Part I

Simple Linear Regression

Chapter 1

Introduction

This course is designed to provide a broad overview of many common procedures encountered when performing a regression analysis. Roughly, the first half of this course will focus on fitting regression models using ordinary least squares. The simple linear regression model is when we have one predictor variable (or independent variable) and one response variable (or dependent variable). We will also discuss multiple linear regression, which is used when we have more than one predictor variable. Certain assumptions are also made when fitting these models and various diagnostic tools are used to assess these assumptions. Throughout, we will define the notion of statistical inference used to assess the statistical significance of the estimated model at hand, discuss various statistical intervals, and introduce other procedures to deal with various characteristics of the data set (such as outliers and collinearity).

The second half of the course will focus on models where the data may have a certain structure which differs from that described in the ordinary least squares models. It may seem that the topics discussed are slightly disjointed, but when integrated with the material learned from the simple and multiple regression models, you should hopefully feel that a thorough treatment of regression has been presented. Later topics in this course include common procedures implemented when various regression assumptions are violated, determining how to select a model when faced with many predictors, and the exploration of more advanced regression models (including polynomial regression models, nonlinear regression models, nonparametric regression models, autocorrelated data, and regression models for censored data). Such models are needed when the data is not necessarily in the form that we analyze in the earlier portion of this course. As you read through the

material, it is suggested that you occasionally revisit the "Steps for Building a Regression Model" in Appendix A to help came the bigger picture in mind when using regression modeling strategies.

Finally, I also wish to acknowledge my colleague, Bob Heckard (retired Senior Lecturer of Statistics from Penn State University - University Park), who graciously helped me organize this online regression course during my first semester teaching this through Penn State's World Campus (web-based courses leading to certification in various fields). Some of the content and examples we will see are borrowed from that course.

1.1 Why Use Regression?

As a researcher, you will be tasked with formulating and/or investigating hypotheses of interest. In order to adequately explore a research hypothesis, you will need to design an experiment, collect the data, and analyze the data. During the analysis stage, the results will either support or discredit your hypothesis and, in some cases, be inconclusive. In this last case, you will need to assess whatever lessons were learned from the current research, implement those, and then proceed to run through another iteration of the research process while obviously considering other factors (e.g., cost) throughout this process.

Fortunately, the field of Statistics gives an excellent toolbox for guiding you through the experimental design, data collection, and analysis. A major tool in that toolbox concerns the area of **regression analysis**. Regression analysis is a set of techniques used to explore the relationship between at least two variables (i.e., at least one independent and one dependent variable). The majority of regression analyses usually implement a linear regression model, which means that the dependent variable(s0 can be written in terms of a linear combination of the independent variable(s). Some reasons why this is usually the common approach are: (1) the linear regression model is something that is easily understood by the majority of researchers; (2) many processes naturally follow a relationship that is well-represented by a linear relationship; and (3) the use of linear regression models allows for the use of techniques that are well-rooted in statistical theory with desirable asymptotic properties (i.e., large sample properties), thus yielding tractable results. Regardless, this text attempts to provide a broad overview of the many regression techniques that are available to you as a researcher. While

a one-semester regression course can by no means cover all of these topics, this book is meant to serve as a good reference for your future work. It is aimed at getting you to start thinking about the structure of your data and how to best analyze it.

Before we begin our survey of regression methods, let us introduce some of the uses that you as a researcher may find for using regression strategies:

1. **Descriptions:** As a researcher, you may wish to seek some sort of descriptive relationship between a set of measured variables. At this point, you are making the fewest assumptions in your search for such a relationship. This relationship may or may not help to justify a possible deterministic relationship at this time; however, you are at least establishing some sort of connection with the sample of data you are currently working with. *Example: A sociologist may be interested in establishing a relationship between the final occupational status of an individual and the educational level of that individual as well as their parents' educational level.*

2. **Coefficient Estimation:** When analyzing the data, the researcher may have a theoretical, deterministic relationship in mind. Whether this is linear or nonlinear, the use of regression analysis can provide evidence for such a theory (but note that we never say that we can *prove* such a theory - we can only provide *evidence* for such a theory). Of particular interest will be the magnitudes and signs of the coefficients, which will yield insight into the research questions at hand. *Example: A botanist may be interested in estimating the coefficients for an established model used for relating a certain plant's weight with the amount of water it receives, the nutrients in the soil, and the amount of sunlight exposure.*

3. **Prediction:** A researcher may be primarily concerned with predicting some response variable at given levels of other input variables. These predictions may be crucial in planning, monitoring, altering, or evaluating a process or system. For prediction to be valid, various assumptions must be made and met in this case. Most notably, you must not extrapolate beyond the range of the data since the estimation is only valid within the domain of the sampled data. *Example: A realtor has a 20-year history of the home selling prices for those properties that she sold throughout her career as well as the home's total square footage, the*

year it was built, and the assessed value. She will put a new home on the market and wants to be able to predict that home's selling given the values of the other variables provided that they do not extend outside the domain of her data.

4. **Control:** Regression models may be used for monitoring and controlling systems such that the functional relationship continues over time. If it does not, then continual modification of the model must occur. *Example: A manufacturer of semiconductors continuously monitors the camber measurement on the substrate to be within certain limits. During this process, a variety of measurements in the system are recorded, such as lamination temperature, firing temperature, and lamination pressure. These inputs are always controlled within certain limits and if the camber measurement exceeds the designed limits, then the manufacturer must take corrective action.*

5. **Variable Selection or Screening:** A researcher may be faced with many independent variables and just one dependent variable. Since it is not feasible, nor necessarily informative, to model the dependent variable as a function of *all* of the independent variables, a search can be conducted to focus on only a subset of the independent variables that explain a significant amount of the variation in the dependent variable. Historical data may also be used to help in this decision process. *Example: A wine producer may be interested in assessing how the composition of his wine relates to sensory evaluations. A score for the wine's aroma is given by a judging panel and 25 elemental concentrations are recorded from that wine. It is then desired to see which elements explain a significant amount of the variation in the aroma scores.*

Chapter 2

The Basics of Regression Models

Statistical methods (of all kinds) are used to make generalizations about populations on the basis of the information in a sample. Thus, it is important to understand the distinction between a *sample* and a *population*.

- A **sample** is the collection of units (e.g., people, animals, cities, fields, whatever you study) that is actually measured or surveyed in a study.
- The **population** is the larger group of units from which the sample was selected (ideally using probability methods for selection). A sample, which is a subset of the population, is used to estimate characteristics of the population.

For instance, suppose that we measure the ages and distances at which $n = 30$ drivers can read a highway sign (we usually use n to denote the sample size of the study). This collection of 30 drivers is the sample which presumably will be used to represent and estimate characteristics of the larger population of all drivers.

Different notations are used for sample and population characteristics to distinguish the difference. For example, the mean of a sample $x_1, x_2, \ldots, x_n$ is usually denoted by $\bar{x}$, whereas the mean of a population is typically denoted by μ . An alternative notation for a population mean is $\mathrm{E}(X)$, which is read as the "expected value of X".

- With the observed sample, we can compute a value for $\bar{x}$. This sample value estimates the unknown value of the population mean μ.

- It is crucial to recognize that we do not, and will not, know the exact value of μ. We only know the value of its sample estimate, $\bar{x}$.

It is important to note that when we use a capital Roman letter (such as X), we are usually talking about a **random variable**, while smaller case Roman letters (such as x) are typically **realizations** of that random variable using the data. This rule is not set in stone as when we introduce matrix notation, sometimes capital Roman letters will be used for matrices of realizations (or the observed data).

The way we characterize a random variable and realizations of that random variable also differ. A measure computed from sample data is called a **statistic**. A measure characterizing the population is called a **parameter**. For example, $\bar{x}$ is a statistic for the sample while μ is a parameter of the population. One way to keep this bit of terminology straight is to note that sample and statistic both begin with "s" while population and parameter both begin with "p".

2.1 Regression Notation

A component of the simple regression model is that the mean value of the y-variable is a straight-line function of an x-variable. The two coefficients of a straight-line model are the **intercept** and the **slope**. The notation used to distinguish the sample and population versions of these coefficients is given in Table 2.1.

Coefficient	**Sample Notation**	**Population Notation**
Intercept	b_0	β_0
Slope	b_1	β_1

Table 2.1: Notation pertaining to the sample and population coefficients in a regression model.

Suppose that the regression equation relating systolic blood pressure and age for a sample of $n = 50$ individuals aged 45 to 70 years old is

$$\text{average blood pressure} = 95 + 0.80 * \text{age}.$$

Then, the sample slope is $b_1 = 0.80$ and the sample intercept is $b_0 = 95$. We do not know the values of β_0 and β_1 , the intercept and slope, respectively,

for the larger population of all individuals in this age group. For example, it would be **incorrect** to write $\beta_1 = 0.80$. One "trick" of notation is that the "hat" notation is used in Statistics to indicate an estimate. As an alternative to writing $b_1 = 0.80$, we could write $\hat{\beta}_1 = 0.80$, but I choose to use the former initially. Later when alternative regression estimates are presented, then the "hat" notation will be utilized.

2.2 Population Model for Simple Regression

This section is about the (theoretical) regression model which we try to estimate. Again, you will **never** know the actual population regression model in practice. If we always knew population models, then there would be little need for the field of Statistics!

A **regression equation** describes how the *mean* value of a y-variable (also called the *response* or *dependent variable*) relates to specific values of the x-variable(s) (also called the *predictor(s)* or *independent variable(s)*) used to predict y. A **regression model** also incorporates a measure of uncertainty or error. One general format for a regression model is:

$$\text{individual's } y = \text{equation for mean} + \text{individual's deviation from mean.}$$

Suppose that $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are realizations of the random variable pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The **simple linear regression equation** is that the mean of Y is a straight-line function of x. This could be written as:

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i,$$

where $\mathrm{E}(Y_i)$ is used to represent the mean value (expected value) and the subscript i denotes the (hypothetical) i^{th} unit in the population. To be completely pedantic, the simple regression equation should actually be written as the mean of a conditional random variable:

$$\mathrm{E}(Y_i \big| X_i = x_i) = \beta_0 + \beta_1 x_i.$$

The overall **simple linear regression model** for individuals in the larger population from which the sample has been taken can be written as:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i is the "error" or "deviation" of y_i from the line $\beta_0 + \beta_1 x_i$. (We will mostly be dealing with real data, so the model is usually written in terms of the realizations, as done above. However, we could write the model in terms of the general random variables, which is $Y = \beta_0 + \beta_1 X + \epsilon$.) For the purpose of statistical inference, the error terms are also assumed to be independent and identically distributed (*iid*) according to a normal distribution with mean 0 and variance σ^2. Thus, $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. We later discuss how to test these assumptions and what to do when one or more of the assumptions appears to be violated.

Recall that a random variable Z is normally distributed if it has the following density function:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(z-\mu)^2}{2\sigma^2}},$$

where $-\infty < \mu < +\infty$ is the mean of Z and $\sigma > 0$ is the standard deviation of Z. Furthermore, a standard normal random variable is the special case where Z has $\mu = 0$ and $\sigma = 1$. While we generally do not utilize the normal distribution directly, it is at the root of many of our regression calculations. However, we will return to the normal distribution function later when we discuss the correlation model.

The term "linear" is used in the above discussion to indicate the behavior of the regression coefficients (i.e., β_0 and β_1). An example of a model that is nonlinear in the parameters is:

$$y_i = \frac{\beta_0}{\beta_1 - \beta_0}[e^{-\beta_2 x_i} - e^{-\beta_1 x_i}] + \epsilon_i.$$

Analysis of nonlinear regression models requires more advanced techniques which we will not cover here. Since our focus is on linear regression, the term "linear" may be dropped throughout the text.

2.3 Sample Estimates of the Regression Model

The first step in a regression problem is to estimate the model. The standard mathematical criterion used is ordinary least squares, which is short for the least sum of squared errors. The "best" sample estimate of the regression equation is the equation for which the observed sample has the smallest sum of squared errors. To illuminate these statements, we introduce some more notation:

- $\hat{y}_i = b_0 + b_1 x_i$; $\hat{y}_i$ is the **predicted value** (or **predicted fit**) of y for the i^{th} observation in the sample.
- $e_i = y_i - \hat{y}_i$; e_i is the **observed error** (or **residual**) for the i^{th} observation in the sample. This is calculated by taking the difference between the observed and predicted values of y for the i^{th} observation in the sample.
- $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$; SSE is the **sum of squared observed errors** for all observations in a sample of size n.

The "best" regression line for a sample *minimizes* the SSE. Letting $\text{S} = \sum_{i=1}^{n} \epsilon_i^2$ and using Calculus, first differentiate with respect to each of the regression coefficients. This yields the system of equations

$$\frac{\partial \text{S}}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$
$$\frac{\partial \text{S}}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i).$$

Setting the above equal to 0 and replacing (β_0, β_1) by (b_0, b_1) yields the **normal equations**

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$
$$\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0.$$

The sample estimates (b_0, b_1) in the normal equations are called the **ordinary least squares estimates**. Thus, the regression line determined by (b_0, b_1) is called the ordinary least squares line. The solutions for the normal equations are:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$b_0 = \bar{y} - b_1\bar{x}$$

and a couple consequences of these equations are that $\sum_{i=1}^{n} e_i = 0$ and that b_0 and b_1 are unbiased estimates of their population quantities (i.e., $\text{E}(b_0) = \beta_0$

and $\mathrm{E}(b_1) = \beta_1$).[1] Also note that the ordinary least squares line goes through the point $(\bar{x}, \bar{y})$. Do you see why?

One other notational convention often used in the literature concerns the following sums of squares:

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

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

So for example, the ordinary least squares estimate for the slope can be written as:

$$b_1 = \frac{S_{xy}}{S_{xx}}.$$

We will primarily use this notation only in the simple linear regression setting as we will yield to the more succinct matrix notation for multiple linear regression.

2.4 Measuring Overall Variation from the Sample Line

Next we discuss some measures of overall variation from the sample regression line.

$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2}$ is called the **mean squared error** for simple regression. MSE is the **sample variance** of the errors (residuals) and estimates σ^2, the population variance for the errors. It is important to note that the divisor $n - 2$ only applies to simple regression. The general rule is that the divisor is $n - p$, where p is the number of parameters in the regression equation. For a straight-line model, we estimate $p = 2$ coefficients, the intercept and the slope.

[1] A famous result is the Gauss-Markov Theorem, which says that the ordinary least squares estimates are the "best" linear unbiased estimates. We formalize this theorem later after we introduce the multiple linear regression setting and matrix notation for regression models.

The **sample standard deviation** of the errors (residuals) from the regression line is given by $s = \sqrt{\text{MSE}}$. The value of s can be interpreted (roughly) as the average absolute size of deviations of individuals from the sample regression line.

Let $\bar{y}$ be the mean of all observed y values. Then, the **total sum of squares** is given by $\text{SSTO} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

2.4.1 R^2

We now define the **coefficient of determination**:

$$R^2 = \frac{\text{SSR}}{\text{SSTO}} = \frac{\text{SSTO} - \text{SSE}}{\text{SSTO}},$$

which is the proportion of variation in y that is explained by x. This value is often expressed as a percentage by multiplying by 100. For example, if your calculation shows that the proportion of variation in y explained by x is 0.437, then it is also correct to say that 43.7% of the total variation in y is explained by x.

While R^2 is a popular measure of how well the regression model fits the data, it should not be used solely to assess the model's adequacy without further justification. Some caveats regarding the use of R^2 include:

1. The value of R^2 is highly sensitive to the sample size.

2. The value of R^2 can be increased by adding more predictors to the model (as we will see when we discuss multiple linear regression). However, this can cause the unwanted situation of an increase in the MSE. This often occurs when dealing with small sample sizes.

3. R^2 is influenced by the range of the predictors in that if the range of X increases or decreases, then R^2 increases or decreases, respectively.

4. The magnitude of the slopes is not measured by R^2.

5. R^2 is only the measure of the *strength* of the linear component of a model. For example, suppose the relationship between the response and predictor is measured perfectly by $Y = \cos(X)$. Then, the value of R^2 will be very small even though the two variables have a perfect relationship.

6. A high or low level of R^2 does not indicate one way or the other about the predictability of the model.

2.5 Regression Through the Origin

Consider a generic assembly process. Suppose we have data on the number of items produced per hour along with the number of rejects in each of those time spans. If we have a period where no items were produced, then there are obviously 0 rejects. Such a situation may indicate deleting β_0 from the model since β_0 reflects the amount of the response (in this case, the number of rejects) when the predictor is assumed to be 0 (in this case, the number of items produced). Thus, the model to estimate becomes

$$y_i = \beta_1 x_i + \epsilon_i,$$

which is called a **regression through the origin** (or **RTO**) model.

The estimate for β_1 when using the regression through the origin model is:

$$\hat{\beta}_{\text{RTO}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Thus, the estimated regression equation is

$$y_i = \hat{\beta}_{\text{RTO}} x_i,$$

Note that we no longer have to center (or "adjust") the x_i's and y_i's by their sample means (compare this estimate for $\hat{\beta}_1$ to that of the estimate found for the simple linear regression model). Since there is no intercept, there is no correction factor and no adjustment for the mean (i.e., the regression line can only pivot about the point (0,0)).

Generally, a regression through the origin is not recommended due to the following:

1. Removal of β_0 is a strong assumption which forces the line to go through the point (0,0). Imposing this restriction does not give ordinary least squares as much flexibility in finding the line of best fit for the data.

2. Generally, $\sum_{i=1}^{n} e_i \neq 0$. Because of this, the SSE could actually be larger than the SSTO, thus resulting in $R^2 < 0$.

- Since R^2 can be negative, the same interpretation of this value as a measure of the strength of the linear component in the simple linear regression model cannot be used here.

If you strongly believe that a regression through the origin model is appropriate for your situation, then statistical testing can help justify your decision (to be discussed later). Moreover, if data has not been collected near $X = 0$, then forcing the regression line through the origin is likely to make for a worse-fitting model. So again, this model is not usually recommended unless there is a strong belief that it is appropriate.

2.6 Distinguishing Regression from Correlation

The **(Pearson) correlation coefficient** (often denoted by ρ) is a bounded index (i.e., $-1 \leq \rho \leq 1$) which provides a unitless measure for the strength and direction of the association between two variables. For instance, suppose we wish to examine the relationship between the two random variables X and Y and we find that $\rho = -0.93$. What this says is that there is a strong negative relationship between these two variables (i.e., as the value of one of the variables *increases*, the value of the other variable tends to *decrease*). Thus, a value of ρ close to 0 indicates that there is no association between the two variables while values close to -1 or $+1$ indicate strong negative or strong positive associations, respectively.

In terms of estimation, suppose we have the samples $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$. Then, the sample Pearson correlation coefficient is given by

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \end{aligned}$$

which provides an estimate of the population parameter ρ. Inference procedures can also be carried out on this quantity, but we will not explore those details here.

So how does the correlation coefficient measurement differ from what we have presented with linear regression? Unlike regression, correlation is

treating the two variables X and Y as having a random bivariate structure. Correlation only establishes what kind of association possibly exists between these two random variables. However, regression is concerned with treating Y as a random variable while fixing X (i.e., Y *depends* on X). Subsequently, a regression analysis provides a model for a cause and effect type of relationship, while correlation simply provides a measure of association. Correlation does not necessarily imply causality. Also note that the estimated slope coefficient (b_1) and the estimated correlation coefficient (r) are related in the following way:

$$\begin{aligned} r &= b_1\sqrt{\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{\sum_{i=1}^n(y_i-\bar{y})^2}} \\ &= b_1\sqrt{\frac{S_{xx}}{S_{yy}}}. \end{aligned}$$

There are various correlation statistics and the Pearson correlation coefficient is probably the most commonly used measure since it is a parametric measure. If a correlation coefficient is presented in a report and the type of coefficient used is not specified, then it is likely that it is the Pearson correlation coefficient being reported. Some common non parametric measures of association include:

- **Spearman's rank correlation coefficient** (often denoted by θ), which measures the association based on the ranks of the variables. The estimate for this measure is given by

$$\hat{\theta} = \frac{\sum_{i=1}^n(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum_{i=1}^n(R_i-\bar{R})^2\sum_{i=1}^n(S_i-\bar{S^2})}},$$

  where R_i and S_i are the rank of the x_i and y_i values, respectively.

- **Hoeffding's D statistic**, which detects more general departures from independence by measuring the distance between the joint cumulative distribution function of X and Y and the product of their respective marginal distribution functions. This statistic is calculated as

$$D^* = 30\left[\frac{(n-2)(n-3)D_1+D_2-2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}\right],$$

where $D_1 = \sum_{i=1}^{n}(Q_i - 1)(Q_i - 2)$, $D_2 = \sum_{i=1}^{n}(R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$, and $D_3 = \sum_{i=1}^{n}(R_i - 2)(S_i - 2)(Q_i - 1)$. Here, R_i and S_i are defined as above and Q_i is 1 plus the number of points with both x and y values less than the i^{th} point.

- **Kendall's $\boldsymbol{\tau_b}$**, which is based on the number of concordant and discordant pairs between the ordinal variables X and Y (ordinal variables are categorical variables where there is an implied meaning in the order of the categories - for example, if asked how you feel today on a scale of 1 to 5 with 1 being "miserable" and 5 being "great", the order has an implied meaning). **Concordant pairs** occur when a pair of observations are in the same direction and **discordant pairs** occur when a pair of observations are in the opposite direction. For all (X_i, Y_i) and (X_j, Y_j) pairs (with $i < j$), concordance occurs if the relationship between X_i and X_j are in the same direction as Y_i and Y_j and discordance occurs if the relationships are in opposite directions. This statistic is calculated as
$$\hat{\tau}_b = \frac{\sum_{i<j} \texttt{sgn}(x_i - x_j)\texttt{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}},$$
where $T_0 = \frac{1}{2}n(n-1)$, $T_1 = \frac{1}{2}\sum_k t_k(t_k - 1)$, and $T_2 = \frac{1}{2}\sum_l u_l(u_l - 1)$. The t_k and u_l are the number of tied values in the k^{th} group of tied x values and the l^{th} group of tied y values, respectively.

- For a sample of size n, other measures based strictly on the number of concordant pairs (n_c), the number of discordant pairs (n_d), and the number of tied pairs (n_t) are:
  - **Somer's D**:
  $$D_S = \frac{n_c - n_d}{n_c + n_d + n_t}.$$
  - **Goodman-Kruskal Gamma**:
  $$\gamma_{GK} = \frac{n_c - n_d}{n_c + n_d}.$$
  - **Kendall's $\boldsymbol{\tau_a}$**:
  $$\hat{\tau}_a = \frac{2(n_c - n_d)}{n(n-1)}.$$

2.7 The Regression Effect

The **regression effect** (or **regression towards the mean**)[2] is the phenomenon that if a variable is extreme on its first measurement, then it will tend to be closer to the average on a second measurement. Moreover, if it is extreme on a second measurement, then it will tend to have been closer to the average on the first measurement (which superficially seems paradoxical). To avoid making wrong inferences, the possibility of regression toward the mean must be considered when designing experiments and interpreting experimental, survey, and other empirical data in the physical, life, behavioral and social sciences.

A common, simple example is about a class of students that takes a 100-item true/false test on a subject. Suppose that all students choose randomly on all questions. Then, each student' score would be a realization of one of a set of independent and identically distributed random variables, with a mean of, say, 50. Naturally, some students will score substantially above 50 and some substantially below 50 just by chance. If one takes only the top scoring 10% of the students and gives them a second test on which they again choose randomly on all items, the mean score would again be expected to be close to 50. Thus the mean of these students would "regress" all the way back to the mean of all students who took the original test. No matter what a student scores on the original test, the best prediction of his score on the second test is 50. If there were no luck or random guessing involved in the answers supplied by students to the test questions then all students would score the same on the second test as they scored on the original test, and there would be no regression toward the mean.

2.7.1 The Regression Fallacy

The **regression fallacy** occurs when the regression effect is mistaken for a real treatment effect. The regression fallacy is often observed where there is no overall treatment effect, which usually results in further, unnecessary subset analysis. Some examples of statements committing a regression fallacy are as follows:

[2]The conditions under which regression toward the mean occurs depend on the way the term is mathematically defined. Sir Francis Galton first observed the phenomenon in the context of simple linear regression of data points collected from his experiments on the size of the seeds of successive generations of sweet peas.

1. *When a subject's pain became unmanageable, he went to the doctor. After the visit, his pain subsided. Therefore, the subject benefited from the doctor's treatment.* The pain subsiding after it has gotten worse is more easily explained by regression towards the mean. Assuming it was caused by the doctor is fallacious.

2. *A student did exceptionally poorly last semester, so they were punished. He did much better this semester. Therefore, punishment is effective in improving students' grades.* Often exceptional performances are followed by more normal performances, so the change in performance might better be explained by regression towards the mean.

3. *The frequency of accidents on a road fell after a speed camera was installed. Therefore, the speed camera has improved road safety.* Speed cameras are often installed after a road incurs an exceptionally high number of accidents, and this value usually falls (regression towards the mean) immediately afterwards.

Another popular variant of the regression fallacy occurs when subjects are enrolled into a study on the basis of an extreme value of some measurement and a treatment is declared effective because subsequent measurements are not as extreme. Similarly, it is fallacious to take individuals with extreme values from one measuring instrument (a food frequency, say), reevaluate them using a different instrument (a diet record), and declare the instruments to be biased relative to each other because the second instrument's measurements are not as extreme as the first's. The regression effect guarantees that such results must be observed in the absence of any treatment effect or bias between the instruments.

While the regression effect is real and complicates the study of subjects who are initially extreme on the outcome variable, it does not make such studies impossible. Randomization and controls are usually enough to safeguard against this outcome. Consider a study of subjects selected for their initially high blood pressure measurement who are enrolled in a controlled diet trial to lower it. Regression to the mean says even the controls will show a decrease over the course of the study, but if the treatment is effective the decrease will be greater in the treated group than in the controls. This is also something that can be assessed through a topic we discuss later called analysis of covariance.

2.8 Examples

Example 1: Made-Up Data
We will first just analyze a set of made-up data in order to illustrate the concepts we provided thus far. The data is provided in Table 2.2.

x_i	4	4	7	10	10
y_i	15	11	19	21	29
$\hat{y}_i$	13	13	19	25	25
e_i	15-13=2	11-13=-2	19-19=0	21-25=-4	29-25=4

Table 2.2: The made-up data and relevant calculations.

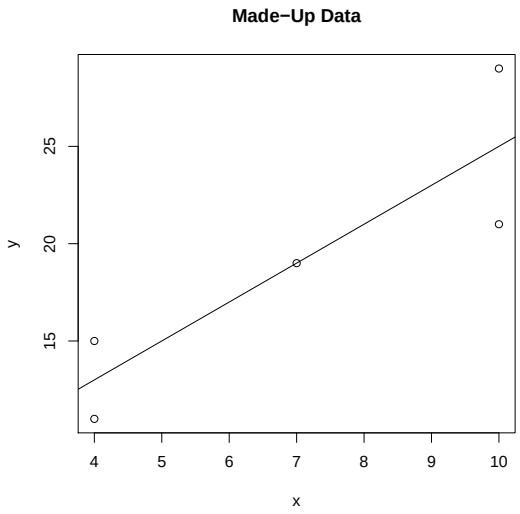


Figure 2.1: A plot of the made-up data along with the ordinary least squares fit.

The least squares regression line for this sample data is $\hat{y}_i = 5+2x_i$ (a plot of this data along with this ordinary least squares fit is given in Figure 2.1). Notice in Table 2.2 that there are additional calculations provided based on this estimated equation. The third row shows predicted values, determined by substituting the value of x_i into the estimated equation for each i (i.e.,

$\hat{y}_i = 5+2x_i$). The fourth row shows residuals, calculated as $e_i = y_i - \hat{y}_i$, such that $i = 1, \ldots, 5$.

Relevant calculations on this data set include:

- The sum of squared errors is $\text{SSE} = 2^2 + (-2)^2 + 0^2 + (-4)^2 + 4^2 = 40$, which is the value minimized in order to obtain the regression coefficient estimates of $b_0 = 5$ and $b_1 = 2$.
- $\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{40}{3} = 13.33$, which estimates the unknown value of σ^2, the population variance for the errors.
- $s = \sqrt{\text{MSE}} = \sqrt{13.33} = 3.65$ is an estimate of σ, the unknown value of the population standard deviation of the errors.
- To calculate SSTO, first calculate $\bar{y} = (15+11+19+21+29)/5 = 19$. Then

  $$\text{SSTO} = (15-19)^2+(11-19)^2+(19-19)^2+(21-19)^2+(29-19)^2 = 184.$$

- Finally, $R^2 = \frac{\text{SSTO}-\text{SSE}}{\text{SSTO}} = \frac{184-40}{184} = 0.783$. So roughly 78.3% of the variation in y is explained by x.
- We can also calculate $r = +0.8847$, which indicates a strong, positive association between the two variables. R^2 is often reported with correlation procedure because it is also a measure of association and is simply the square of the sample correlation coefficient (i.e., $R^2 = r^2$).

Example 2: Steam Output Data

This data set of size $n = 25$ contains observations taken at a large steam plant. The variables are y = steam used per month and x = average atmospheric pressure in degrees Fahrenheit. Table 2.3 gives the data used for this analysis.

Figure 2.2 is a plot of the data with the least squares regression line overlaid. Some noticeable features from the scatterplot in Figure 2.2 is that the relationship appears to be linear and it appears that there is a negative slope. In other words, as the atmospheric pressure increases, then the steam used per month tends to decrease.

Next, we fit a simple linear regression line to this data using ordinary least squares. Basic output pertaining to an ordinary least squares analysis includes the following:

i	y_i	x_i	i	y_i	x_i	i	y_i	x_i
1	10.98	35.3	10	9.14	57.5	19	6.83	70.0
2	11.13	29.7	11	8.24	46.4	20	8.88	74.5
3	12.51	30.8	12	12.19	28.9	21	7.68	72.1
4	8.40	28.8	13	11.88	28.1	22	8.47	58.1
5	9.27	61.4	14	9.57	39.1	23	8.86	44.6
6	8.73	71.3	15	10.94	46.8	24	10.36	33.4
7	6.36	74.4	16	9.58	48.5	25	11.08	28.6
8	8.50	76.7	17	10.09	59.3			
9	7.82	70.7	18	8.11	70.0			

Table 2.3: The atmospheric pressure data. i corresponds to the observation number.

```
##########
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.05668    0.65534  19.923 5.26e-16 ***
temp        -0.07067    0.01207  -5.855 5.75e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 23 degrees of freedom
Multiple R-Squared: 0.5985,     Adjusted R-squared: 0.581
F-statistic: 34.28 on 1 and 23 DF,  p-value: 5.745e-06
##########
```

The regression line fitted to this data is $\hat{y}_i = 13.0567 - 0.0707x_i$. Recall that we use the "hat" notation to denote estimates. In this case, we interpret the equation as saying that the predicted amount of steam used per month decreases, on average, by about 0.0707 pounds for every 1 degree Fahrenheit increase in temperature.

Interpretation of the estimated regression coefficients is:

- The interpretation of the slope (value=-0.0707) is that the predicted amount of steam used per month decreases, on average, by about 0.0707 pounds for every 1 degree Fahrenheit increase in temperature.

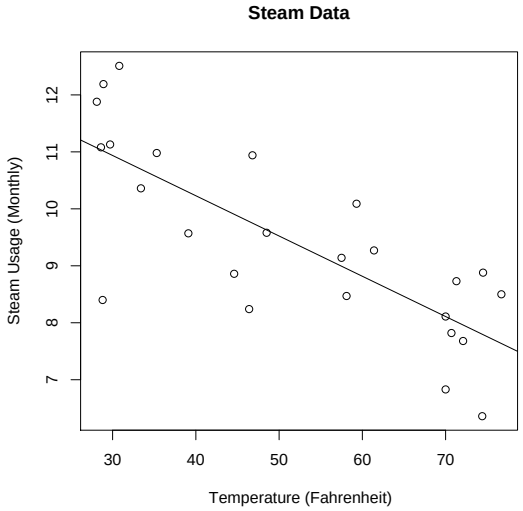


Figure 2.2: This is a scatterplot of y on x with the least squares regression line overlaid. What type of relationship do you see?

- The interpretation of the intercept (value=13.0567) is that if the temperature was 0 degrees Fahrenheit, then the average amount of steam used per month would be about 13.0567 pounds.

Some other statistics from the output worth noting are:

- The value of $s = \sqrt{\text{MSE}} = 1.0555$ tells us roughly the average difference between the y values of individual observations and predictions of y based on the regression line.
- The value of R^2 can be interpreted to mean that atmospheric temperature explains 59.85% of the observed variation in the steam used per month.
- We can obtain the value of the correlation coefficient r by hand:

$$\begin{aligned} r &= \texttt{sgn}(b_1)\sqrt{R^2} \\ &= -\sqrt{0.5985} \\ &= -0.7736. \end{aligned}$$

Thus, there is a strong, negative relationship between steam used per month and the atmospheric temperature.

- The other values given in the output will be discussed in later chapters.

Chapter 3

Statistical Inference

Statistical inference concerns statistical methods for using sample data to make judgments about a larger population. Two statistical inference tools we discuss are:

1. hypothesis testing

2. confidence intervals.

3.1 Hypothesis Testing and Confidence Intervals

Hypothesis testing requires the formulation of a null and alternative hypotheses which specify possibilities for the value(s) of one or more population parameters. The **null hypothesis** (denoted by H_0) is the hypothesis being tested and is usually a statement regarding no change or difference in the situation at hand (i.e., the status quo). The **alternative hypothesis** (denoted by H_A or H_1) is the anticipated situation should the null hypothesis be false (i.e., what we usually hope to show). Occasionally, one tests multiple alternative hypotheses against a single null hypothesis.

After establishing H_0 and H_A, a **test statistic** is then calculated using the sample data (we will introduce the general formula for a test statistic later in this chapter). The value of the test statistic is affected by the degree to which the sample supports one or the other of the two hypotheses. Two completely equivalent strategies for making a decision based on the test statistic are:

1. The **p-value** approach, which is used by all statistical software. We find the probability that the test statistic would be as extreme as the value found, if the null hypothesis were true (some probabilities associated with certain distributions can be found in Appendix B). We decide in favor of the alternative hypothesis when the p-value is less than the **significance level** (or **α-level**). The correct way to phrase your decision is to state that "we reject the null hypothesis if $p < \alpha$" or "fail to reject the null hypothesis if $p \geq \alpha$". The significance level is usually set at $\alpha = 0.05$, in which case we say a result is **statistically significant** if $p < 0.05$ and a result is **marginally significant** if $0.05 \leq p < 0.10$.

2. The **critical value** (or **critical region**) approach, which is what some textbooks use. We decide in favor of the alternative hypothesis when the value of the test statistic is more extreme than a critical value. The critical region is such that if the null hypotheses were true, the probability that the test statistic ends up in the critical region is the α-level (significance level) of the test. The α-level, again, is usually $\alpha = 0.05$.

Once we have used one of the two methods above, we can also include a **decision rule**, which is a statement in terms of the test statistic as to which values result in rejecting or failing to reject the null hypothesis.

A $100\times(1-\alpha)\%$ **confidence interval** is an interval of values that is likely to include the unknown value of a population parameter. The **confidence level** is the probability that the procedure used to determine the interval will provide an interval that "captures" the population value. As an example, suppose $\alpha = 0.05$ which corresponds to the 95% confidence level. Then, the calculated confidence interval will "capture" the population value for about 95% of all random samples. The way we interpret this for the interval (a, b) is by saying "with 95% confidence, we expect the true value to be between a and b". It is **INCORRECT** to say "there is a 95% chance that the true value will be between a and b". The reason why is because the true value (which is unknown) is either in or out of the interval (a, b), which would correspond to a probability of 1 or 0, respectively, to belonging to the interval. Thus, always make sure you use the former statement and not the latter when interpreting a confidence interval!

A general format for a confidence interval in many situations is

sample statistic $\pm$ (multiplier $\times$ standard error of the statistic).

1. The **sample statistic** is the value that estimates the parameter of interest. For example, it might be the sample slope (b_1) in regression.

2. The **multiplier** is determined by the confidence level and a relevant probability distribution. Again, some of these can be found in Appendix B.

3. The **standard error** of the statistic is a measure of the accuracy of the statistic (as an estimate of the true population value). The "by hand" formula is different for different types of statistics. We normally rely on software to provide this value.

We have just introduced some of the basic terminology regarding hypothesis testing and confidence intervals. At the end of this chapter, we will use an example to illustrate these concepts with the proper statistical notation as well as some calculations.

3.2 Power

In hypothesis testing, you can commit one of two types of errors:

- A **Type I error** is when you reject the null hypothesis (H_0) when it is actually true (also called a false positive).

- A **Type II error** is when you fail to reject the null hypothesis when it is actually false (also called a false negative).

Earlier we introduced the significance level α, which is actually the probability of making a Type I error. Hypothesis tests are constructed to minimize the probability of committing a Type II error. In other words, we wish to maximize 1 minus the probability of committing a Type II error, which is called the **power** of a test.[1] Various factors affecting the power of a test are:

[1]The power of a statistical test is also written as $1 - \beta$. However, do not get this value of β confused with the β used to represent a regression coefficient!

- n: Increasing the sample size provides more information regarding the population and thus increases power.
- α: A larger α increases power because you are more likely to reject the null hypothesis with a larger α.
- σ: A smaller σ results in easier detection of differences, which increases power.
- population effect (δ): The more similar populations are, the more difficult it becomes to detect differences, thus decreasing power.

The two types of errors and their associated probabilities help us interpret α and $1 - \beta$. The distinction made between tests of significance and tests as decision rules between two hypothesis really lies behind the meaning of the calculations. In significance testing, we focus on a single hypothesis (H_0) along with a single probability (the p-value).[2] The goal is to measure the strength of evidence from the sample against the null hypothesis. Subsequent power calculations can then check the sensitivity of the test. If we fail to reject the null hypothesis, then we can only claim there is not sufficient evidence against H_0, *not* that it is actually true. If looking at this as a decision problem, then we construct a decision rule for deciding between the two hypotheses based on sample evidence. We therefore focus equally on two probabilities, which are the probabilities of the two types of errors. We must decide upon one hypothesis or the other. Figure 3.1 shows the types of errors and decisions in hypothesis testing.

The basic thought process is to first state H_0 and H_A in terms of a test of significance. Then, think of this as a decision problem so that the probabilities of a Type I error and Type II error are incorporated. Statisticians view Type I errors as more serious, so choose a significance level (α) and consider only tests with probability of a Type I error no greater than α. Then, among all of these tests, select the one that maximizes the power (i.e., minimizes a Type II error).

A common example illustrating power is to consider a legal trial where the null hypothesis would be that the individual on trial is innocent versus the alternative hypothesis that the individual is guilty. In the legal system, an individual is innocent until proven guilty. Hypothesis testing must show statistically significant evidence in order to reject the null hypothesis. Similarly,

[2]Multiple hypothesis can be tested, but this is a more advanced topic.

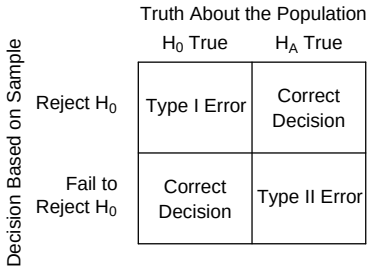


Figure 3.1: The two types of errors that can be made in hypothesis testing. Remember that the probability of making a Type I error is α and the probability of making a Type II error is $1 - \beta$.

the justice system must show beyond a reasonable doubt that the individual is guilty. It is often viewed as more serious to find someone guilty given that they are innocent (a Type I error) as opposed to finding someone not guilty who is actually guilty (a Type II error). The Type I error in this case also means that the truly guilty individual is still free while an innocent person has been convicted. This justice example illustrates a case where a Type I error is more serious, and while Type I errors usually are more serious errors, this is not true in all hypothesis tests.

In the simple linear regression setting, we are interested in testing that the slope is 0 versus the alternative that it is not equal to 0. This is written as:

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

More generally, one could also test

$$H_0 : \beta_1 = \beta_1^*$$
$$H_A : \beta_1 \neq \beta_1^*,$$

where β_1^* is any real number. However, β_1^* is typically 0. (We will go into more detail about constructing such hypothesis tests in the example that follows.)

The power of this test is calculated by first finding the tabled $100 \times (1-\alpha)^{\text{th}}$ percentile of the $F_{1,n-2}$ distribution. Some of these values are given in Table E.4 of Appendix E. Next we calculate $F_{1,n-2;1-\alpha}(\delta)$, which is the $100 \times (1-\alpha)^{\text{th}}$ percentile of a non-central $F_{1,n-2}$ distribution with non-centrality parameter δ. This is essentially a shifted version of the $F_{1,n-2}$-distribution, but is not tabulated in Appendix E. The non-centrality parameter is calculated as:

$$\delta = \frac{b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\text{MSE}}.$$

If β_1^* is a number other than 0 in the hypothesis test, then we would replace b_1 by $(b_1 - \beta_1^*)$ in the equation for δ. Finally, power is simply the probability that the calculated $F_{1,n-2;1-\alpha}(\delta)$ value is greater than the calculated $F_{1,n-2;1-\alpha}$ value under the $F_{1,n-2}(\delta)$ distribution.

We can also test

$$H_0 : \beta_0 = \beta_0^*$$
$$H_A : \beta_0 \neq \beta_0^*,$$

where β_0^* is any real number. We can find the power in a similar manner, but where the non-centrality parameter is defined as:

$$\delta = \frac{(b_0 - \beta_0^*)^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\text{MSE}}.$$

Luckily, statistical software will calculate these quantities for us!

There is also a relationship between the t-distribution and F-distribution which can be exploited here. If the random variable Z is distributed as t_n, then Z^2 is distributed as $F_{1,n}$. Because of this relationship, we can calculate the power in a similar manner using the t-distribution and non-central t-distribution. However, the non-central t-distribution has non-centrality parameter $\delta^* = \sqrt{\delta}$, such that δ is as defined above.

3.3 Inference on the Correlation Model

Earlier we introduced the (Pearson) correlation coefficient ρ. Now we will formally develop this quantity in the context of a probability model. Suppose

now that we no longer have the X values at fixed constants. Instead we have two random variables, say, Y_1 and Y_2. For example, suppose these random variables are the height and weight of a person, respectively. The objective will be to develop a formal probability model to allow us to make inferences on the correlation coefficient.

The **correlation model** assumes that the random variables Y_1 and Y_2 are jointly normally distributed where their joint distribution is the following bivariate normal distribution:

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]\right\},$$

where

1. $-\infty < \mu_1 < +\infty$ is the mean of Y_1 and $\sigma_1 > 0$ is the standard deviation of the marginal distribution of Y_1;
2. $-\infty < \mu_2 < +\infty$ is the mean of Y_2 and $\sigma_2 > 0$ is the standard deviation of the marginal distribution of Y_2; and
3. ρ is the coefficient of correlation between the random variables Y_1 and Y_2.

Note that if Y_1 and Y_2 are jointly normally distributed, then you can integrate out the other variable in the bivariate normal distribution above, which will yield the marginal distributions of Y_1 and Y_2 (both of which are normal distributions with the parameters specified above).

When the population is bivariate normal, it is often of interest to test the following hypothesis:

$$H_0 : \rho = 0$$
$$H_A : \rho \neq 0.$$

This test is of interest because if $\rho = 0$, then this implies that Y_1 and Y_2 are independent of each other. The test statistic for this test is

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

which is distributed according to a t_{n-2} distribution. Thus you can get a p-value just as in the case of testing the regression coefficients.

However, a confidence interval for ρ is actually more complicated since the sampling distribution of r is complex. The $(1-\alpha/2)\times 100\%$ confidence intervals for ρ are estimated using the following transformation, which is known as **Fisher's z transformation**:

$$z' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right).$$

When $n \geq 25$, then

$$\mathrm{E}(z') = \zeta = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$
$$\mathrm{Var}(z') = \frac{1}{n-3}.$$

Moreover, the standardized statistic

$$\frac{z'-\zeta}{\sqrt{\mathrm{Var}(z')}}$$

is approximately a standard normal random variable. This yields the following (approximate) $(1-\alpha/2)\times 100\%$ confidence intervals for ζ:

$$z' \pm z^*_{1-\alpha/2}\sqrt{\frac{1}{n-3}},$$

where $z^*_{1-\alpha/2}$ is the $(1-\alpha/2)\times 100^{\text{th}}$ percentile of the standard normal distribution. We can then transform these limits back to limits for ρ using the definition of $\mathrm{E}(z')$. Then, the interpretation of this interval would be similar to how we have interpreted other confidence intervals.

Note that we have the stipulation that $n \geq 25$. The above approximation can be used for $n < 25$, but you will likely have very wide intervals. There are other procedures out there for calculating these intervals (e.g., by simulation), but for our purposes, the method outlined above will be sufficient.

3.4 Example

Example: Steam Output Data (*continued*)
For this example, we will utilize the following output:

```
##########
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.05668    0.65534  19.923 5.26e-16 ***
temperature -0.07067    0.01207  -5.855 5.75e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 23 degrees of freedom
Multiple R-Squared: 0.5985,     Adjusted R-squared: 0.581
F-statistic: 34.28 on 1 and 23 DF,  p-value: 5.745e-06

                Power
(Intercept) 1.0000000
temperature 0.9998623
##########
```

Hypothesis Test for the Intercept (β_0)
This test is rarely a test of interest, but does show up when one is interested in performing a regression through the origin. For a hypothesis test about the intercept, the null and alternative hypotheses are written as:

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

In other words, the null hypothesis is testing if the population intercept is equal to 0 versus the alternative hypothesis that the population intercept is not equal to 0. In our example, the intercept is the mean steam used per month for a temperature of 0. In most problems, we are not particularly interested in hypotheses about the intercept. In particular, the intercept does not give information about how the value of y changes when the value of x changes. Nevertheless, to test whether the population intercept is 0 is as follows:

1. The sample intercept is $b_0 = 13.0567$.

2. The standard error (SE) of the sample intercept, written as s.e.(b_0), is 0.6553. The SE of any statistic is a measure of its accuracy. In this case, the s.e.(b_0) gives, very roughly, the average difference between the

sample b_0 and the true population intercept β_0, for random samples of this size (and with these x-values).

3. The test statistic is $t^* = b_0/\text{s.e.}(b_0) = 13.0567/0.6553 = 19.9248$.

4. The p-value for the test is $p = 5.26 \times 10^{-16}$, which is very small.

5. The power of the test is nearly 1.0000.

6. The decision at the 0.05 significance level is to reject the null hypothesis since $p < 0.05$. Thus, we conclude that there is statistically significant evidence that the population intercept is not equal to 0.

So how exactly is the p-value found? For simple regression, the p-value is determined using a t distribution with $n - 2$ degrees of freedom (df), which is written as t_{n-2}, and is calculated as $2 \times$ area past $|t^*|$ under a t_{n-2} curve since we have a two-sided alternative hypothesis (a one-sided alternative hypothesis is when "$\neq$" in H_A is replaced by either "$>$" or "$<$".). In this example, df $= 25 - 2 = 23$. The p-value region for the two-sided alternative is twice the unshaded region as shown in Figure E.2 (you can also imagine another unshaded region of the same area on the other side of this curve). The negative and positive versions of the calculated t^* provide the interior boundaries of the two unshaded regions. As the value of $|t^*|$ increases, the p-value (area in the unshaded regions) decreases.

Hypothesis Test for the Slope (β_1)

This test can be used to test whether or not x and y are related. The slope directly tells us about the link between the mean y and x. When the true population slope does not equal 0, the variables y and x are linearly related. When the slope is 0, there is not a linear relationship because the mean y does not change when the value of x is changed (i.e., you just have a straight line). For a hypothesis test about the slope, the null and alternative hypotheses are written as:

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

In other words, the null hypothesis is testing if the population slope is equal to 0 versus the alternative hypothesis that the population slope is not equal to 0. To test whether the population slope is 0, the following information is used:

1. The sample slope is $b_1 = -0.0707$.

2. The SE of the sample slope, written as s.e.(b_1), is 0.0121. Again, the SE of any statistic is a measure of its accuracy. In this case, the s.e.(b_1) gives, very roughly, the average difference between the sample b_1 and the true population slope β_1, for random samples of this size (and with these x-values).

3. The test statistic is $t^* = b_1/\text{s.e.}(b_1) = -0.0707/0.0121 = -5.8430$.

4. The p-value for the test is $p = 5.75 \times 10^{-6}$.

5. The power of the test is 1.0000.

6. The decision at the 0.05 significance level is to reject the null hypothesis since our $p < 0.05$. Thus, we conclude that there is statistically significant evidence that the variables of steam used per month and temperature are related.

Standard Errors of b_0 and b_1

The SE of each estimated regression coefficient measures, roughly, the average difference between the estimate and the true, unknown population quantity. These quantities are obtained by looking at the variances of the estimated regression coefficients and are easier to write in matrix notation (which we will discuss later). Regardless, it can be shown that the variances of b_0 and b_1 are

$$\begin{aligned}\text{Var}(b_0) &= \frac{\sigma^2 \sum_{i=1}^n x_i^2/n}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]\end{aligned}$$

and

$$\text{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Then, the square root of each of the above is the corresponding standard error of that estimate.

However, σ^2 is unknown, so we use the estimate MSE. This results in the following (estimated) standard error formulas:

$$\begin{aligned}\text{s.e.}(b_0) &= \sqrt{\frac{\text{MSE}\sum_{i=1}^n x_i^2/n}{\sum_{i=1}^n (x_i-\bar{x})^2}} \\ &= \sqrt{\text{MSE}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}\end{aligned}$$

and

$$\text{s.e.}(b_1) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (x_i-\bar{x})^2}}.$$

To be completely pedantic, we should probably put "hats" on the estimated standard errors (i.e., $\hat{\text{s.e.}}(b_0)$ and $\hat{\text{s.e.}}(b_1)$). However, it is rarely of interest to talk about the standard error quantities where σ^2 is known, so I will avoid the more burdensome notation and omit the hat as well as drop the word "estimated".

Confidence Interval for the Slope (β_1)

As mentioned earlier, inference procedures regarding β_0 are rarely of interest, so we will focus our attention on β_1. However, it should be noted that most of our discussion here is applicable to procedures concerning the intercept.

A confidence interval for the unknown value of the population slope β_1 can be computed as

$$\text{sample statistic } \pm \text{ multiplier } \times \text{ standard error of statistic}$$

$$\Rightarrow b_1 \pm t^*_{n-2;1-\alpha/2} \times \text{s.e.}(b_1).$$

This $t^*_{n-2;1-\alpha/2}$ multiplier is not the same value as the t^* value calculated for the test statistic.

To find the $t^*_{n-2;1-\alpha/2}$ multiplier, you can do one of the following:

1. In simple regression, the $t^*_{n-2;1-\alpha/2}$ multiplier is determined using a t_{n-2} distribution. The value of $t^*_{n-2;1-\alpha/2}$ is such that the confidence level is the area (probability) between $-t^*_{n-2;1-\alpha/2}$ and $+t^*_{n-2;1-\alpha/2}$ under the t_{n-2} curve.

- For example, suppose $\alpha = 0.05$ and $n = 8$. Then we want $t^*_{6;0.975}$ such that $1 - \alpha = 0.95$ is the area between $-t^*_{6;0.975}$ and $+t^*_{6;0.975}$. This is equivalent to finding the value of $+t^*_{6;0.975}$ such that all of the area under the curve to the right of this value is $1 - \alpha/2 = 0.975$. Hence, $t^*_{6;0.975} = 2.4469$ (see Figure E.2).
- Note that we write $1 - \alpha/2$ because we have a two-sided interval (i.e., we have a two-sided alternative hypothesis). If we have a one-sided alternative (which is rarely used in regression hypothesis testing), then we would use $1 - \alpha$.

2. Figure E.2 can be used to look up the $t^*_{n-2;1-\alpha/2}$ multiplier.

For the regression through the origin setting, a confidence interval for the unknown value of the population slope β_1 can be computed as

$$\Rightarrow \hat{\beta}_{\text{RTO}} \ \pm \ t^*_{n-1;1-\alpha/2} \ \times \ \text{s.e.}(\hat{\beta}_{\text{RTO}}),$$

where

$$\text{s.e.}(\hat{\beta}_{\text{RTO}}) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n x_i^2}}.$$

Note that the degrees of freedom for the t-multiplier are $n-1$ and *not* $n-2$.

Bonferroni Joint Confidence Intervals for (β_0, β_1)
Sometimes we may want to find the confidence interval for more than one parameter simultaneously. While we are usually only interested in a confidence interval for the slope, this section will serve as more of a foundation for this type of interval, which is more commonly used in the multiple regression setting.

Let A and B be two events with complements A^{C} and B^{C}, respectively. Then, **Bonferroni's Inequality** says that

$$\text{P}(\text{A} \cap \text{B}) \geq 1 - \text{P}(\text{A}^{\text{C}}) - \text{P}(\text{B}^{\text{C}}).$$

Now, let A be the event that the confidence interval for β_0 covers β_0 and B be the event that the confidence interval for β_1 covers β_1. If $\text{P}(\text{A}) = 1 - \alpha$ and $\text{P}(\text{B}) = 1-\alpha$, then by Bonferroni's Inequality, $\text{P}(\text{A} \cap \text{B}) \geq 1-\alpha-\alpha = 1-2\alpha$. Thus, to get the joint confidence interval with at least $100(1-\alpha)\%$ confidence, the Bonferroni joint confidence intervals for β_0 and β_1 are

$$b_0 \ \pm \ t^*_{n-2;1-\alpha/(2p)} \ \times \ \text{s.e.}(b_0) \Rightarrow b_0 \ \pm \ t^*_{n-2;1-\alpha/4} \ \times \ \text{s.e.}(b_0)$$

and

$$b_1 \pm t^*_{n-2;1-\alpha/(2p)} \times \text{s.e.}(b_1) \Rightarrow b_1 \pm t^*_{n-2;1-\alpha/4} \times \text{s.e.}(b_1),$$

respectively. Here, $p = 2$ corresponds to the number of parameters for which we are trying to jointly estimate confidence intervals. You can also imagine an extension to the general case of $q = n$ joint confidence intervals by an application of Bonferroni's Inequality for n sets:

$$\text{P}(\text{A}_1 \cap \text{A}_2 \cap \ldots \cap \text{A}_n) \geq 1 - \sum_{i=1}^{n} \text{P}(\text{A}_i).$$

95% Confidence Interval
In our example, $n = 25$ and $\text{df} = n - 2 = 23$. For 95% confidence, $t^*_{23;0.975} = 2.0687$ (even though $\text{df} = 23$ is not included in Figure E.2, see if you can understand why this value would be correct). A 95% confidence interval for β_1 , the true population slope, is:

$$\begin{gathered} -0.0707 \pm (2.0687 \times 0.0121) \\ \Rightarrow -0.0707 \pm 0.0250 \\ \text{or about } -0.0957 \text{ to } -0.0457. \end{gathered}$$

Interpretation: With 95% confidence, we can say the mean steam used per month decreases somewhere between 0.0457 and 0.0957 pounds per each degree increase in temperature. Remember, it is incorrect to say with 95% probability that the mean steam used per month decreases somewhere between 0.0457 and 0.0957 pounds per degree increase!

99% Confidence Interval
For 99% confidence, $t^*_{23;0.995} = 2.8073$. A 99% confidence interval for β_1 , the true population slope, is:

$$\begin{gathered} -0.0707 \pm (2.8073 \times 0.0121) \\ \Rightarrow -0.0707 \pm 0.0340 \\ \text{or about } -0.1047 \text{ to } -0.0367. \end{gathered}$$

Interpretation: With 99% confidence, we can say the mean steam used per month decreases somewhere between 0.0367 and 0.1047 pounds per each degree increase in temperature. Notice that as we increase our confidence,

the interval becomes wider. So as we approach 100% confidence, our interval grows to become the whole real line (i.e., $(-\infty, +\infty)$).

Bonferroni Joint 95% Confidence Intervals for (β_0, β_1)
For joint 95% confidence $t_{23;1-0.05/4} = t_{23;0.9875} = 2.3979$. Joint 95% confidence intervals for β_0 and β_1 are, respectively,

$$\begin{gathered} 13.0567 \pm (2.3979 \times 0.6553) \\ \Rightarrow 13.6230 \pm 1.5713 \\ \text{or about } \ 11.4787 \ \text{ to } \ 14.6213 \end{gathered}$$

and

$$\begin{gathered} -0.0707 \pm (2.3979 \times 0.0121) \\ \Rightarrow -0.0707 \pm -0.0290 \\ \text{or about } \ -0.0997 \text{ to } \ -0.0417. \end{gathered}$$

Interpretation: With 95% joint confidence, we can say that the true population intercept and slope terms are in the intervals (11.4787,14.6213) and (-0.0997, -0.0417), respectively.

The Duality Principle
Recall that the null hypothesis of interest is $H_0 : \beta_1 = 0$ and that at the $\alpha = 0.05$ significance level we rejected this null hypothesis. Also, the 95% confidence interval that was calculated did not include the null value of 0. In other words, we were 95% confident that our interval would not include 0 (in fact, it was entirely below 0). Thus, we arrived at the same statistical conclusion using both a hypothesis test and constructing a confidence interval. This illustrates what is called the **duality principle**, which means for a given hypothesis, the test and confidence interval for a given α will lead you to the same conclusion.

Practical Significance Versus Statistical Significance
For statistical inference, the larger your sample size n, the smaller your p-values. This means that a larger data set will typically yield a statistically significant predictor. However, is it an *important* effect? For example, suppose it had been established that for each degree increase in temperature, the amount of steam used per month should decrease by about 0.080 pounds (assume that no upper or lower limits were set on this quantity). If the data

set had been considerably larger and we obtained, say, an estimate of 0.078 pounds, this would still likely be acceptable as an average decrease in steam usage. However, the large sample size may claim that the plant's usage differs significantly from 0.080 pounds, while from a practical standpoint, this would likely be an acceptable amount.

Mainly, we test if a regression coefficient is equal to 0. In actuality, it is highly unlikely that a predictor which has been measured and analyzed will have no effect on the response. The effect may be minute, but it is not completely 0. One will need to defer to examining the practicality of building such a model and ask themselves if a regression coefficient is close to 0, but deemed statistically significant, does that mean anything in the context of the data? Or, if we fail to reject the null hypothesis, but a relationship between a predictor and response seems to make sense, did the size of the data affect the result? While this mainly boils down to an issue of sample size and the power of a test, it is suggested to always report confidence intervals which help with the bigger picture by giving probabilistic bounds of the estimates.

Chapter 4

Statistical Intervals

For this chapter, we will discuss two general groups of statistical intervals. The first group (confidence intervals) is concerned with estimating the mean $\mathrm{E}(Y)$ given a particular value or particular values of x. The second group (prediction and tolerance intervals) is concerned with predicting a new y given a particular value or particular values of x.

4.1 Intervals for a Mean Response

Confidence Intervals
A $100 \times (1-\alpha)\%$ **confidence interval** for $\mathrm{E}(Y|X = x_h)$ (i.e., the mean of the random variable Y) estimates the mean value of Y for individuals with a particular value of x. This confidence interval estimates the location of the line at a specific location of x. Note that here it will always be appropriate to say "$1-\alpha$".

Suppose we have x_h, a specified level of x, and the corresponding fit $\hat{y}_h = b_0 + b_1 x_h$. The standard error of the fit at x_h is given by the formula:

$$\text{s.e.}(\hat{y}_h) = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}. \tag{4.1}$$

Formula (4.1) applies only to simple linear regression. (We will not be calculating this quantity by hand thankfully!) The answer describes the accuracy of a particular $\hat{y}_h$ as an estimate of $\mathrm{E}(Y|X = x_h)$. It is also important to note that the subscript "i" is associated with the index of the "observed"

data set, while the subscript "h" is used to denote *any* possible level of x.[1]

A $100 \times (1-\alpha)\%$ confidence interval for $\mathrm{E}(Y)$ at x_h is calculated as

$$\hat{y}_h \pm t^*_{n-2;1-\alpha/2}\text{s.e.}(\hat{y}_h), \tag{4.2}$$

where the $t^*_{n-2;1-\alpha/2}$ multiplier is found by calculating the $(1-\alpha/2)^{\text{th}}-$percentile using a t_{n-2}-table. For example, if $n = 10$ and $\alpha = 0.05$, then $t^*_{8;0.975} = 2.3060$ from Table E.2.

Bonferroni Joint Confidence Intervals

Bonferroni joint $100 \times (1-\alpha)\%$ **confidence intervals** for $\mathrm{E}(Y|X = x_h)$ at q different values of x_h (i.e., $x_{h_1}, x_{h_2}, \ldots, x_{h_q}$) estimate the mean value of Y for individuals with these q different values of x. The confidence intervals estimate the location of the line at q specific locations of x.

Bonferroni joint $100\times(1-\alpha)\%$ confidence intervals for $\mathrm{E}(Y)$ at the values $x_{h_1}, x_{h_2}, \ldots, x_{h_q}$ are calculated as

$$\hat{y}_{h_i} \pm t^*_{n-2;1-\alpha/(2q)}\text{s.e.}(\hat{y}_{h_i}), \tag{4.3}$$

where $i = 1, 2, \ldots, q$ and $t^*_{n-2;1-\alpha/(2q)}$ is a multiplier found using a t_{n-2}-table.

Working-Hotelling Confidence Bands

A **Working-Hotelling** $100 \times (1-\alpha)\%$ **confidence band** for $\mathrm{E}(Y)$ at all possible values of $X = x_h$ estimates the mean value of Y for all different values of x_h. This confidence band contains the entire regression line (for all values of X) with confidence level $1-\alpha$.

A Working-Hotelling $100 \times (1-\alpha)\%$ confidence band for $\mathrm{E}(Y)$ at all possible values of $X = x_h$ is calculated as

$$\hat{y}_h \pm \sqrt{2F^*_{2,n-2;1-\alpha}}\text{s.e.}(\hat{y}_h), \tag{4.4}$$

where $F^*_{2,n-2;1-\alpha}$ is a multiplier found using a $F_{2,n-2}$-table.

[1] Note that the equivalent formula to (4.1) for the regression through the origin setting is

$$\text{s.e.}(\hat{y_h}) = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{x_h^2}{\sum_{i=1}^n x_i^2}\right)}.$$

Any of the intervals discussed in this chapter can be calculated for the regression through the origin setting by using this quantity in place of (4.1).

4.2 Intervals for a New Observation

Prediction Intervals

A $100 \times (1-\alpha)\%$ **prediction interval** for $Y = y$ estimates the value of y for an individual observation with a particular value of x. Equivalently, a prediction interval estimates the range of a future value for a response variable at a particular value of x given a specified confidence level. Since our data is just one sample, it is reasonable to consider taking another sample, which would yield different values for b_0 and b_1. This illustrates the sample variability in the slope and intercept of the regression line as well as the variability of the observations about the regression line. Since it will not always be feasible to get another sample of data, a prediction interval will allow us to assess the limits for the response at any given value of x (i.e., x_h). You can think of a prediction interval as pertaining to some characteristic of interest in a *future* sample of data from the population.

A prediction interval for $Y = y_h$ at x_h is calculated as

$$\hat{y}_h \pm t^*_{n-2;1-\alpha/2}\sqrt{\text{MSE} + [\text{s.e.}(\hat{y}_h)]^2}. \tag{4.5}$$

The $t^*_{n-2;1-\alpha/2}$ value is found in the same way as in the confidence interval. Also, if you look at the formulas for both the confidence intervals (4.2) and prediction intervals (4.6), you will see that a prediction interval will always be wider than a confidence interval at the same α-level. This is because a prediction interval considers a larger source of possible variation than do confidence intervals. Confidence intervals generally bound the estimate of a true parameter (e.g., a mean) around an average value. Thus, you are bounding a statistic and NOT an individual value.

It is also possible to construct a $100 \times (1-\alpha)\%$ prediction interval for the mean $\bar{Y} = \bar{y}_h$ when m new observations are drawn at the level x_h. These are calculated as

$$\hat{y}_h \pm t^*_{n-2;1-\alpha/2}\sqrt{\text{MSE}/m + [\text{s.e.}(\hat{y}_h)]^2}. \tag{4.6}$$

Note that the case of $m = 1$ is the original prediction interval from above.

Bonferroni Joint Prediction Intervals

Bonferroni joint $100 \times (1-\alpha)\%$ **prediction intervals** for $Y = y$ estimate the values of y for q different values of x. Equivalently, Bonferroni joint prediction intervals estimate the range of a future value for a response variable at a particular value of x given a specified confidence level.

Bonferroni joint $100 \times (1-\alpha)\%$ prediction intervals for $Y = y_{h_i}$ at x_{h_i} for $i = 1, 2, \ldots, q$ are calculated as

$$\hat{y}_{h_i} \pm t^*_{n-2;1-\alpha/(2q)} \sqrt{\text{MSE} + [\text{s.e.}(\hat{y}_{h_i})]^2}. \tag{4.7}$$

The $t^*_{n-2;1-\alpha/(2q)}$ value is found in the same manner as discussed earlier.

Scheffé Joint Prediction Intervals

Scheffé joint $100 \times (1-\alpha)\%$ **prediction intervals** for $Y = y$ estimate the values of y for q different values of x. Equivalently, Scheffé joint prediction intervals estimate the range a future value for a response variable at a particular value of x given a specified confidence level. As can be seen, Scheffé joint prediction intervals accomplish the same goal as Bonferroni joint prediction intervals - but it follows a slightly different formula (as to be shown). As to which interval to choose, you would want the interval that produces the tighter interval at the same confidence level.

Scheffé joint $100 \times (1-\alpha)\%$ prediction intervals for $Y = y_{h_i}$ at x_{h_i} for $i = 1, 2, \ldots, q$ are calculated as

$$\hat{y}_{h_i} \pm \sqrt{qF^*_{q,n-2;1-\alpha}(\text{MSE} + [\text{s.e.}(\hat{y}_{h_i})]^2)}. \tag{4.8}$$

The $F^*_{q,n-2;1-\alpha}$ value is found in the same manner as discussed earlier.

Tolerance Intervals

A $[100\times(1-\alpha)\%]/[100\times P\%]$ **tolerance interval** for y estimates the value of y for an individual with a particular value of x. Equivalently, for a given value of x, tolerance intervals pertain to a specified proportion of individual values (P) in the entire population with a specified degree of confidence ($1-\alpha$). As we can see, tolerance intervals, like prediction intervals, are concerned with probabilistic bounds of individual data points rather than of statistics.

To construct a one-sided tolerance interval for y at x_h, first set the values of α and P. Then, find the value of z^*_P (found in Figure E.1 of Appendix B) such that:

$$\begin{aligned} P &= \text{P}(Z \leq z^*_P) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z^*_P} e^{-\frac{x^2}{2}}\, dx. \end{aligned}$$

Next, calculate

$$n^* = \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]^{-1} = \frac{\text{MSE}}{\text{s.e.}(\hat{y}_h)^2}.$$

Finally, find $K_{\alpha,P}$ such that:

$$\text{P}(\sqrt{n^*} K_{\alpha,P} \leq t^*_{n-2;1-\alpha}(\delta)) = 1 - \alpha,$$

where $t^*_{n-2;1-\alpha}(\delta)$ is from a non-central t-distribution with $n-2$ degrees of freedom and non-centrality parameter $\delta = \sqrt{n^*} z^*_P$. (A non-central t-distribution is basically a t-distribution shifted from 0, but is not included in Appendix E.) Finally, the upper and lower one-sided tolerance intervals are calculated as

$$(-\infty, \hat{y}_h + K_{\alpha,P}\sqrt{\text{MSE}}) \tag{4.9}$$

and

$$(\hat{y}_h - K_{\alpha,P}\sqrt{\text{MSE}}, +\infty), \tag{4.10}$$

respectively.

Unlike a one-sided tolerance interval, two-sided tolerance intervals do not have an exact solution. Also, it would be incorrect to take the intersection of the two one-sided tolerance intervals just presented (although sometimes this appears to be a decent approximation). First, define $z^*_{(1+\omega)/2}$ such that

$$\frac{1}{\sqrt{2\pi}} \int_{-z^*_{(1+\omega)/2}}^{z^*_{(1+\omega)/2}} e^{(-\frac{x^2}{2})} dx = \omega,$$

where $\omega \in \{\alpha, P\}$ and let $f = n - 2$ (i.e., the error degrees of freedom). We also utilize χ^2_n, which is called a chi-square distribution with n degrees of freedom (a table for select values of n is given in Figure E.3). It can be shown that

$$K_{\alpha/2,P/2} = \begin{cases} z^*_{(1+P)/2}\left[\frac{f(1+1/n^*)}{\chi^{2*}_{f;1-\alpha}}\left\{1 + \frac{f-2-\chi^{2*}_{f;1-\alpha}}{2(n^*+1)^2}\right\}\right]^{\frac{1}{2}}, & f \leq n^{*2}\left(1 + \frac{1}{z^{*2}_{(1+\alpha)/2}}\right); \\ z^*_{(1+P)/2}\left[V\left\{1 + \frac{n^* V}{2f}\left(1 + \frac{1}{z^{*2}_{(1+\alpha)/2}}\right)\right\}\right]^{\frac{1}{2}}, & f > n^{*2}\left(1 + \frac{1}{z^{*2}_{(1+\alpha)/2}}\right) \end{cases}$$

such that

$$V = 1 + \frac{z^{*2}_{(1+\alpha)/2}}{n^*} + \frac{(3 - z^{*2}_{(1+P)/2})z^{*4}_{(1+\alpha)/2}}{6n^{*2}}.$$

For these equations, $\chi^{2*}_{f;1-\alpha}$ is the $100 \times (1-\alpha)^{\text{th}}$ percentile of the χ^2_f distribution. Finally an approximate two-sided tolerance interval is given by

$$\hat{y}_h \pm K_{\alpha/2,P/2}\sqrt{\text{MSE}}. \tag{4.11}$$

Calibration Intervals

Consider the calibration of a venturi that measures the flow rate of a chemical process. Let X denote the actual flow rate and Y denote a reading on the venturi. In this calibration study, the flow rate is controlled at n levels of X. Then, the corresponding Y readings on the venturi are observed. Suppose we assume the simple linear regression model with the standard assumptions on the error terms. Sometime in the future, the experimenter may be interested in estimating the flow rate from a particular venturi reading.

The method of **calibration** (or **inverse regression**) involves predicting a new value of a predictor that produced a given response value. Conditioning the simple linear regression model on $Y = y_h$ (i.e., the *given* response value), the predicted value for $X = x_h$ (i.e., the *unknown* predictor value) is

$$\hat{x}_h = \frac{y_h - b_0}{b_1},$$

where $b_1 \neq 0$. Calibration is concerned with finding an interval for this value.

We do not delve into the theory behind calibration as it is beyond the scope of this course. Basically, estimation and construction of intervals in the regression setting assume the predictor values are fixed while the response is random. However, calibration switches these roles. What results is that the statistical intervals for calibration need to be solved numerically, but various formulas for approximate intervals exist in the literature.

One approximation for a $100\times(1-\alpha)\%$ **calibration confidence interval** for the mean value of $X = x_h$ is calculated as

$$\frac{(\hat{x}_h - g\bar{x}) \pm t^*_{n-2,1-\alpha/2}\sqrt{\frac{\text{MSE}}{b_1^2}\left[\frac{1-g}{n} + \frac{(\hat{x}_h-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}}{1-g}, \tag{4.12}$$

where $t^*_{n-2;1-\alpha/2}$ is found as before and

$$g = \frac{\text{MSE} \times t^*_{n-2;1-\alpha/2}}{b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}.$$

A $100\times(1-\alpha)\%$ **calibration prediction interval** for a new value of $X = x_h$ is calculated as

$$\frac{(\hat{x}_h - g\bar{x}) \pm t^*_{n-2,1-\alpha/2}\sqrt{\frac{\text{MSE}}{b_1^2}\left[\frac{(n+1)(1-g)}{n} + \frac{(\hat{x}_h-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}}{1-g}. \tag{4.13}$$

Notice that these intervals are not centered at $\hat{x}_h$, partly because the quantity which describes the "standard error" of $\hat{x}_h$ is also a function of $\hat{x}_h$. Compare this to the traditional confidence and prediction intervals where the standard error of $\hat{y}_h$ does not depend on $\hat{y}_h$. However, estimates for calibration confidence and prediction intervals which are centered at $\hat{x}_h$ can also be used, but do not have as much theoretical development as the intervals presented above. The respective formulas for these "centered" intervals are

$$\hat{x}_h \pm t^*_{n-2;1-\alpha/2}\sqrt{\frac{\text{MSE}}{b_1^2}\left[\frac{1}{n} + \frac{(\hat{x}_h - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}. \tag{4.14}$$

and

$$\hat{x}_h \pm t^*_{n-2;1-\alpha/2}\sqrt{\frac{\text{MSE}}{b_1^2}\left[1 + \frac{1}{n} + \frac{(\hat{x}_h - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}. \tag{4.15}$$

Sometimes, you may wish to restrict your calibration to a region of y values. Thus a $100 \times (1-\alpha)\%$ calibration prediction interval for a new value of x_h corresponding to $100P\%$ of the mean of m independent y values can also be calculated. Let

$$\hat{x}_h = \frac{P\bar{y}_m - b_0}{b_1}$$

be the predicted value of $X = x_h$ based on the $100P\%$ of the mean of the m independent y values (i.e., $\bar{y}_m$). b_0 and b_1 are still the ordinary least squares estimates. The calibration prediction interval is then calculated as

$$\frac{(\hat{x}_h - g\bar{x}) \pm t^*_{n-2,1-\alpha/2}\sqrt{\frac{\text{MSE}}{b_1^2}\left[\left(\frac{(Ps_{\bar{y}})^2}{\text{MSE}} + \frac{1}{n}\right)(1-g) + \frac{(\hat{x}_h-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}}{1-g}, \tag{4.16}$$

where $s_{\bar{y}}$ is the standard error of the m independent y values.

One final note is that the method for constructing tolerance intervals can actually be inverted to provide an estimate for calibration intervals. But again, the theory for this method goes beyond the scope of this course.

4.3 Examples

Example 1: Steam Output Data (*continued*)
Suppose we wish to calculate 95% confidence intervals and 95% prediction intervals for the values of $x_{h_1} = 40$ and $x_{h_2} = 50$. Table 4.1 provides the various intervals from our earlier discussion. The upper and lower 95%/99% tolerance intervals for $x_{h_1} = 40$ and $x_{h_2} = 50$ are also given.

x_h	40	50
$\hat{y}_h$	10.2297	9.5229
95% CI	(9.7084, 10.7510)	(9.0849, 9.9610)
Bonferroni 95% CI	(9.6255, 10.8339)	(9.0151, 10.0307)
Working-Hotelling 95% CB	(9.5705, 10.8889)	(8.9689, 10.0770)
95% PI	(7.9849, 12.4745)	(7.2960, 11.7499)
Bonferroni 95% PI	(7.6277, 12.8317)	(6.9416, 12.1043)
Scheffé 95% PI	(7.3908, 13.0686)	(6.7066, 12.3393)
Upper 95%/99% TI	$(-\infty,\ 13.6215)$	$(-\infty,\ 12.8773)$
Lower 95%/99% TI	$(6.8379,\ +\infty)$	$(6.1686,\ +\infty)$
95%/99% TI	(6.5280, 13.9313)	(5.8491, 13.1968)

Table 4.1: Statistical intervals for the given levels of x. Note that $q = 2$ for the joint intervals.

Figure 4.1 gives a plot of the 95% confidence intervals and 95% prediction intervals. Notice how this plot illustrates that the prediction intervals are wider than the confidence intervals for the same confidence level. Similar plots for each of the different types of intervals can be constructed, but would be too cumbersome to include those here.

Example 2: Paper Data
A study was performed regarding the relationship between the tensile strength of paper (y) and the percentage of hardwood in the pulp (x). The data set

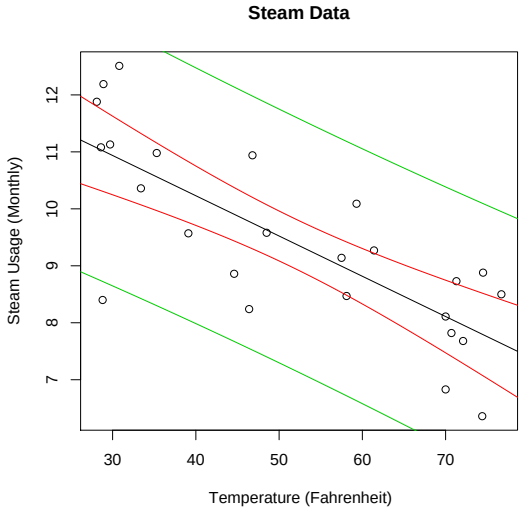


Figure 4.1: The steam output with the ordinary least squares line, 95% confidence intervals (red bands), and 95% prediction intervals (green bands).

consists of $n = 10$ samples. The data is provided in Table 4.2 and the estimated regression equation is given by

$$y_i = 143.8244 + 1.8786x_i + e_i$$

x_i	10	15	15	20	20	20	25	25	28	30
y_i	160	171	175	182	184	181	188	193	195	200

Table 4.2: The paper data set.

Suppose we want to predict the percentage of hardwood pulp for paper with a tensile strength of $y_h = 200$ along with a 95% calibration prediction interval? Let us use the formula for the calibration prediction interval that is centered at $\hat{x}_h$ since it is more computationally simple.

The expected percentage of hardwood pulp for this tensile strength is

$$\hat{x}_h = \frac{y_h - b_0}{b_1} = \frac{200 - 143.8244}{1.8786} = 29.9029.$$

For calculating the estimated standard error of this predicted value, $\sum_{i=1}^{10}(x_i - \bar{x})^2 = 357.6$ and MSE=4.8541. Then,

$$\begin{aligned}
\text{s.e.}(\hat{x}_h) &= \sqrt{\frac{\text{MSE}}{b_1^2}\left[1 + \frac{1}{n} + \frac{(\hat{x}_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} \\
&= \sqrt{\frac{4.8541}{1.8786^2}\left[1 + \frac{1}{10} + \frac{(29.9029 - 20.8)^2}{357.6}\right]} \\
&= 1.3534.
\end{aligned}$$

Thus, an estimated 95% calibration prediction interval is:

$$\begin{gathered}
29.9029 \pm (2.3060 \times 1.3534) \\
\Rightarrow 29.9029 \pm 3.1209 \\
\text{or about } 26.7820 \text{ to } 33.0238,
\end{gathered}$$

where 2.3060 is the $t^*_{8,0.975}$ multiplier. So with (approximately) 95% confidence, we would expect the percentage of hardwood pulp for a tensile strength of 200 to be between 26.7820 and 33.0238.

Chapter 5

Assessing Regression Assumptions

5.1 Theoretical Assumptions in Regression

When we fit a regression model to sample data and use the model to make statistical inferences about a larger population, we make several assumptions that may or may not be correct for the data at hand.

Theoretical assumptions that we make include the following:

1. The equation that is used for the connection between the mean of y and the x-variables describes the actual pattern of the data. For instance, when we use a straight-line equation, we assume the "average" pattern in the data is indeed linear.

2. The errors have a mean of 0, regardless of the values of the x-variables (and thus, regardless of the values of the mean y).

3. The errors have the same theoretical variance, σ^2, regardless of the values of the x-variables (and thus, regardless of the values of the mean y). For a straight-line, this means that the vertical variation of data points around the line has about the same magnitude everywhere.

4. The errors have a normal distribution.

5. We also assume that the errors are independent of each other (i.e., that they are a random sample) and are independent of any time order in

the data. A discussion assessing any time order dependence will require an introduction to time series analysis.

This chapter will deal with various regression diagnostic procedures to assess these assumptions. In the meantime, Figure 5.1 shows how all of these assumptions translate to the simple linear regression setting. The values for the dependent variable are distributed normally, with mean value falling on the regression line. Moreover, the same standard deviation holds at all values of the independent variable. Thus, the distribution curves (shown in red at three different values of the independent variable) all look the same, but they are just translated along the x-axis based on the regression relationship. Note that the only assumption that is not visualized here is the assumption of independence, which will usually be satisfied if there is no temporal component and the experimenter did a proper job of designing the experiment.

The way these assumptions are usually written is that the $\epsilon_i \sim_{iid} N(0, \sigma^2)$, such that *iid* stands for independent and identically distributed, and $N(0, \sigma^2)$ is a normal distribution with mean 0 and variance σ^2. Also, these assumptions play a role when we discuss **robustness**, which is the ability to produce estimators not heavily affected by small departures from the model assumptions.

5.2 Consequences of Invalid Assumptions

- Using the wrong equation (such as using a straight line for curved data) is a disaster. Predicted values will be wrong in a biased manner, meaning that predicted values will systematically miss the true pattern of the mean of y (as related to the x-variables).

- It is not possible to check the assumption that the overall mean of the errors is equal to 0 because the least squares process causes the residuals to sum to 0. However, if the wrong equation is used and the predicted values are biased, the sample residuals will be patterned so that they may not average 0 at specific values of x.

- The principal consequence of nonconstant variance (the phrase that means the variance is not the same for all x) are prediction intervals for individual y values will be wrong because they are determined assuming

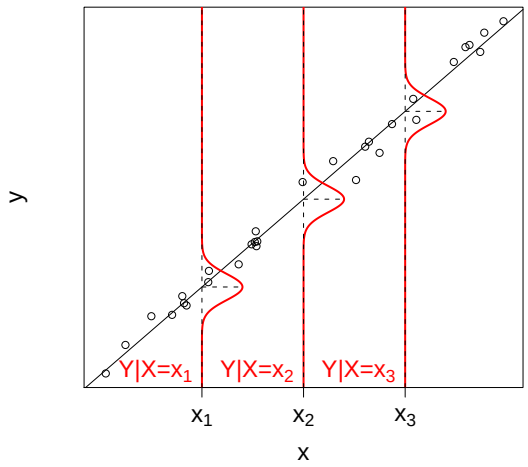


Figure 5.1: This figure shows data fit with a simple linear regression line. The distribution of the errors is shown at three different values of the predictor.

constant variance. There is a small effect on the validity of t-test and F-test results, but generally regression inferences are robust with regard to the variance issue.

- If the errors do not have a normal distribution, it usually is not particularly serious. Simulation results have shown that regression inferences tend to be robust with respect to normality (or nonnormality of the errors). In practice, the residuals may appear to be nonnormal when the wrong regression equation has been used.

5.3 Diagnosing Validity of Assumptions

Diagnosing Whether the Right Type of Equation Was Used

1. Examine a **plot of residuals versus fits** (predicted values). A curved pattern for the residuals versus fits plot indicates that the wrong type of equation has been used.

2. If the regression has repeated measurements at the same x_i values, then you can perform a formal **lack of fit test** (also called a **pure error test**) in which the null hypothesis is that the type of equation used as the regression equation is correct. Failure to reject this null hypothesis is a good thing since it means that the regression equation is okay. We will discuss this test later after we introduce analysis of variance.

Diagnosing Whether the Variance is Constant or Not

1. Examine a plot of residuals versus fits. Obvious differences in the vertical spread of the residuals indicate nonconstant variance. The most typical pattern for nonconstant variance is a plot of residuals versus fits with a pattern that resembles a sideways cone.

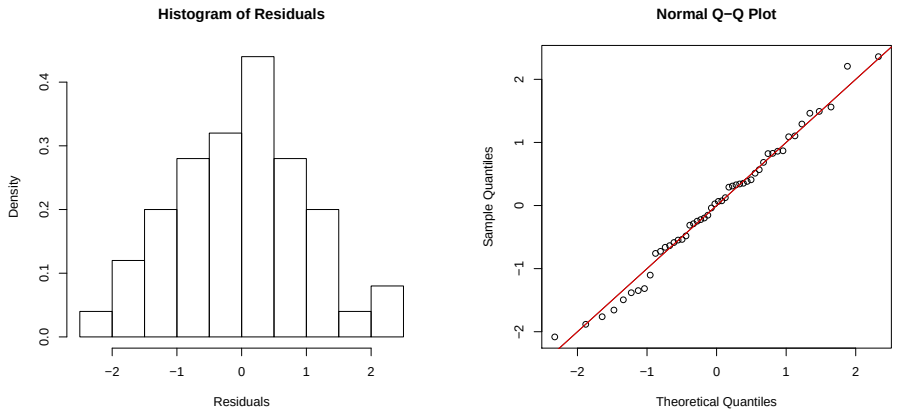
2. Do a hypothesis test to test the null hypothesis that the variance of the errors is the same for all values of the x-variable(s). There are various statistical tests that can be used, such as the modified Levene test. In practice, these tests are not used very often because non constant variance tends to be obvious from the plot of residuals versus fits plot.

Diagnosing Whether the Errors Have a Normal Distribution

1. Examine a **histogram of the residuals** to see if it appears to be bell-shaped (such as the residuals from the simulated data given in Figure 5.2(a)). The difficulty is that the shape of a histogram may be difficult to judge unless the sample size is large.

2. Examine a **normal probability plot of the residuals**. Essentially, the ordered (standardized) residuals are plotted against theoretical expected values for a sample from a standard normal curve population. A straight-line pattern for a normal probability plot (NPP) indicates that the assumption of normality is reasonable (such as the NPP given in Figure 5.2(b)).

3. Do a **hypothesis test in which the null hypothesis is that the errors have a normal distribution**. Failure to reject this null hypothesis is a good result. It means that it is reasonable to assume that the errors have a normal distribution. We discuss some testing procedures later in this chapter.



(a) (b)

Figure 5.2: (a) Histogram of the residuals from a regression fit with normal errors. (b) A normal probability plot (NPP) or quantile-quantile (QQ) plot of those errors.

Diagnosing Independence of the Error Terms

1. Usually, experiments are constructed in a way such that independence of the observations will be assumed.

2. If the observations in the data set represent successive time periods at which they were recorded and you suspect a temporal component to the study, then time series can be used for analyzing and remedying the situation. Time series analysis is a topic to be discussed later.

5.4 Plots of Residuals Versus Fits

A **predicted value** for an individual is $\hat{y}_i = b_0 + b_1 x_i$, where x_i is a specified value for the x-variable. **Fit** and **fitted value** both are synonyms for predicted value. A **residual** for an individual is $e_i = y_i - \hat{y}_i$. This is the difference between the actual and predicted values for an individual. **Error** and **deviation** both are synonyms for residual. For example, suppose that a sample regression line is $\hat{y}_i = 5 + 3x_i$ and an observation has $x_i = 4$ and $y_i = 20$. The fit is $\hat{y}_i = 5 + 3(4) = 17$ and the residual is $e_i = 20 - 17 = 3$.

We can also define the **Studentized residuals** for the j^{th} observation as:

$$r_j = \frac{e_j}{s\sqrt{(1-h_{j,j})}},$$

where $s = \sqrt{\text{MSE}}$ is the standard deviation of the residuals and

$$h_{j,j} = \frac{\sum_{i=1}^{n}(x_i - x_j)^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}.$$

The value $h_{j,j}$ is actually the diagonal entry of a matrix called the hat matrix. This value is a measure of leverage, which is a topic we will take up in the sequel to this course. We use r_j because while the variance of the true errors (ϵ_j) is σ^2, the variance of the computed residuals (e_j) is not. Using r_j now gives us residuals with constant variance. A value of $|r_j| > 3$ usually indicates the value can be considered an outlier. Other methods can be used in outlier detection, which will be discussed later.

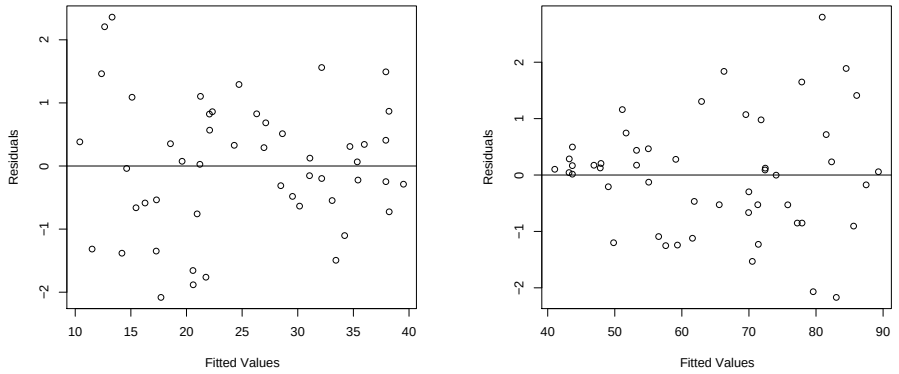
Plots of residuals versus fitted values are made by plotting all pairs of e_i and $\hat{y}_i$ for the observed sample, with residuals on the vertical axis. The idea is that properties of the residuals should be the same for all different values of the predicted values. Specifically, as we move across the plot (across predicted values), the average residual should always be about 0 and the vertical variation in residuals should maintain about the same magnitude.

5.4.1 Ideal Appearance of Plots

The usual assumptions for regression imply that the pattern of deviations (residuals) from the regression line should be similar regardless of the value of the predicted value (and value of an x-variable). The consequence is that a plot of residuals versus fits (or residuals versus an x-variable) ideally has a random "zero correlation" appearance.

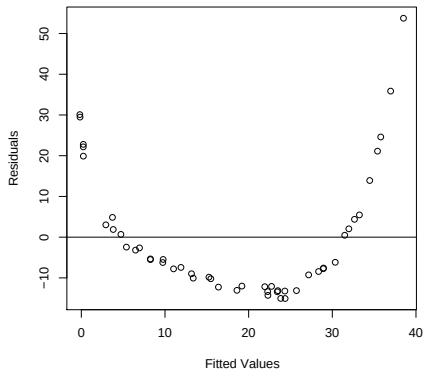
Figure 5.3(a) provides a plot of residuals versus fitted values from simulated data that more or less has the ideal appearance for this type of plot. However, the other plots of Figure 5.3 show other situations requiring further actions.

- The residuals are plotted on the vertical axis and the fitted values are on the horizontal. This is the usual convention.

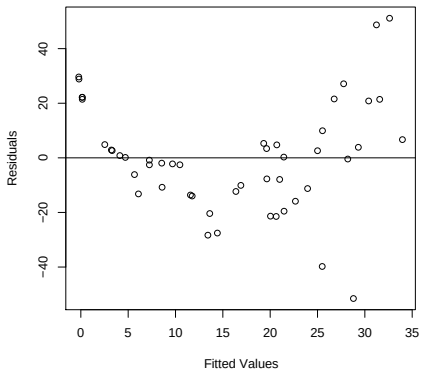


(a)

(b)



(c)



(d)

Figure 5.3: (a) Ideal scatterplot of residuals versus fitted values showing constant variance. (b) A fit demonstrating nonconstant variance. (c) A nonlinear fit. (d) A nonlinear fit with nonconstant variance.

- Notice that the residuals are averaging about 0 all the way across the plot. This is a good thing - the regression line is describing the center of the data in all locations.
- We are interested in the vertical range as we move across the horizontal axis. In Figure 5.3(a), we see roughly the same type of residual variation (vertical variation in the plot) for all predicted values. This means that deviations from the regression line are about the same as we move across the fitted values.

5.4.2 Difficulties Possibly Seen in the Plots

Three different difficulties may show up in plots of residuals versus fits:

1. Outliers in the data.
2. The regression equation for the average does not have the right form.
3. The residual variance is not constant.

Difficulty 1: Outliers in the y Values
An unusual value for the y-variable often will lead to a large residual. Thus an outlier may show up as an extreme point in the vertical direction of the residuals versus fitted values plot. Figure 5.4 gives a plot of data simulated from a simple linear regression model. However, the point at roughly (6, 35) is an outlier, especially since the residual at this point is so large. This is also confirmed by observing the plot of Studentized residuals versus fitted values (not shown here).

Difficulty 2: Wrong Mathematical Form of the Regression Equation (Curved Shape in Residual Plot)
A curved appearance in a plot of residuals versus fits indicates that we used a regression equation that does not match the curvature of the data. Figures 5.3(c) and 5.3(d) show the case of nonlinearity in the residual plots. When this happens, there is often a pattern to the data similar to that of a trigonometric function.

Difficulty 3: Nonconstant Residual Variation
Most often, although not always, non constant residual variance leads to a

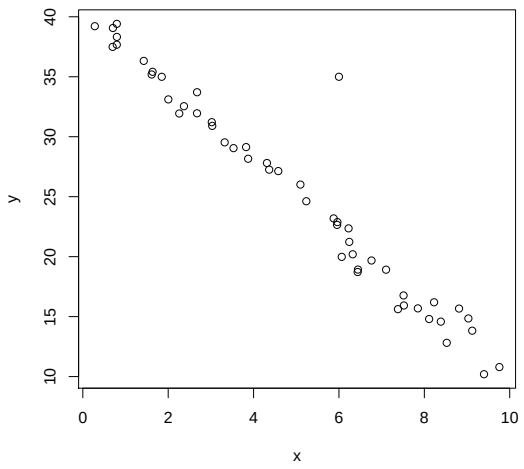


Figure 5.4: A data set with an outlier.

sideways cone or funnel shape for the plot of residuals versus fitted values. Figures 5.3(b) and 5.3(d) show plots of such residuals versus fitted values. Nonconstant variance is noted by how the data "fans out" as the predicted value increases. In other words, the residual variance (the vertical range in the plot) is increasing as the size of the predicted value increases. This actually will not hurt our regression estimates much, but is an interesting feature of the data and is also something that we should take into account when reporting the accuracy of the predictions.

5.5 Data Transformations

Transformations of the variables are used in regression to describe curvature and sometimes are also used to adjust for nonconstant variance in the errors (and y-variable).

What to Try?
When there is curvature in the data, there might possibly be some theory in the literature of the subject matter to suggests an appropriate equation.

Or, you might have to use trial and error data exploration to determine a model that fits the data. In the trial and error approach, you might try polynomial models or transformations of the x-variable(s) and/or y-variable such as square root, logarithmic, or reciprocal. One of these will often end up working out.

Transform x or Transform y?
In the data exploration approach, remember the following: If you transform the y-variable, you will change the variance of the y-variable and the errors. You may wish to try transformations of the y-variable (e.g., $\ln(y)$, $\sqrt{y}$, y^{-1}) when there is nonconstant variance and possible curvature to the data. Try transformations of the x-variable(s) (e.g., x^{-1}, x^2, x^3) when the data are curved, but the variance looks to be constant in the original scale of the y-variable. Sometimes it will be just as much art as it is science!

Why Might Logarithms Work?
If you know about the algebra of logarithm, you can verify the relationships in this section. If you don't know about the algebra of logarithms, take a leap of faith by accepting what is here as truth.

Logarithms often are used because they are connected to common exponential growth and power curve relationships. The **exponential growth equation** for variables y and x may be written as

$$y = a \times 10^{bx},$$

where a and b are parameters to be estimated. Taking logarithms on both sides of the exponential growth equation gives

$$\log_{10} y = \log_{10} a + bx.$$

Thus, an equivalent way to express exponential growth is that the logarithm of y is a straight-line function of x.

A general **power curve** equation is

$$y = a \times x^b,$$

where again a and b are parameters to be estimated. Taking logarithms on

both sides of the exponential growth equation gives[1]

$$\log y = \log a + b \log x.$$

Thus an equivalent way to write a power curve equation is that the logarithm of y is a straight-line function of the logarithm of x. This regression equation is sometimes referred to as a **log-log regression equation**.

Box-Cox Transformation
It is often difficult to determine which transformation on Y to use. **Box-Cox transformations** are a family of power transformations on Y such that $Y' = Y^\lambda$, where λ is a parameter to be determined using the data. The normal error regression model with a Box-Cox transformation is

$$Y_i^\lambda = \beta_0 + \beta_1 X_i + \epsilon_i.$$

The method of maximum likelihood (which we discuss later) can be used to estimate λ or a simple search over a range of candidate values may be performed (e.g., $\lambda = -4.0, -3.5, -3.0, \ldots, 3.0, 3.5, 4.0$). For each λ value, the Y_i^λ observations are standardized so that the analysis using the SSEs does not depend on λ. The standardization is

$$W_i = \begin{cases} K_1(Y_i^\lambda - 1), & \lambda \neq 0; \\ K_2(\log Y_i), & \lambda = 0, \end{cases}$$

where $K_2 = \prod_{i=1}^n Y_i^{1/n}$ and $K_1 = \frac{1}{\lambda} K_2^{\lambda - 1}$. Once the W_i have been calculated for a given λ, then they are regressed on the X_i and the SSE is retained. Then the maximum likelihood estimate $\hat{\lambda}$ is that value of λ for which the SSE is a minimum.

5.6 Tests for Normality

As mentioned earlier, a hypothesis test can be performed in which the null hypothesis is that the errors have a normal distribution. Failure to reject this null hypothesis is a good result. It means that it is reasonable to assume that the errors have a normal distribution. Typically, assessment of

[1]Note that here log refers to the natural logarithm. In Statistics, log and ln are used interchangeably. If any other base is ever used, then the appropriate subscript will be used (e.g., $\log_{10}$).

the appropriate residual plots is sufficient to diagnose deviations from normality. However, more rigorous and formal quantification of normality may be requested. So this section provides a discussion os some common testing procedures (of which there are many) for normality. For each test discussed below, the formal hypothesis test is written as:

$$H_0 : \text{the errors follow a normal distribution}$$
$$H_A : \text{the errors do not follow a normal distribution.}$$

Anderson-Darling Test

The **Anderson-Darling Test** measures the area between the fitted line (based on chosen distribution) and the nonparametric step function (based on the plot points). The statistic is a squared distance that is weighted more heavily in the tails of the distribution. Smaller Anderson-Darling values indicates that the distribution fits the data better. The test statistic is given by:

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n}[\log \mathrm{F}(e_i) + \log(1 - \mathrm{F}(e_{n+1-i}))],$$

where $\mathrm{F}(\cdot)$ is the cumulative distribution of the normal distribution. The test statistic is compared against the critical values from a normal distribution in order to determine the p-value.

Kolmogorov-Smirnov Test

The **Kolmogorov-Smirnov Test** compares the empirical cumulative distribution function of your sample data (which we discuss in more detail later) with the distribution expected if the data were normal. If this observed difference is sufficiently large, the test will reject the null hypothesis of population normality. The test statistic is given by:

$$D = \max(D^+, D^-),$$

where

$$D^+ = \max_i(i/n - \mathrm{F}(e_{(i)}))$$
$$D^- = \max_i(\mathrm{F}(e_{(i)}) - (i-1)/n),$$

where $e_{(i)}$ pertains to the i^{th} largest value of the error terms. The test statistic is compared against the critical values from a normal distribution in order to determine the p-value.

Shapiro-Wilk Test

The **Shapiro-Wilk Test** uses the test statistic

$$W = \frac{\left(\sum_{i=1}^{n} a_i e_{(i)}\right)^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2},$$

where the a_i values are calculated using the means, variances, and covariances of the $e_{(i)}$ (the form of which requires more of technical discussion and is beyond the scope of what we have discussed thus far). W is compared against tabulated values of this statistic's distribution. Small values of W will lead to rejection of the null hypothesis.

Ryan-Joiner Test

The **Ryan-Joiner Test** is sort of a simpler alternative to the Shapiro-Wilk test. The test statistic is actually a correlation coefficient calculated by

$$R_p = \frac{\sum_{i=1}^{n} e_{(i)} z_{(i)}}{\sqrt{s^2(n-1)\sum_{i=1}^{n} z_{(i)}^2}},$$

where the $z_{(i)}$ values are the z-score values (i.e., normal values) of the corresponding $e_{(i)}$ value and s^2 is the sample variance. Values of R_p closer to 1 indicate that the errors are normally distributed.

5.6.1 Skewness and Kurtosis

Suppose we have the residuals $e_1, \ldots, e_n$ from our simple linear regression fit. We have already introduced the notion of the mean and variance of the residuals. More broadly, for a univariate data set, the mean and variance of the data are based on the first and second moments of the data, respectively. In particular, the **$\mathbf{r^{th}}$ (sample) moment** of the residuals is given by

$$m_r' = \frac{\sum_{i=1}^{n} e_i^r}{n}.$$

Moreover, the **r^{th} central moment** of the residuals is given by

$$m_r = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^r}{n},$$

where $\bar{e}$ is the mean of the residuals, which will be 0 when using an ordinary least squares fit to the data. In other words, the r^{th} moment and r^{th} central moment are the same for our setting. Also note that the mean and variance are given by m_1 and $\frac{(n-1)m_2}{n}$, respectively.

Two other moment estimators can be used to assess the degree of non-normality - both of which can be estimated in a variety of ways. We just present one formulation for each of these estimators, but it should be noted that there is "confusing" references to each of these estimators in the literature. The first one we discuss is **skewness**, which is a statistic that measures the direction and degree of asymmetry of your data. A value of 0 indicates a symmetrical distribution. A positive value indicates skewness (i.e., long-tailedness) to the right, while a negative value indicates skewness to the left. Values between -3 and +3 are typical values of samples from a normal distribution. The formulation we present for the sample skewness is

$$g = \frac{m_3}{m_2^{3/2}}.$$

Note that this particular estimate of the skewness, like the others found in the literature, has a degree of bias to it when estimating the "true" population skewness.

The other moment estimator is **kurtosis**, which is a statistic that measures the heaviness of the tails of a distribution. The usual reference point in kurtosis is the normal distribution. If this kurtosis statistics equals three and the skewness is zero, then the distribution is normal. Unimodal distributions that have kurtosis greater than 3 have heavier or thicker tails than the normal. These same distributions also tend to have higher peaks in the center of the distribution (*leptokurtic*). Unimodal distributions whose tails are lighter than the normal distribution tend to have a kurtosis that is less than 3. In this case, the peak of the distribution tends to be broader than the normal (*platykurtic*). The formulation we present for the sample kurtosis is

$$k = \frac{m_4}{m_2^2}.$$

Be forewarned that this statistic is an unreliable estimator of kurtosis for small sample sizes and there are other formulations in the literature as with skewness.

There are some formal tests of these statistics with how they translate to the normal distribution. However, simply using the rules of thumb above with respect to the raw residuals is generally acceptable. While these statistics are not commonly reported with assessing the normality of the residuals, they can give further justification or clarity in a setting where it appears that the normality assumption is violated.

5.7 Tests for Constant Error Variance

There are various tests that may be performed on the residuals for testing if they have constant variance. The following should not be used solely in the determination if the residuals do indeed have constant variance. In fact, it is usually sufficient to "visually" interpret a residuals versus fitted values plot. However, the tests we discuss can provide an added layer of justification to your analysis. Also, some of the following procedures require you to partition the residuals into a certain number of groups, say $k \geq 2$ groups of sizes $n_1, \ldots, n_k$ such that $\sum_{i=1}^{k} n_i = n$. For these procedures, the sample variance of group i is given by:

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(e_{i,j} - \bar{e}_{i,\cdot})^2}{n_i - 1},$$

where $e_{i,j}$ is the j^{th} residual from group i. Moreover, the pooled variance is given by:

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n - k}.$$

F-Test

Suppose we partition the residuals of observations into two groups - one consisting of the residuals associated with the lowest predictor values and those belonging to the highest predictor values. Treating these two groups as if they could (potentially) represent two different populations, we can test

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_A : \sigma_1^2 \neq \sigma_2^2$$

using the F-statistic $F^* = s_1^2/s_2^2$. This test statistic is distributed according to a F_{n_1-1,n_2-1} distribution, so if $F^* \geq F_{n_1-1,n_2-1;1-\alpha}$, then reject the null hypothesis and conclude that there is statistically significant evidence that the variance is not constant.

Modified Levene Test

Another test for nonconstant variance is to use the **modified Levene test** (sometimes called the **Brown-Forsythe test**). This test does not require the error terms to be drawn from a normal distribution and hence it is a nonparametric test. The test is constructed by grouping the residuals into k groups according to the values of X. It is typically recommended that each group has at least 25 observations. Usually two groups are used, but we develop this test from the more general perspective of k groups.

Begin by letting group 1 consist of the residuals associated with the n_1 lowest values of the predictor. Then, let group 2 consists of the residuals associated with the n_2 lowest remaining values of the predictor. Continue on in this manner until you have partitioned the residuals into k groups. The objective is to perform the following hypothesis test:

$$H_0 : \text{the variance is constant}$$
$$H_A : \text{the variance is not constant.}$$

While hypothesis tests are usually constructed to *reject* the null hypothesis, this is a case where we actually hope we *fail to reject* the null hypothesis. The test statistic for the above is computed as follows:

$$L^2 = \frac{(n-k)\sum_{i=1}^{k} n_i(\bar{d}_{i,\cdot} - \bar{d}_{\cdot,\cdot})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(d_{i,j} - \bar{d}_{i,\cdot})^2},$$

where $d_{i,j}$ is typically one of the following quantities:

- $d_{i,j} = |e_{i,j} - \bar{e}_{i,\cdot}|$, where $\bar{e}_{i,\cdot}$ is the mean of the i^{th} group of residuals (this is how the *original* Levene test was defined).
- $d_{i,j} = |e_{i,j} - \tilde{e}_{i,\cdot}|$, where $\tilde{e}_{i,\cdot}$ is the median of the i^{th} group of residuals.
- $d_{i,j} = |e_{i,j} - \check{e}_{i,\cdot;\gamma}|$, where $\check{e}_{i,\cdot}$ is the $100 \times \gamma\%$ trimmed mean of the i^{th} group of residuals. The $100\times\gamma\%$ **trimmed mean** removes the $100\times\gamma\%$ smallest and $100 \times \gamma\%$ largest values (rounded to the nearest integer) and then calculates the mean of the remaining values. Typically, for the modified Levene test, $\gamma = 0.10$.

In each case, L is approximately distributed according to a t_{n-k} distribution (i.e., L^2 is approximately distributed according to a a $F_{1,n-k}$ distribution).

So why are there multiple ways to define $d_{i,j}$? The choice of how you define $d_{i,j}$ determines the robustness and power of the Levene test. By robustness, we mean the ability of the test to *not* falsely detect unequal variances when the underlying distribution of the residuals is normal and the variance is constant. By power, we mean the ability of the test to detect that the variance is not constant when it is, in fact, not constant. Although the optimal choice inevitably depends on the underlying distribution, the definition based on the median of the residuals is typically recommended as it provides robustness against many non-normal distributions while retaining high power.

As an example of one of the formulations above, consider using the medians of the residual groups when $k = 2$. Then, let

$$s_L = \sqrt{\frac{\sum_{j=1}^{n_1}(d_{1,j} - \bar{d}_1)^2 + \sum_{j=1}^{n_2}(d_{2,j} - \bar{d}_2)^2}{n_1 + n_2 - 2}}.$$

The test statistic is computed as

$$L = \frac{\bar{d}_1 - \bar{d}_2}{s_L\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

such that L is approximately distributed according to a $t_{n_1+n_2-2}$ distribution.

The modified Levene test is usually the preferred test used in textbooks and statistical packages regarding variances. Since the Levene test does not make any distributional assumptions, the median will provide a more informative measure of the center of the data (compared to the mean or trimmed mean) if the distribution of our data is skewed. But if the data is (approximately) symmetric, then all versions of the Levene test will provide similar results.

One additional note is that $d_{i,j}$ can also be formulated by the squared versions of the quantities, namely:

- $d_{i,j} = (e_{i,j} - \bar{e}_{i,\cdot})^2$;
- $d_{i,j} = (e_{i,j} - \tilde{e}_{i,\cdot})^2$;
- $d_{i,j} = (e_{i,j} - \check{e}_{i,\cdot;\gamma})^2$.

The testing is still done similarly, it is just that this version utilizes an L_2-norm, which results in larger residuals having a (potentially) bigger effect on the calculated $d_{i,j}$. The earlier definitions of $d_{i,j}$ using the absolute values (i.e., L_1-norm) are typically the quantities used.

Bartlett's Test

Bartlett's test is another test that can be used to test for constant variance. The objective is to again perform the following hypothesis test:

$$H_0 : \text{the variance is constant}$$
$$H_A : \text{the variance is not constant.}$$

Bartlett's test is highly sensitive to the normality assumption, so if the residuals do not appear normal (even after transformations), then this test should not be used. Instead, the Levene test is the alternative to Bartlett's test that is less sensitive to departures from normality.

This test is carried out similarly to the Levene test. Once you have partitioned the residuals into k groups, the following test statistic can be constructed:

$$B = \frac{(n-k)\ln s_p^2 - \sum_{i=1}^{k}(n_i-1)\ln s_i^2}{1+\left[\frac{1}{3(k-1)}\left(\left(\sum_{i=1}^{k}\frac{1}{n_i-1}\right)-\frac{1}{n-k}\right)\right]}.$$

The test statistic B is distributed according to a χ^2_{k-1} distribution, so if $B \geq \chi^2_{k-1;1-\alpha}$, then reject the null hypothesis and conclude that there is statistically significant evidence that the variance is not constant. The p-value would correspond to the unshaded region shown in Figure E.3.

Breusch-Pagan Test

The **Breusch-Pagan test** is an alternative to the modified Levene test. Whereas the modified Levene test is a nonparametric test, the Breusch-Pagan test assumes that the error terms are normally distributed, with $\mathrm{E}(\epsilon_i) = 0$and $\mathrm{Var}(\epsilon_i) = \sigma_i^2$ (i.e., nonconstant variance). The σ_i^2 values depend on the X_i values in the following way:

$$\log \sigma_i^2 = \gamma_0 + \gamma_1 X_i.$$

We are interested in testing the null hypothesis of constant variance versus the alternative hypothesis of nonconstant variance. Specifically, the hypoth-

esis test is formulated as:

$$H_0 : \gamma_1 = 0$$
$$H_A : \gamma_1 \neq 0.$$

This test is carried out by first regressing the residuals on the predictor (i.e., regressing e_i on x_i). The sum of squares resulting from this analysis is denoted by SSR*, which provides a measure of the dependency of the error term on the predictor. The test statistic is given by

$$X^{2*} = \frac{\text{SSR}^*/2}{(\text{SSE}/n)^2},$$

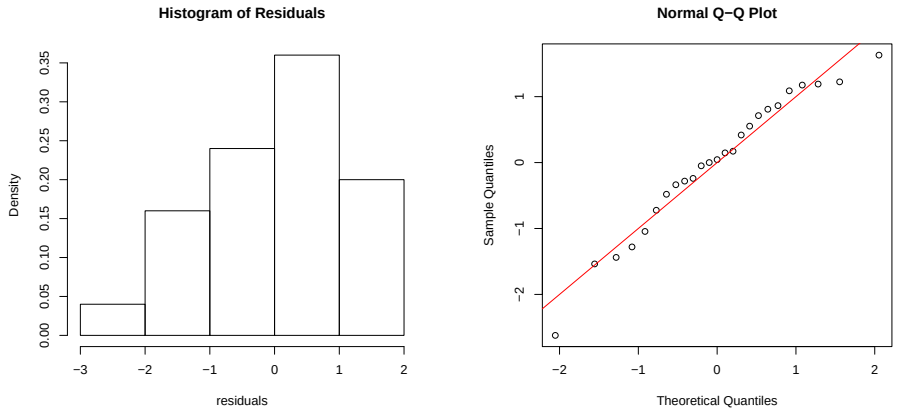
where SSE is from the regression analysis of the response on the predictor. The p-value for this test is found using a χ^2 distribution with 1 degree of freedom (written as χ^2_1). The p-value would correspond to the unshaded region shown in Figure E.3.

5.8 Examples

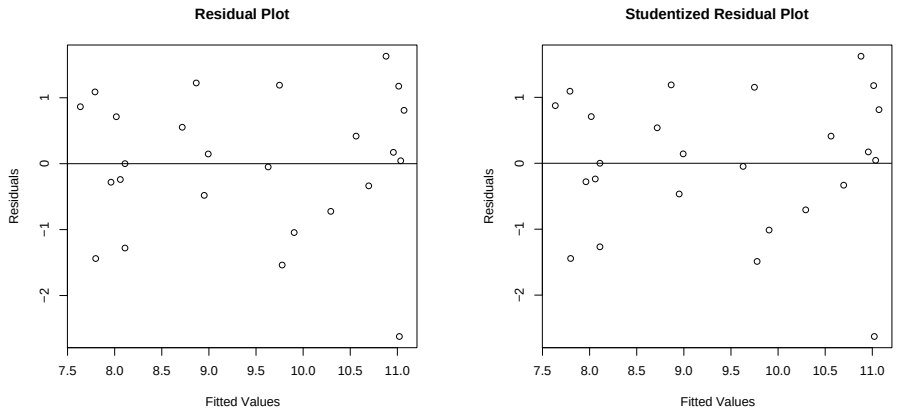
There is one thing to keep in mind when assessing regression assumptions - it is just as much an art as it is a science. When dealing with real data sets, it is somewhat subjective as to what you are willing to claim as satisfying a certain assumption. If you look at data sets in textbooks, often times they are chosen because they provide very nice plots or can clearly demonstrate some concept (e.g., outliers). However, real data sets will require you to study them a little closer. The important thing is to be able to justify your decision based on the statistical tools you have at your disposal.

Example 1: Steam Output Data (*continued*)
We have already performed statistical tests on this data set to assess the significance of the slope and intercept terms as well as construction of various statistical intervals. However, all of these procedures are technically invalid if we do not demonstrate that the underlying assumptions are met.

Figure 5.5(a) gives a histogram of the residuals. While the histogram is not completely symmetrical and bell-shaped, chances are you will not see such a histogram unless you have a very large data set. The important thing is to not see very erratically shaped histograms.



(a) (b)



(c) (d)

Figure 5.5: (a) Histogram of the residuals for the steam data. (b) NPP of the residuals. (c) Plot of the residuals versus the fitted values. (d) Plot of the Studentized residuals versus the fitted values.

Figure 5.5(b) provides a normal probability plot which seems to indicate the residuals lie on a straight line. Thus the assumption of normality appears to be met. Using the Kolmogorov-Smirnov test (with using MSE = 1.1140 as an estimate of the variance) yields $D = 0.0951$ with a p-value of 0.9616. Thus we fail to reject the null hypothesis of normality.

Finally, Figures 5.5(c) and 5.5(d) show the residual plots which show random scatter about 0. In fact, there is not much difference between the two plots (which is often the case). So the assumption of constant variance appears to be valid.

Example 2: Catapult Data
This data set of size $n = 9$ are from a data set obtained during a class demonstration of launching a ball from different angles using a catapult. The variables are x = start angle and y = distance. Table 5.1 gives the data used for this analysis. A scatterplot of the data with the least squares line is provided in Figure 5.6(a).

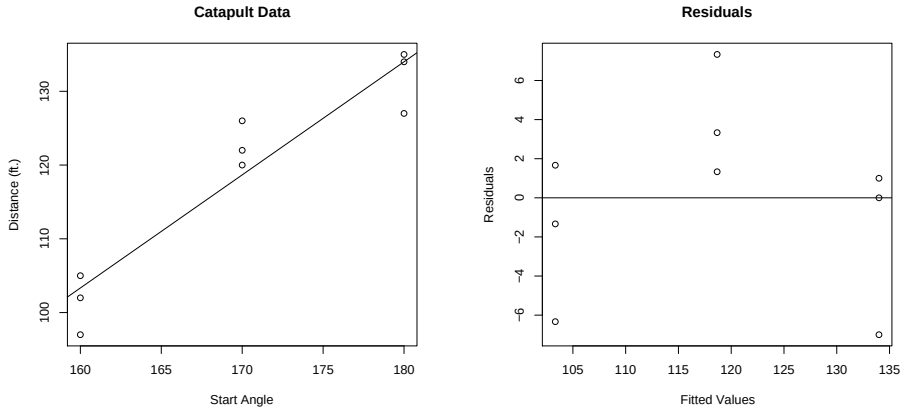
i	1	2	3	4	5	6	7	8	9
y_i	102	97	105	126	122	120	134	135	127
x_i	160	160	160	170	170	170	180	180	180

Table 5.1: The catapult data. i corresponds to the observation number.

Notice in Figure 5.6(b) that the shape of the residuals versus the fitted values do not seem to indicate constant variance, but we do have a small data set so this may be an artifact of the data. Using the modified Levene test (with the 5 smallest values comprising the first group), the output is

```
##########
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  1   0.405 0.5447
       7
##########
```

From this output, we see that $F^*_{1,7} = 0.405$ with a p-value of 0.5447, so there is not statistically significant evidence to reject the hypothesis of constant variance.



(a) (b)

Figure 5.6: (a) Catapult data with ordinary least squares line. (b) Plot of the residuals.

Chapter 6

ANOVA I

6.1 Constructing the ANOVA Table

An **analysis of variance** (**ANOVA**) table for regression displays quantities that measure how much of the variability in the y-variable is explained and how much is not explained by the regression relationship with the $x$variable(s). The table also gives the construction and value of the mean squared error (MSE) and a significance test of whether the variables are related in the population represented by the sample.

An underlying conceptual idea for the construction of the ANOVA table is:

$$\text{overall variation in } y \;=\; \text{regression variation} \;+\; \text{error variation.}$$

More formally, we can partition our linear regression equation to reflect the above concept:

$$\begin{aligned} y_i &= b_0 + b_1 x_i + e_i \\ &= \hat{y}_i + e_i \\ &= \hat{y}_i + (y_i - \hat{y}_i) \\ \Rightarrow (y_i - \bar{y}) &= (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i). \end{aligned}$$

The above relationship is also illustrated in Figure 6.1, where the different color lines represent the different sources of error.

Some quantities in the ANOVA table include the following:

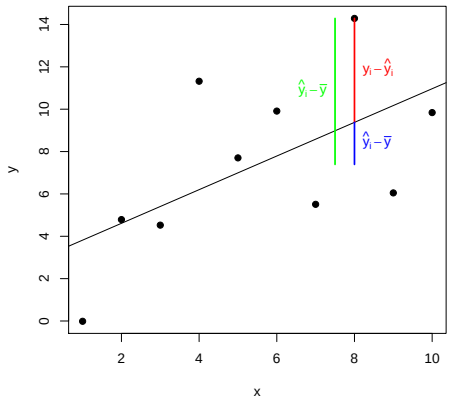


Figure 6.1: How the regression equation can be partitioned into it's various error sources. These quantities are the basis of an ANOVA table, which is the platform for our statistical testing and variance component analysis.

- The **sum of squares for total** is $\text{SSTO} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, which is the sum of squared deviations from the overall mean of y. The value $n-1$ is called the **total degrees of freedom** ($\mathbf{df}_T$). SSTO is a measure of the overall variation in the y-values.
- The **sum of squared errors** is $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, which is the sum of squared observed errors (residuals) for the observed data. SSE is a measure of the variation in y that is not explained by the regression. For simple regression, the value $n-2$ is called the **error degrees of freedom** ($\mathbf{df}_E$).
- The **mean squared error** is $\text{MSE} = \frac{\text{SSE}}{\text{df}_E} = \frac{\text{SSE}}{n-2}$, which estimates σ^2, the variance of the errors.
- The **sum of squares due to the regression** is $\text{SSR} = \text{SSTO} - \text{SSE}$, and it is a measure of the total variation in y that can be explained by the regression with the x-variable. Also, for the **regression degrees of freedom** ($\mathbf{df}_R$), we have $\text{df}_R = \text{df}_T - \text{df}_E$. For simple regression, $\text{df}_R = (n-1) - (n-2) = 1$.
- The **mean square for the regression** is $\text{MSR} = \frac{\text{SSR}}{\text{df}_R} = \frac{\text{SSR}}{1}$.

Source	df	SS	MS	F
Regression	1	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MSR	MSR/MSE
Error	$n-2$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE	
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$		

Table 6.1: ANOVA table for simple linear regression.

The uses of the ANOVA table in Table 6.1 include:

- The error line tells the value for the estimated error variance, MSE.
- The "Total" and "Error" lines give the SS values used in the calculation of R^2.
- The F-statistic can be used to test whether the y-variable and x-variable are related.

Specifically, in the simple linear regression model, $F^* = \text{MSR/MSE}$ is a test statistic for:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

and the corresponding degrees of freedom are 1 and $n-2$. Statistical software will report a p-value for this test statistic. The same decision rules apply here as when we discussed statistical inference.

As a special case of the ANOVA table in Table 6.1, consider the regression through the origin setting. When performing a regression through the origin, we force $\bar{y} = 0$ to get the uncorrected sums of squares totals in Table 6.2. By forcing $\bar{y} = 0$, the sums of squares can now be shown to be additive. Table 6.2 gives the ANOVA table for the regression through the origin. Notice that (generally) $\sum_{i=1}^{n} e_i \neq 0$. Because of this, the SSE could actually be larger than the SSTO in Table 6.1, hence why we use the uncorrected values in Table 6.2. Note that by forcing $\bar{y} = 0$, we are not actually asserting that the mean value of our response is 0. It is just a way to get additivity in the ANOVA table.

Source	df	SS	MS	F
Regression	1	$\sum_{i=1}^{n} \hat{y}_i^2$	MSR	MSR/MSE
Error	$n-1$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE	
Total	n	$\sum_{i=1}^{n} y_i^2$		

Table 6.2: ANOVA table for regression through the origin.

6.2 Formal Lack of Fit

It was just shown that every regression divides up the variation into two segments:

$$\text{SSTO} = \text{SSR} + \text{SSE}.$$

If the regression also has repeated measurements with the same x values (called **replicates**), then you can estimate a lack of fit term to detect non-linearities in the data. A lack of fit test cannot be used with data which has a continuous spread of x values with no replicates, such as what might come from an existing process data set. The formal test for lack of fit is written as:

$$H_0 : \text{there is no linear lack of fit}$$
$$H_A : \text{there is a linear lack of fit.}$$

To perform a lack of fit test, first partition the SSE into two quantities: lack of fit and pure error. The **pure error sum of squares** (**SSPE**) is found by considering only those observations that are replicates. The x values are treated as the levels of the factor in a one-way ANOVA. With x held constant, the SSE from this analysis measures the underlying variation in y. Thus, it is called pure error. Subtracting the pure error sum of squares from the SSE of the linear regression measures the amount of nonlinearity in the data, called the **lack of fit sum of squares** (**SSLOF**).

Formally, let $j = 1, \ldots, m < n$ be the index for the unique levels of the predictors and $i = 1, \ldots, n_j$ be the index for the i^{th} replicate of the j^{th} level. (Notice that $\sum_{j=1}^{m} n_j = n$.) We assume that the random variable $Y_{i,j}$ can be written as

$$Y_{i,j} = \mu_j + \epsilon_{i,j},$$

such that the μ_j are the mean responses at the m unique levels of the predictor and $\epsilon_{i,j}$ are again assumed to be *iid* normal with mean 0 and variance σ^2.

Then,

$$\text{SSE} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j} - \hat{y}_j)^2,$$

where $y_{i,j}$ is the i^{th} observed response at the j^{th} predictor level and $\hat{y}_j$ is the expected response at the j^{th} predictor level. Then,

$$\begin{aligned}\text{SSE} &= \text{SSPE} + \text{SSLOF} \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j} - \bar{y}_j)^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(\bar{y}_j - \hat{y}_{i,j})^2\end{aligned}$$

An F-statistic can be constructed from these two values that will test the statistical significance of the lack of fit

$$F^*_{\text{df}_{LOF},\text{df}_{PE}} = \frac{\frac{\text{SSLOF}}{\text{df}_{LOF}}}{\frac{\text{SSPE}}{\text{df}_{PE}}} = \frac{\text{MSLOF}}{\text{MSPE}},$$

where $\text{df}_{LOF} = m - 2$ and $\text{df}_{PE} = n - m$ are the degrees of freedom for the SSLOF and SSPE, respectively. The ANOVA table including lack of fit and pure error is given in Table 6.3. The actual formulas for the sum of squares have been omitted to simplify the table.

Source	**df**	**SS**	**MS**	**F**
Regression	1	SSR	MSR	MSR/MSE
Error	$n-2$	SSE	MSE	
Lack of Fit	$m-2$	SSLOF	MSLOF	MSLOF/MSPE
Pure Error	$n-m$	SSPE	MSPE	
Total	$n-1$	SSTO		

Table 6.3: ANOVA table for simple linear regression which includes a lack of fit test.

6.3 Examples

Example 1: Steam Output Data (*continued*)
The following is the ANOVA table for the steam output data set:

```
##########
Analysis of Variance Table

Response: steam
            Df Sum Sq Mean Sq F value    Pr(>F)
temperature  1 38.193  38.193  34.284 5.745e-06 ***
Residuals   23 25.623   1.114
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##########
```

Some things to note are:

- The MSE is 1.114 and this estimates the variance of the errors.
- $R^2 = \frac{\text{SSTO-SSE}}{\text{SSTO}} = \frac{\text{SSR}}{\text{SSTO}} = \frac{38.193}{63.816} = 0.5985$ or 59.85%. The interpretation is that atmospheric temperature explains 59.85% of the observed variation in steam output.
- The value of the F-statistic is $F^* = 38.193/1.114 = 34.285$ (different due to rounding error), and the p-value for this F-statistic is $p = 5.75 \times 10^{-6}$. Thus, we reject the null hypothesis $H_0 : \beta_1 = 0$ because the p-value is so small. In other words, the observed relationship is statistically significant.

Example 2: Catapult Data (*continued*)
For the catapult data set, we are interested in testing whether or not there is a linear lack of fit. The ANOVA table for this data set is given below:

```
##########
Analysis of Variance Table

Response: distance
          Df  Sum Sq Mean Sq F value    Pr(>F)
angle      1 1410.67 1410.67  61.207 0.0001051 ***
Residuals  7  161.33   23.05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##########
```