -23.6345 + 0.8565x_2,$$

with $SSE(x_2) = 113.42$ and $SSR(x_2) = 381.97$ with $R^2 = 0.771$ and $R^2_{adjusted} = 0.7583$. For x_3 we obtain

$$\hat{y} = 14.6868 + 0.1994x_3,$$

with $SSE(x_3) = 485.34$ and $SSR(x_3) = 10.05$ with $R^2 = 0.02029$ and $R^2_{adjusted} = -0.03414$.

Step 2. By comparing the values of SSE, SSR, R^2 and $R^2_{adjusted}$ for the three models we see that the reduction in the uncertainty in predicting y is highest when x_2 is in the model. So, we select the model,

$$\hat{y} = -23.6345 + 0.8565x_2, \tag{8.1}$$

as our starting model. Next, we examine which of the other variables, x_1 or x_3, should be added to our starting model which already contains x_2. Later, we will discuss how to determine whether the variable added to the model has significantly improved the predictive capability of the model. Again, using the R software, we obtain the following results. Adding x_1, we obtain,

$$\hat{y} = -19.1742 + 0.6594x_2 + 0.2224x_1. \tag{8.2}$$

For this model, $SSE(x_2, x_1) = 109.95$ with $R^2 = 0.7781$ and $R^2_{adjusted} = 0.7519$. We note that $SSE(x_2, x_1)$ is smaller than $SSE(x_2)$. This means, adding x_1 to the starting model reduced the uncertainty in predicting y. It is this marginal reduction in SSE that is referred to as the extra sum of squares commonly denoted by $SSR(x_1|x_2)$ and interpreted as the SSE of x_1 given that x_2 was in the model. We compute the extra sum of squares after adding x_1 to the model as,

$$SSR(x_1|x_2) = SSE(x_2) - SSE(x_2, x_1) = 113.42 - 109.95 = 3.47.$$

Notice that the ANOVA table for this model, from the R software shown in Table 6,

Table 6. Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x2	1	381.97	381.97	59.057	6.281e-07	***
x1	1	3.47	3.47	0.537	0.4737	
Residuals	17	109.95	6.47			

decomposes $SSR(x_2, x_1)$ into two components, namely, $SSR(x_2) = 381.97$ and $SSR(x_1|x_2) = 3.47$. That is,

$$SSR(x_2, x_1) = SSR(x_2) + SSR(x_1|x_2) = 381.97 + 3.47 = 385.44.$$

It follows that, the extra sum of squares $SSR(x_1|x_2)$, can also be computed as

$$SSR(x_1|x_2) = SSR(x_2, x_1) - SSR(x_2) = 385.44 - 381.97 = 3.47.$$

Clearly, the marginal reduction in SSE, 3.47, after adding x_1 is quite small in magnitude. We also notice that $R^2_{adjusted} = 0.7519$ for this model is actually smaller than $R^2_{adjusted} = 0.7583$ for the starting model. This means that based on this criteria the starting model with only x_2, (8.1) is better than the model (8.2) with (x_1, x_2). In addition to the $R^2_{adjusted}$ statistic, one can also conduct a formal statistical test to determine if the contribution of x_1 to the model (8.2) is statistically significant. Before fitting the model to obtain (8.2), suppose the FULL MODEL is represented as

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_1 x_{i1} + \epsilon_i, i = 1, \ldots, n.$$

Then, the null hypothesis for examining the contribution of x_1 is, $H_0 : \beta_1 = 0$. Under $H_0 : \beta_1 = 0$, the REDUCED MODEL becomes,

$$y_i = \beta_0 + \beta_2 x_{i2} + \epsilon_i, i = 1, \ldots, n.$$

Using the general linear test approach, the test statistic becomes,

$$\begin{aligned} F_0^* &= \frac{SSE(\mathit{Reduced}) - SSE(\mathit{Full})}{\mathit{df}(\mathit{Reduced}) - \mathit{df}(\mathit{Full})} \div \frac{SSE(\mathit{Full})}{\mathit{df}(\mathit{Full})} \\ &= \frac{SSE(x_2) - SSE(x_2, x_1)}{(n-2) - (n-3)} \div \frac{SSE(x_2, x_1)}{n-3} \\ &= \frac{SSR(x_1|x_2)}{1} \div \frac{SSE(x_2, x_1)}{n-3} \\ &= \frac{3.47}{1} \div \frac{109.95}{17} \approx 0.537 \end{aligned}$$

Students should notice that the numerator of F_0^*, is actually the extra sum of squares. The p-value corresponding to $F_0^* = 0.537$ is,

$$p - value = 1 - P[F_{1,17} < 0.537] \approx 0.4737.$$

This result agrees with the conclusion based on the $R^2_{adjusted}$. It indicates a strong evidence in the data that x_1 does not contribute significantly to the prediction of y.

Step 3. Now, adding x_3, we obtain,

$$\hat{y} = -25.99695 + 0.85088 x_2 + 0.09603 x_3. \tag{8.3}$$

For this model, $SSE(x_2, x_3) = 111.11$ with $R^2 = 0.7757$ and $R^2_{adjusted} = 0.7493$, and extra sum of squares,

$$SSR(x_3|x_2) = SSE(x_2) - SSE(x_2, x_3) = 113.42 - 111.11 = 2.31.$$

The ANOVA table for this model from the R software is shown in Table 7.

Table 7. Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x2	1	381.97	381.97	58.441	6.737e-07	***
x3	1	2.31	2.31	0.354	0.5597	
Residuals	17	111.11	6.54			

Again, we see that

$$SSR(x_2, x_3) = SSR(x_2) + SSR(x_3|x_2) = 381.97 + 2.31 = 384.28.$$

Also,

$$SSR(x_3|x_2) = SSR(x_2, x_3) - SSR(x_2) = 384.28 - 381.97 = 2.31.$$

The marginal reduction in SSE, 2.31, for this model (8.3) is even smaller than that of (8.2). Thus, the two predictor model (8.2) is better than the two predictor model (8.3). In addition, to conduct a formal statistical test, we proceed as in Step 2 to test the null hypothesis $H_0 : \beta_3 = 0$. Under $H_0 : \beta_3 = 0$, the REDUCED MODEL is also,

$$y_i = \beta_0 + \beta_2 x_{i2} + \epsilon_i, i = 1, \ldots, n.$$

Again, using the general linear test approach, the test statistic becomes,

$$\begin{aligned} F_0^* &= \frac{SSE(Reduced) - SSE(Full)}{df(Reduced) - df(Full)} \div \frac{SSE(Full)}{df(Full)} \\ &= \frac{SSE(x_2) - SSE(x_2, x_3)}{(n-2) - (n-3)} \div \frac{SSE(x_2, x_3)}{n-3} \\ &= \frac{SSR(x_3|x_2)}{1} \div \frac{SSE(x_2, x_3)}{n-3} \\ &= \frac{2.31}{1} \div \frac{109.95}{17} \approx 0.354 \end{aligned}$$

Students should again notice that the numerator of F_0^*, is actually the extra sum of squares. The p-value corresponding to $F_0^* = 0.354$ is,

$$p - value = 1 - P[F_{1,17} < 0.354] \approx 0.5597.$$

This result also agrees with the conclusion based on the $R^2_{adjusted}$. It indicates a very strong evidence in the data that x_3 does not contribute significantly to the prediction of y.

Step 4. How about a 3-predictor variable model? The fitted model after adding x_3 to (8.2) is,

$$\hat{y} = 117.085 - 2.857x_2 + 4.334x_1 - 2.816x_3, \tag{8.4}$$

with $SSE(x_2, x_1, x_3) = 98.40$, $R^2 = 0.8014$ and $R^2_{adjusted} = 0.764$. For this model, the extra sum of squares is,

$$SSR(x_3|x_2, x_1) = SSE(x_2, x_1) - SSE(x_2, x_1, x_3) = 109.95 - 98.40 = 11.55.$$

Based on the $R^2_{adjusted}$ criteria, the 3-predictor model (8.4), is better than the model (8.1) with only x_2 ($R^2_{adjusted} = 0.7583$). From the R software, we obtain Table 8 below.

Table 7. Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x2	1	381.97	381.97	62.1052	6.735e-07	***
x1	1	3.47	3.47	0.5647	0.4633	
x3	1	11.55	11.55	1.8773	0.1896	
Residuals	16	98.40	6.15			

We again notice that the SSR for this model is decomposed into 3 components as follows,

$$\begin{aligned} SSR(x_2, x_1, x_3) &= SSR(x_2) + SSR(x_1|x_2) + SSR(x_3|x_2, x_1) \\ &= 381.97 + 3.47 + 11.55 = 396.99. \end{aligned}$$

This implies that we can also compute the extra sum of squares as,

$$\begin{aligned} SSR(x_3|x_2, x_1) &= SSR(x_2, x_1, x_3) - SSR(x_2) - SSR(x_1|x_2) \\ &= 396.99 - 381.97 - 3.47 = 11.55. \end{aligned}$$

Using the general linear test approach again, we test whether x_3 is a useful predictor in the 3-predictor variable model (8.4). In this case, under $H_0 : \beta_3 = 0$, the reduced model is,

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_1 x_{i1} + \epsilon_i, i = 1, \ldots, n,$$

and the full model is

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i, i = 1, \ldots, n.$$

The test statistic is computed as,

$$\begin{aligned} F_0^* &= \frac{SSE(Reduced) - SSE(Full)}{df(Reduced) - df(Full)} \div \frac{SSE(Full)}{df(Full)} \\ &= \frac{SSE(x_2, x_1) - SSE(x_2, x_1, x_3)}{(n-3) - (n-4)} \div \frac{SSE(x_2, x_1, x_3)}{n-4} \\ &= \frac{SSR(x_3|x_2, x_1)}{1} \div \frac{SSE(x_2, x_1, x_3)}{n-4} \\ &= \frac{11.55}{1} \div \frac{98.40}{16} \approx 1.8773 \end{aligned}$$

Students should again notice that the numerator of F_0^*, is actually the extra sum of squares. The p-value corresponding to $F_0^* = 1.8773$ is,

$$p - value = 1 - P[F_{1,16} < 1.8773] \approx 0.1896.$$

The p-value indicates evidence in support of H_0. That is, x_3 does not contribute significantly to the prediction of y. The results of our analysis in Steps 1-4 clearly show that based on the $R^2_{adjusted}$ criterion, the 3-predictor model (8.4),

$$\hat{y} = 117.085 - 2.857x_2 + 4.334x_1 - 2.816x_3,$$

is better than either the 1-predictor variable or 2-predictor variable models (8.2), (8.1) and (8.3). However, the statistical analysis indicates that x_3 can be dropped from the 3-predictor variable model. This contradictory conclusions is an indication that pairs of the 3 predictor variables may be strongly correlated. Furthermore, we can use the extra sum of squares to test the hypothesis, $H_0 : \beta_1 = \beta_3 = 0$ in the 3-predictor model (8.4). We first note that for this test, the FULL MODEL is,

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i, i = 1, \ldots, n,$$

and the REDUCED MODEL is,

$$y_i = \beta_0 + \beta_2 x_{i2} + \epsilon_i, i = 1, \ldots, n.$$

It follows that the general linear test statistic for $H_0 : \beta_1 = \beta_3 = 0$ is,

$$\begin{aligned} F_0^* &= \frac{SSE(Reduced) - SSE(Full)}{df(Reduced) - df(Full)} \div \frac{SSE(Full)}{df(Full)} \\ &= \frac{SSE(x_2) - SSE(x_2, x_1, x_3)}{(n-2)-(n-4)} \div \frac{SSE(x_2, x_1, x_3)}{n-4} \\ &= \frac{SSR(x_1, x_3|x_2)}{2} \div \frac{SSE(x_2, x_1, x_3)}{n-4}. \end{aligned}$$

Now, we note that

$$SSR(x_1, x_3|x_2) = SSR(x_1|x_2) + SSR(x_3|x_1, x_2) = 3.47 + 11.55 = 15.02.$$

Therefore,

$$F_0^* = \frac{15.02}{2} \div \frac{98.40)}{16} \approx 1.2211,$$

with $p - value = 1 - P(F_{2,16} < 1.2211) = 0.32095$, indicating strong evidence in the data in support of $H_0 : \beta_1 = \beta_3 = 0$.

Next, we summarize some of the results and notations we have introduced through the example in Steps 1 - 4 above.

$$\begin{aligned} SSR(x_1|x_2) &= SSE(x_2) - SSE(x_2, x_1) \\ SSR(x_2, x_1) &= SSR(x_2) + SSR(x_1|x_2) \\ SSR(x_1|x_2) &= SSR(x_2, x_1) - SSR(x_2) \\ SSR(x_3|x_2) &= SSE(x_2) - SSE(x_2, x_3) \\ SSR(x_2, x_3) &= SSR(x_2) + SSR(x_3|x_2) \\ SSR(x_3|x_2) &= SSR(x_2, x_3) - SSR(x_2) \\ SSR(x_2, x_1, x_3) &= SSR(x_2) + SSR(x_1|x_2) + SSR(x_3|x_2, x_1) \\ SSR(x_3|x_2, x_1) &= SSR(x_2, x_1, x_3) - SSR(x_2) - SSR(x_1|x_2) \\ SSR(x_1, x_3|x_2) &= SSR(x_1|x_2) + SSR(x_3|x_1, x_2) \end{aligned}$$

In general, the general linear test statistic F_0^* for the hypothesis $H_0 : \beta_k = 0$, for some $k \leq p-1$, in the model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \quad i = 1, \ldots, n,$$

can be written in terms of extra sum of squares as

$$F_0^* = \frac{SSR(x_k|x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{p-1})}{1} \div \frac{SSE(x_1, x_1, \ldots, x_{p-1})}{n-p}.$$

A similar general expression, based on extra sum of squares can also be written for testing whether some β_k's are zero.

Example 34 (Patient Satisfaction, Problems 7.5 and 7.6 Pages 289 and 290)

The ANOVA table for this data which partitions the regression sum of squares into single degrees of freedom extra sum of squares with x_2 as the first variable to enter the model is shown in Table 9.

Table 9. Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X2	1	4860.3	4860.3	48.0439	1.822e-08 ***
X1	1	3896.0	3896.0	38.5126	2.008e-07 ***
X3	1	364.2	364.2	3.5997	0.06468 .
Residuals	42	4248.8	101.2		

From the ANOVA table we find that,

$$\begin{aligned} SSR(x_2) &= 4860.3, \;\; SSR(x_1|x_2) = 3896.0, \\ SSR(x_3|x_2, x_1) &= 364.2, \;\; \text{and} \;\; SSE(x_2, x_1, x_3) = 4248.8. \end{aligned}$$

Next, we test if x_3 can be dropped from the model given that x_1 and x_2 are retained.

$H_0: \; \beta_3 = 0$
$H_a: \beta_3 \neq 0$

Significance level: $\alpha = 0.025$.

Test Statistic Value:

$$\begin{aligned} F_0^* &= \frac{SSE(Reduced) - SSE(Full)}{df(Reduced) - df(Full)} \div \frac{SSE(Full)}{df(Full)} \\ &= \frac{SSE(x_2, x_1) - SSE(x_2, x_1, x_3)}{(n-3) - (n-4)} \div \frac{SSE(x_2, x_1, x_3)}{n-4} \\ &= \frac{SSR(x_3|x_2, x_1)}{1} \div \frac{SSE(x_2, x_1, x_3)}{n-4} \\ &= \frac{364.2}{1} \div \frac{4248.8}{42} \approx 3.5997 \end{aligned}$$

p-value: p-value $= 1 - P(F_{1,42} < 3.5997)] = 0.06468$.

Critical value: $F(0.9875, 1, 42) \approx 6.8113$.

Conclusion: We cannot reject H_0 since $3.5997 < 6.8113$ (alternatively, 0.025 or $0.01 < 0.06468$). That means, there is evidence in the data in support of H_0, that, x_3 can be dropped from the model.

Next, we test if x_3 and x_2 can be dropped from the model given that x_1 is retained. The ANOVA table for this test is,

Table 9. Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	8275.4	8275.4	81.8026	2.059e-11 ***
X2	1	480.9	480.9	4.7539	0.03489 *
X3	1	364.2	364.2	3.5997	0.06468 .
Residuals	42	4248.8	101.2		

$H_0: \ \beta_2 = \beta_3 = 0$
$H_a: \beta_2 \neq 0, \ \text{or } \beta_3 \neq 0 \text{ or both } \neq 0$

Significance level: $\alpha = 0.025$.

Test Statistic Value:

$$\begin{aligned}
F_0^* &= \frac{SSE(Reduced) - SSE(Full)}{df(Reduced) - df(Full)} \div \frac{SSE(Full)}{df(Full)} \\
&= \frac{SSE(x_1) - SSE(x_1, x_2, x_3)}{(n-2) - (n-4)} \div \frac{SSE(x_1, x_2, x_3)}{n-4} \\
&= \frac{SSR(x_2, x_3|x_1)}{2} \div \frac{SSE(x_1, x_2, x_3)}{n-4} \\
&= \frac{SSR(x_2|x_1) + SSR(x_3|x_1, x_2)}{2} \div \frac{SSE(x_1, x_2, x_3)}{n-4} \\
&= \frac{480.9 + 364.2}{2} \div \frac{4248.8}{42} \approx 4.1769
\end{aligned}$$

p-value: p-value $= 1 - P(F_{2,42} < 4.1769) \approx 0.02216$.

Critical value: $F(0.975, 2, 42) \approx 4.03271$.

Conclusion: We reject H_0 since $4.1769 > 4.03271$ (alternatively, $0.025 > 0.02216$). That means, there is evidence in the data in support of H_a, that, x_2 and x_3 cannot be dropped from the model.

8.2 Coefficient of partial determination

Recall that the extra sum of squares measure the marginal reduction in SSE when one or more predictor variables are added to a model that contains other variables. The coefficient of partial determination measures the marginal proportion of variation explained by introducing these additional variable or variables into the model. To fix ideas, let's begin by considering a model with only two predictor variables, say x_1 and x_2, given by

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

Now, $SSE(x_1)$ measures the unexplained variation in y when x_1 is in the model and $SSE(x_1, x_2)$ measures the unexplained variation in y when x_1 and x_2 are in the model. Thus, the marginal reduction in the unexplained variation or the contribution of x_2 in further reducing the unxplained variation is given by,

$$SSR(x_2|x_1) = SSE(x_1) - SSE(x_1, x_2) = SSR(x_1, x_2) - SSR(x_1),$$

the extra sum of squares. The relative marginal reduction in unexplained variation due to the addition of x_2 when x_1 is already in the model is the coefficient of partial determination between y and x_2 given that x_1 is already in the model defined by,

$$R^2_{Y2|1} = \frac{SSR(x_2|x_1)}{SSE(x_1)}.$$

Similarly, we can define the coefficient of partial determination between y and x_1 given that x_2 is already in the model as,

$$R^2_{Y1|2} = \frac{SSR(x_1|x_2)}{SSE(x_2)}.$$

This idea can be generalized to 3 or more predictor variables as follows,

$$\begin{aligned} R^2_{Y1|23} &= \frac{SSR(x_1|x_2, x_3)}{SSE(x_2, x_3)}, \\ R^2_{Y3|12} &= \frac{SSR(x_3|x_1, x_2)}{SSE(x_1, x_2)}, \\ R^2_{Y3|124} &= \frac{SSR(x_3|x_1, x_2, x_4)}{SSE(x_1, x_2, x_4)}. \end{aligned}$$

Example 35 (Patient Satisfaction)

For the purpose of illustration, we use results from Example 34 to compute,

$$R^2_{Y3|12} = \frac{SSR(x_3|x_1, x_2)}{SSE(x_1, x_2)}.$$

From the ANOVA table in Example 34, we have that,

$$SSR(x_3|x_1, x_2) = 364.2, \quad SSE(x_1, x_2, x_3) = 4248.8.$$

Since,

$$SSR(x_3|x_1, x_2) = SSE(x_1, x_2) - SSE(x_1, x_2, x_3),$$

we have that,

$$SSE(x_1, x_2) = SSR(x_3|x_1, x_2) + SSE(x_1, x_2, x_3) = 364.2 + 4248.8 = 4613.$$

Therefore,

$$R^2_{Y3|12} = \frac{SSR(x_3|x_1, x_2)}{SSE(x_1, x_2)} = \frac{364.2}{4613} \approx 0.07895.$$

8.3 Criteria for model selection

Suppose we have $p - 1$ potential predictor variables that can be used to build the best, in some sense, regression model. Then, we can construct a total of 2^{p-1} possible models usig the $p - 1$ predictor variables. So far, we have discussed how we can use $R^2_{adjusted}$ and the extra sum of squares for model selection. There are several other criteria that can also be used to select the best model among several competing models. We discuss three of these criteria below.

Mallows' C_k Criterion: Under this criterion, it is assumed that the model which includes all $p-1$ potential x variables has been carefully chosen so that $MSE(x_1,\ldots,x_{p-1})$ is an unbiased estimator of the constant error variance σ^2. For $k=2,\ldots,p$, the C_k criterion is then defined as,

$$C_k = \frac{SSE(x_1,\ldots,x_{k-1})}{MSE(x_1,\ldots,x_{p-1})} - (n-2k),$$

where $SSE(x_1,\ldots,x_{k-1})$ is the error sum of squares for the model with predictor variables $x_1,\ldots,x_{k-1}$. We observe that by this definition, when $k=p$, we have $C_p=p$. When using this criterion, we seek the model with a C_k value that is (a) small in value, and (b) near k. The C_k value is a measure of total mean squared error of y. So, a small C_k value indicates that the total mean squared error of y for that model is small. Recall that the mean squared error of y can be partitioned into the squared bias of y and the variance of y. Now, when the C_k value of a model is near k, it indicates that the bias of the model is small.

AIC and SBC Criterion: The Akaike Information Criterion (AIC) and Schwarz' Bayesian Criterion (SBC) are derived by maximizing the loglikelihood function of the responses. Similar to the $R^2_{adjusted}$ criterion, the model with the smallest AIC and SBC criteria is selected as the best model. These criteria are defined as,

$$\begin{aligned} AIC_k &= nlog_eSSE(x_1,\ldots,x_{k-1}) - nlog_en + 2k \\ SBC_k &= nlog_eSSE(x_1,\ldots,x_{k-1}) - nlog_en + (log_en)k, \quad k=2,\ldots,p. \end{aligned}$$

For the purpose of illustration, we have used the results for the body fat example to compute the C_k, AIC_k and SBC_k statistics for the various models. Note that since there are 3 possible predictor variables, we can construct a total of $2^3=8$ subset models with various combinations of the predictor variables. Now, for the model with only x_1, we had $SSE(x_1)=143.12$. Also, when all predictors are in the model we had $MSE(x_1,x_2,x_3)=6.15$. Therefore, for the model with only x_1 as predictor

$$\begin{aligned} C_2 &= \frac{SSE(x_1)}{MSE(x_1,x_2,x_3)} - (n-2\times 2) = \frac{143.12}{6.15} - (20-2\times 2) \approx 7.27 \\ AIC_2 &= nlog_eSSE(x_1) - nlog_en + 2k = 20log_e(143.12) - 20log_e20 + 2\times 2 \approx 43.36 \\ SBC_k &= nlog_eSSE(x_1) - nlog_en + (log_en)k = 20log_e(143.12) - 20log_e20 + 2log_e20 \approx 45.35. \end{aligned}$$

The values of these criteria for models involving all possible combinations of the predictor variables are shown in Table 11, below.

Table 11. C_k, AIC_k and SBC_k Values for Body Fat Data

Index	k	Predictors	C_k	AIC_k	SBC_k
1	2	x_1	7.27034	43.35898	45.35045
2	2	x_2	2.441959	38.70796	40.69942
3	2	x_3	62.91281	67.78226	69.77373
4	3	x_2 x_1	3.877289	40.08601	43.07321
5	3	x_2 x_3	4.065734	40.29573	43.28293
6	3	x_1 x_3	3.224212	39.34171	42.32891
7	4	x_2 x_1 x_3	4.0000	39.86716	43.85009

The results in Table 11 show that the models with C_k values that are small and closest to k are Models 2 and 6. However, the AIC and SBC values for Model 2 are smaller than that of all the other models. Thus, based on these criteria the best model is Model 2, the model with only x_2 as predictor.

Example 35 (Patient Satisfaction)

The "olsrr" package in the R software can be used to compute the values of the C_k statistic for all possible combinations of the predictor variables under consideration. Using the "*ols_step_all_possible(lm.object)*" command we obtain the results in Table 12 for the patient satisfaction data with 3 predictor variables.

Table 12. C_k Values for Patient Satisfaction Data

Index	k	Predictors	R-Square Adj.	R-Square	Mallow's Cp
1	2	X1	0.6189843	0.6103248	8.353606
2	2	X3	0.4154975	0.4022134	35.245643
3	2	X2	0.3635387	0.3490737	42.112324
4	3	X1 X3	0.6760864	0.6610206	2.807204
5	3	X1 X2	0.6549559	0.6389073	5.599735
6	3	X2 X3	0.4684545	0.4437314	30.247056
7	4	X1 X2 X3	0.6821943	0.6594939	4.000000

In this case, we see that the $R^2_{adjusted}$ is largest for Models 4 and 7, the models with $(X1, X3)$ and the models with $(X1, X2, X3)$, respectively. The R^2 value is however better for Model 4, though it has one less predictor variable. The C_k value for Model 4 is also small and close to the value of k. Thus, based on the C_k criterion, Model 4 which contains $(X1, X3)$ as predictor variables is selected as the best model. Students should compute the AIC_k and SBC_k values for this data as practice.

8.4 Search algorithms for model selection

We had noted earlier that when building a model involving $p-1$ potential predictor variables we have to search among 2^{p-1} possible models in order to select the best model. When,

$p-1 = 3$, there are $2^3 = 8$ possible models to compare. As the number of potential predictor variables increases, the number of possible models we can choose from increases exponentially. So some statisticians have developed search algorithms that does not require us to fit all the possible models. We will discuss the Best Subsets and Stepwise Algorithms.

Best Subsets Algorithms: These algorithms simply identify the best subsets of predictor variables based on a specified criterion without fitting all possible models. When $p - 1 > 30$, these algorithms usually require a lot of computational time. As a result, it may not be the most efficient approach when $p - 1 > 30$. Again, the "olsrr" package in R software can be used to implement the best subset algorithm using the command, "$ols\_step\_best\_subset(lm.object)$". As an example, consider the data on survival times in a surgical unit g iven in Table 9.1 with eight (8) potential predictor variables. For this data, using the "$ols\_step\_best\_subset(lm.object)$" command we obtain the results shown below.

Surgical Unit Example - Table 9.1, Page

Table 13. Best Subsets Regression

Model Index	Predictors								
1	X9								
2	X3	X9							
3	X2	X3	X9						
4	X2	X3	X8	X9					
5	X2	X3	X6	X8	X9				
6	X2	X3	X5	X6	X8	X9			
7	X2	X3	X5	X6	X7	X8	X9		
8	X2	X1	X3	X5	X6	X7	X8	X9	
9	X2	X1	X3	X4	X5	X6	X7	X8	X9

Table 14. Subsets Regression

Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C_k	AIC	SBC
1	0.8610	0.8583	0.8165	26.9140	-25.0117	-19.0447
2	0.8817	0.8770	0.8299	17.4669	-31.7134	-23.7575
3	0.9041	0.8984	0.8602	7.0464	-41.0738	-31.1289
4	0.9099	0.9026	0.8671	5.8323	-42.4490	-30.5151
5	0.9139	0.9049	0.8679	5.6413	-42.8770	-28.9541
6	0.9160	0.9053	0.8665	6.4594	-42.2336	-26.3217
7	0.9179	0.9054	0.8655	7.4454	-41.4251	-23.5243
8	0.9192	0.9049	0.86	8.6800	-40.3424	-20.4526
9	0.9205	0.9042	0.8553	10.0000	-39.1706	-17.2917

In this example, we see that Model 7 with $X_2, X_3, X_5, X_6, X_7, X_8, X_9$ as predictor variables is selected based on the $R^2_{adjusted}$ criterion because this model has the largest value of $R^2_{adjusted}$. The C_k criterion leads to Model 5 with predictor variables X_2, X_3, X_6, X_8, X_9, because the C_k value for this model is (a) near $k = 6$, and (b) small. This 5 predictor variable model is also selected by the AIC criterion. Model diagnostics and the generalized linear test approach can then be used to select the best among these two competing models.

Stepwise Regression Algorithms: As the name itself suggests, stepwise algorithms develop the best subset model sequentially by adding and/or removing predictor variables at each step. There are two main approaches, namely, forward stepwise regression and backward stepwise regression.

Forward stepwise regression procedure

1. Fit a simple linear regression model to each of the $p-1$ potential predictor variables. Compute the value of the t-statistic,
$$t_0 = \frac{\hat{\beta}_k}{s_{\hat{\beta}_k}},$$
for testing $H_0 : \beta_k = 0$, $k = 1, \ldots, p-1$, for each of the $p-1$ SLR models. The model with the largest t_0 value (or smallest p-value), such that H_0 is rejected, is selected as the starting model. If H_0 is not rejected for all $p-1$ models, the procedure terminates with no model selected.

2. Suppose the SLR model with X_{10} was selected as the starting model. Fit all possible two predictor models by adding the other variables that are not in the starting model to the starting model one at a time. Now, compute t_0 for each of the variables X_k, $k = 1, 2, \ldots, 9, 11, \ldots, p-1$ that have been added to the starting model in the two predictor model. The two predictor model with the largest t_0 value (or smallest p-value), such that $H_0 : \beta_k = 0$, $k = 1, 2, \ldots, 9, 11, \ldots, p-1$ is rejected, is selected as the two predictor model in step 2. If $H_0 : \beta_k = 0$ is not rejected for all of the two predictor models, the procedure terminates.

3. Suppose X_5 entered the model in Step 2, so that the two predictor model selected was,
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_{10} x_{10} + \hat{\beta}_5 x_5.$$
Next, examine whether x_{10} should be retained in the two predictor model by computing t_0 for testing $H_0 : \beta_{10} = 0$. If H_0 is rejected, we retain x_{10} in the model. Otherwise, x_{10} is dropped from the model. At later stages, when there will be more than 2 predictors in the selected model, there will be a number of these t_0 statistics, one for each of the variables in the model besides the one last added.

4. Assuming x_{10} was retained in the model, we now determine which of the other variables should enter the model by fitting all possible three predictor variables with x_{10} and x_5 already in the model. After deciding which variable

should enter the model we also check whether any of the other variables already in the model should be dropped or retained in the model. This process is repeated until no predictor variables can either be added or deleted from the model, at which point the procedure terminates.

For the purpose of illustration, we have used the surgical unit data to obtain the output in Table 15. We again used the "olsrr" package to obtain the results in Table 15.

Table 15. Stepwise Regression - Forward

Selection Summary

	Variable		Adj.			
Step	Entered	R-Square	R-Square	C_k	AIC	RMSE
1	X9	0.8610	0.8583	26.9140	-25.0117	0.1850
2	X3	0.8817	0.8770	17.4669	-31.7134	0.1724
3	X2	0.9041	0.8984	7.0464	-41.0738	0.1567
4	X8	0.9099	0.9026	5.8323	-42.4490	0.1534
5	X6	0.9139	0.9049	5.6413	-42.8770	0.1516
6	X5	0.9160	0.9053	6.4594	-42.2336	0.1513

The variables that entered the model at each stage of the forward regression process is shown in column 2 of Table 15. The final model selected by the procedure was a model with six predictor variables, given by,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_9 x_9 + \hat{\beta}_3 x_3 + \hat{\beta}_2 x_2 + \hat{\beta}_8 x_8 + \hat{\beta}_6 x_6 + \hat{\beta}_5 x_5.$$

Backward stepwise regression procedure: The backward search procedure is the opposite of forward stepwise regression. It begins with the model containing all potential x variables and identifies the one with the smallest t_0 value or largest p-value for testing the significance of each parameter in the model. The corresponding x variable is dropped from the model if the smallest t_0 value is such that we cannot reject H_0. The procedure terminates if we reject H_0 for all variables. The model with the remaining $p-2$ x variables is then fitted, and the next candidate for dropping is identified. We then determine whether the previously dropped variable can be added to the model with $p-3$ variables. This process continues until no further x variables can be dropped and/or added. In Table 16, we display the results of the backward procedure applied to the surgical unit data involving nine potential predictor variables.

Table 16. Stepwise Regression - Backward

Elimination Summary

	Variable		Adj.			
Step	Removed	R-Square	R-Square	C_k	AIC	RMSE
1	X4	0.9192	0.9049	8.6800	-40.3424	0.1516
2	X1	0.9179	0.9054	7.4454	-41.4251	0.1512
3	X7	0.916	0.9053	6.4594	-42.2336	0.1513

Simpler versions of the forward and backward stepwise procedures can also be applied by omitting the test whether a variable once entered into the model should be dropped or once dropped from the model can re-enter the model. These are referred to as forward and backward elimination procedures.