Notes on regression analysis for social research

E. C. Hedberg
hedbergec@gmail.com

January 30, 2026

Contents

List of Tables

List of Figures

About

This document is a collection of my notes for an advanced regression course. It comes with no assurance of absolute accuracy. These notes are updated constantly and dated.

NOTE: These notes are based on the work of many statisticians and social scientists, and in no way reflect my own intellectual contributions.

I will eventually build a larger bibliography, but for right now these are just notes for a class. These notes are not designed to replace a proper text.

I assume you have had a basic regression course and, of course, introduction to statistics. This is not about theory, and I go through the basics quickly.

Chapter 1

Univariate statistics

1.1 The mean, statistics, and parameters

To calculate the mean, simply add up the numbers for each of the cases and divide by the total number of cases. For example, the mean of the numbers 1,2,3,4,5,6,7,8,9 is (1+2+3+4+5+6+7+8+9) / 9 = 41/9 = 4.556. Statisticians and mathematicians symbolize the addition of values in a set by using the summation symbol, Σ. The $\sum_{i=1}^{N}$ phrase represents the summation of each i^{th} case, starting with the first one $i = 1$ and ending with the N^{th} case. To calculate the mean, this sum is divided by N, or the total number of cases.

The formula of the mean is:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}, \tag{1.1}$$

which is estimated by the statistic:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}. \tag{1.2}$$

In the previous paragraph, the term statistic is used. It is important to note that there are two concepts: parameters and statistics, which are often conflated by students. Frequentists generally think that the population at large can be described with parameters. For example, one parameter is the mean, or μ. For example, a statistician might say that the population of objects (e.g., people) has some mean income. The statistician could then use a sample of objects to estimate this parameter. This estimate is the statistic.

For example, the estimate of the population mean, μ, of some variable x is the mean statistic $\bar{x}$. If the variable was w, the statistic would be $\bar{w}$. The estimate of the population variance σ^2 is the statistic s^2. One way to think about this is with the mnemonic **PPSS**: Populations have Parameters, Samples have Statistics.

Of course, these statistics are not going to be absolutely correct, and will vary from one sample to the next. Thus, one of the things that researchers fret most about is how well a certain statistic estimates a parameter. This is the sampling variation of that statistic, or the variance of the estimate. The square root of the variance of a statistic is the standard error. More on that later.

1.2 Deviance

The next important idea is that of deviance. This is central to most of what statistics is about since statisticians often wonder how one population differs from another that has experienced some effect or treatment. The most basic deviation is the difference between a case's value of x and the mean of x, $\bar{x}$. Deviance can be symbolized as:

$$e_i = x_i - \bar{x}. \tag{1.3}$$

Even this basic idea has substantive interest. It is often important to know how some case differs from what is typical. Here, the mean is considered typical, and it is valuable to know whether a case is typical, above typical, or below typical. Negative deviance values are below the mean, positive deviance values are above the mean, and 0 deviance is exactly the mean.

1.3 The sum of squares

It is often important to understand how much deviance there is in our sample. One method might be to add up each deviance score, but because of properties of the mean this will wind up at 0 which makes this a rather useless method. As luck would have it, there is a simple solution to this problem.

By squaring each deviance, the negative values of deviance are effectively made positive (since the square of any negative number is positive) and then

by adding up each squared deviance term, a measurement of total deviance is obtained. We call this total deviance the sum of squares, or SS:

$$SS = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2. \tag{1.4}$$

One interesting way to view SS is to consider it a measure of how much information is available in a variable. The larger the SS, the more information that is contained in the variable that can be modeled and correlated with other variables. If the SS is small, then little will be correlated with it.

1.4 The variance

The next quantity is the variance. This is simply the average of SS. In the population, this is

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}. \tag{1.5}$$

The estimate of the population variance is the sum of squared deviations divided by N-1:

$$s_x^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}. \tag{1.6}$$

We divide by N-1 instead of N because we lose a degree of freedom since this statistic employs another statistic (the mean). This is also a correction to remove bias, since this estimate of the population variance without this correction is slightly smaller than the expected parameter.

1.5 The standard deviation

A problem with the variance is that it in units of squared deviations. To get back to simple deviations, we take the square root of the variance to get the standard deviation. In the population this is

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}, \tag{1.7}$$

and the statistic is

$$s_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}. \tag{1.8}$$

Chapter 2

A little probability theory

At the end of the day, all analyses are attempting to make some statement(s) with two parts:

- Some assertion (or set of assertions)
- The probability that this assertion is (or these assertions are) false

Most of these notes focus on the methods to make such statements using a regression model of some sort. These regression models produce some *statistic* of interest, whether it is a coefficient, predicted value, or measure of model fit. Generally, all statistics have an associated *variance*, or measure of uncertainty that can generally be used to produce some estimated probability of observing the data on hand if we assume some (null) worldview. Thus, we should never lose sight of keeping track of the probabilities associated with our assertions. The probability associated with an assertion depends on several factors:

- The statistic being estimated, for example
  - a coefficient from a OLS regression
  - a coefficient from a GLS regression
  - a coefficient from a GLM regression
  - a coefficient from a Mixed regression model
- The degrees of freedom associated with the statistic, if applicable

- The distribution of the statistic, for example
  - z
  - t
  - F
  - χ^2
- How the data were collected, for example
  - Census or administrative data
  - Simple random sample
  - Stratified random sample
  - Cluster randomized sample
  - Stratified cluster randomized sample

Therefore, it is important to remind ourselves of the basics of probability, random variables, and samples. This will not be a very in-depth discussion of these concepts, but should put you on notice to spend some time to better understand them as you start your research career.

The study of probability as a formal science started in 17^{th} century France when a gambler started talking to his mathematician buddies, namely, Pascal (yup, that Pascal). We generally care about probability not because the events we study are random, but that the observations we have in our data set, our *sample*, generally arrived there through some random process.

Here are some basic concepts, quickly.

The universe of possible events is considered the **sample space**, which we can sometimes call Ω. An element of the sample space is sometimes called ω. For example, if we roll a dice, we have 6 possible sides than can land upright:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

From a sample space, we are generally concerned with a particular event occurring, for example, we can say that the event A occurs when we roll a dice and a number greater than or equal to 5 shows up

$$A = \{5, 6\}$$

or we can say that the even B occurs when we roll a dice and a number less than or equal to 2 shows up

$$B = \{1, 2\}$$

We can also say that the event C is the union of A and B when either A or B occur

$$C = A \cup B$$

or

$$C = \{1, 2, 5, 6\}$$

Suppose we had another even D, that is the number on the dice was even

$$D = \{2, 4, 6\}$$

We could say that the intersection of events C and D is event E

$$E = C \cap D$$

or

$$E = \{2, 6\}$$

Of course, we could always get an empty set, that is a set with nothing in it. For example if we have event F, that the number of the dice is odd, then

$$E \cap F = \emptyset$$

in which case we can say that E and F are disjoint.

Here are some important laws

- Communicative law

$$A \cup B = B \cup A \tag{2.1}$$

$$A \cap B = B \cap A \tag{2.2}$$

- Associative law

$$(A \cup B) \cup C = A \cup (B \cup C) \tag{2.3}$$

$$(A \cap B) \cap C = A \cap (B \cap C) \tag{2.4}$$

- Distributive law

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C) \tag{2.5}$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C) \tag{2.6}$$

Next, we have the concept of a **probability measure** on Ω. A probability measure is a function P that maps subsets of Ω to a real number. The rules (or axioms) are

- The probability of the whole sample space is 1

$$P(\Omega) = 1 \tag{2.7}$$

- If A a subset of the sample space, then the probability of greater than or equal to 0

$$\text{if } A \subset \Omega, \text{ then } P(A) \geq 0 \tag{2.8}$$

  and the big one that makes a lot of things work:

- If A and B are disjoint, then the probability of either happening is the sub of the probabilities of each

$$P(A \cup B) = P(A) + P(B) \tag{2.9}$$

  which generalizes to (if all S sets $1 \ldots N$ are disjoint)

$$P\left(\bigcup_{i=1}^{N} S_i\right) = \sum_{i=1}^{N} P(S_i) \tag{2.10}$$

These axioms lead to several useful observations. First, the probability of a complement of a set A, A^c, is $1 - P(A)$. Second, the probability of a empty set is 0, or $P(\emptyset) = 0$. Third, if $A \subset B$, then $P(A) \leq P(B)$. Finally, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Other important observations and laws include the definition of conditional probabilities

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{2.11}$$

assuming, of course, that $P(B) \neq 0$. This can work backwards,

$$P(A \cap B) = P(A \mid B) P(B) \tag{2.12}$$

This makes life easier down the road when you consider the situation in which $B_1 \ldots B_n$ are a set of mutually disjoint events with non-zero probabilities. In that case, you can get the total probability of another event A with

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i) \tag{2.13}$$

Which leads to Bayes' rule. If we have the same circumstances, then we can actually find the probability of an event B_j in the other way

$$P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)} \tag{2.14}$$

The frequentist approach produces probabilities of statistics on the basis of a hypothetical set of repeated samples or experiments. Bayesians, on the other hand, assert that the probability of a statistic should be based on the researchers quantification of the probability (usually based on computer simulations). Neither is without merit or warts. We take the frequentist approach here because it is more common, but not because it is necessarily "better."

This is all well and good, but how to you compute probabilities? Basically, if you have a sample space Ω with elements $\{\omega_1 \ldots \omega_n\}$, each element has a certain probability $P(\omega_i) = p_i$. If we want the probability of A, we add all the probabilities in the set A. For example, dice again, the probability of an even number roll when all sides are equally possible are

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

where

$$p_1 = \frac{1}{6} \dots p_6 = \frac{1}{6}$$

so

$$P(even) = p_2 + p_4 + p_6 = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} = 0.5$$

In general, if the sample space has N elements, and if A can occur in any n mutually exclusive ways, then $P(A) = \frac{n}{N}$.

In order to talk samples, we now need to discuss permutations. A permutation is an ordered arrangement of elements–in other words, order matters in permutations. It is important to note whether our permutation occurs with or without replacement.

If we have a sample space of N elements and a sample size of n, there are a possible N^n number of unique ordered samples that can be drawn if we replace elements after drawing them. If we do not replace, then the expression is more complicated in that we have $N(N-1)\dots(N-n)(N-n+1)$ different ordered samples.

Of course, order in sample rarely matters, so we really care about combinations of elements. If we sample without replacement and do not care about order, then the number of possible unordered samples are

$$\frac{N!}{(N-n)!n!} \tag{2.15}$$

which we can codify as $\binom{N}{n}$.

We always talk about independent sample elements. Therefore, we can say that events A and B are independent when $P(A \cap B) = P(A)P(B)$.

Chapter 3

Simple random samples

These notes assume your data come from a simple random sample. Here, I describe what that actually means.

Simple random samples are the most basic of "good" samples. Essentially, you start with a sampling frame and each element has an equal chance of being selected. You then randomly selected n elements.

Other types of more complex samples are those usually used in large-scale surveys. While the estimation procedures for complex samples are more difficult, the cost savings and quality assurance that come with this extra effort are worth it. Some examples of complex samples include stratified random samples, cluster samples, and stratified cluster samples.

3.1 Back to probability

We think of the universe of objects to which we want to generalize to as a finite population. This is the sampling space Ω, defined in Chapter **??**, where each element is indexed by i:

$$\Omega = \{i = 1, i = 2, i = 3, \ldots, i = N\}$$

From this population, we have a set of $\frac{N!}{(N-n)!n!}$ unordered samples of size n, each with a known probability. In simple probability samples, each possible sample S_s has an equal chance of being selected.

Because each sample has a known probability, we can determine the chance of any one unit being selected by adding up the probability of each

sample in which they appear. That is, if we have $\{S_1, S_2, \ldots, S_m\}$ possible samples and

$$I_{ij} = \begin{cases} 1 & \text{unit } i \text{ in sample j;} \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

then

$$P\left(\text{unit } i \text{ in sample}\right) = \pi_i = \sum_{j=1}^{m} I_{ij} P\left(S_j\right) \tag{3.2}$$

.

3.2 Expectation

The backbone of statistical inference depends on the idea of a sampling distribution and expectation. A sampling distribution is a statistic's distribution based on all possible samples. That is, if we took the mean of each possible sample, the sampling distribution is the distribution of each sample's mean.

A statistic's expected value is simply the mean of the sampling distribution. That is, for some statistic θ, the expected value is the sum of each sample's estimate of θ, or $\hat{\theta}$, times the probability of drawing that sample.

$$E\left[\hat{\theta}\right] = \sum_{j=1}^{m} P\left(S_j\right) \hat{\theta}_j \tag{3.3}$$

We worry about bias when there is a difference between the expected value and the true value, Bias $= E\left[\hat{\theta}\right] - \theta$. Finally, standard errors come from the variance of the sampling distribution, which is

$$var\left[\hat{\theta}\right] = \sum_{j=1}^{m} P\left(S_j\right) \left(\hat{\theta}_j - E\left[\hat{\theta}\right]\right)^2 \tag{3.4}$$

3.3 Simple random sampling

Life is pretty simple with simple random sampling (without replacement), the probability of each possible sample is the same,

$$P\left(S\right) = \frac{1}{\binom{N}{n}} = \frac{n!\left(N-n\right)!}{N!} \tag{3.5}$$

in these cases, since $P(S)$ is the same for all samples, the estimates for various parameters are very simple. The population mean of a variable y is

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{3.6}$$

with a variance

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2 \tag{3.7}$$

and standard deviation of $\sqrt{\sigma^2}$.

If we have a sample, the chance of being in the sample is a random variable,

$$I_i = \begin{cases} 1 & \text{unit } i \text{ in sample;} \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

and so we think of the estimated mean as

$$\bar{y} = \frac{1}{\sum_{i=1}^{N} I_i} \sum_{i=1}^{N} y_i \times I_i \tag{3.9}$$

which reduces to

$$\bar{y} = \frac{1}{n} \sum_{i \in S} y_i \tag{3.10}$$

and the sample variance is

$$s^2 = \frac{1}{n-1} \sum_{i \in S1} (y_i - \bar{y})^2 \tag{3.11}$$

There are other details, like the finite population correction term, that are important to sampling. Since this is mainly about regression, we will skip these for now.

Chapter 4

Some distributions

Statistics is all about comparing our data to the expected distribution of some variable. Whether we are modeling a function of a variable to a specific distribution or assuming a specific sampling distribution, it is always a good idea to have a good working knowledge of distributions.

4.1 Discrete distributions

A discrete random variable can only take on positive integer values. Such variables are dichotomous variables such as whether a respondent voted (0 or 1) or a count variable such as the number of children (0, 1, 2, ... etc.). We will revisit these distributions when it comes time to model variables that follow the distribution in Chapter **??**.

There are two important concepts for any probability distribution:

- The probability mass function
- The cumulative distribution function

The probability mass function (PMF) specifies a probability for observing any one possible value of a random variable, X. For example, we can speak of the probability that X is equal to a specific value x_i

$$P\left(X = x_i\right) \tag{4.1}$$

and that the sum of the probabilities of all possible values is 1

$$\sum_{i=0}^{\infty} P\left(X = x_i\right) = 1 \tag{4.2}$$

We also can talk about the cumulative distribution function, which adds the probabilities for each value of X up to and including a specific value, x_i

$$F(x_i) = P\left(X \leq x_i\right) \tag{4.3}$$

4.1.1 Bernoulli distribution

A Bernoulli variable takes on two values, 0 or 1. We are generally interested in the probability of a 1 (p) or the probability of a 0 ($1-p$). The probability mass function of this distribution is

$$P(x) = p^x \left(1-p\right)^{1-x} \tag{4.4}$$

4.1.2 Binomial distribution

We then extend the Bernoulli distribution to consider n Bernoulli events. That is, out of n times, how many instances are 0 or 1? The probability mass function of k events of 1 is

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.5}$$

Figure **??** illustrates likely distributions of a binomial variable given values for n and p.

4.1.3 The Poisson distribution

The Poisson distribution is an extension of the binomial distribution. If we conceive of a parameter $\lambda = np$, where n approaches infinity and p approaches 0 (that is, the first couple trials have a high probability, but as trials increase the probability decreases), then this frequency distribution function becomes

$$P\left(X = k\right) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{4.6}$$

This distribution if often used to model count data.

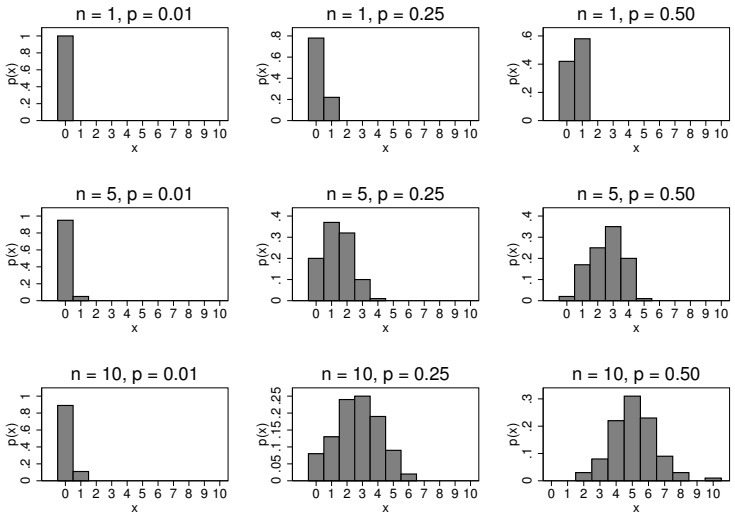

Figure 4.1: Distributions of random binomial variables with different values of n and p.

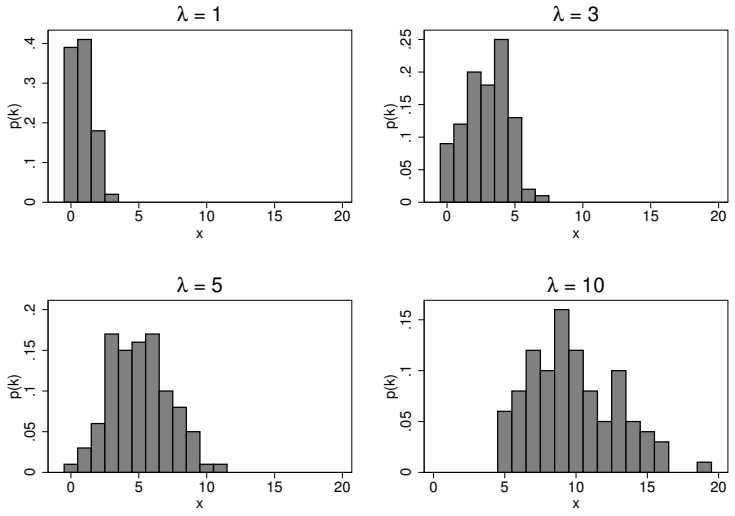

Figure 4.2: Distributions of random Poisson variables with different values of λ.

4.1.4 Negative binomial distribution

We can extend the Bernoulli distribution to consider a series of Bernoulli events, each with a probability p, but ending after r successes (or 1s). That is, out of n times, how many instances are 0 or 1? The probability mass function of k events to get r successes is

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \tag{4.7}$$

4.2 Continuous distributions

The analogous idea to the probability mass function for discrete variables is the probability density function (PDF) that describes continuous variables. These functions have the property that the cumulation of all possible values equals 1. More importantly, we can find the probability that a variable has a value with the interval between a and b by integrating the function

$$P\left(a < X < b\right) = \int_a^b f(x)dx \tag{4.8}$$

Of course, the chance of a specific value is 0

$$P\left(X = c\right) = \int_c^c f(x)dx = 0 \tag{4.9}$$

4.2.1 The uniform distribution

A useful distribution is the uniform distribution that spreads the chance of a variable falling between a and b evenly

$$f(x) = \frac{1}{b-a} \tag{4.10}$$

4.2.2 The normal distribution

The big one is the normal distribution. We use this distribution (and others based on it) all statistical tests in regression. We define it with two parameters, μ and σ, or the mean and standard deviation.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-\left(x-\mu\right)^2}{2\sigma^2}\right] \tag{4.11}$$

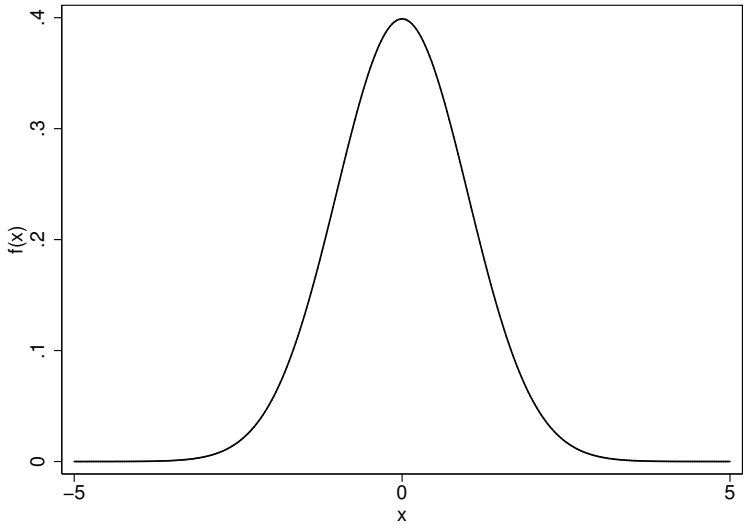

Figure 4.3: A variable x distributed $N \sim (0, 1)$.

The standard normal is shown in Figure **??**. A standard normal curve is one with $\mu = 0$ and $\sigma = 1$.

4.3 Distributions based on the normal distribution

From here, we move to the major distributions used for statistical tests. The central idea of such distributions is less about describing variables in data, but more describing the sampling distributions of statistics. We encounter these distributions when we consider the probability of observing a statistic (or the data that generated a statistic) assuming a specific null hypothesis in Chapter **??**. Note that some details, such as moment generating functions and the gamma distribution are left out, for more details, check out the mathematical statistics text in the bibliography.

4.3.1 The χ^2 distribution

The χ^2 distribution (written chi-square and rhymes with sky-square). This distribution (with 1 degree of freedom) is essentially the square of the normal

distributions. That is, if Z is standard normal, then $U = Z^2$ and U is the χ^2 distribution with 1 degree of freedom.

We note the distribution as χ^2_{df}. We can take any normally distributed variable with mean μ and standard deviation σ and transform it into a chi^2_1 by transforming it into standard normal and squaring it. That is, if

$$X \sim N(\mu, \sigma)$$

then

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

so

$$\left(\frac{X - \mu}{\sigma}\right)^2 \sim \chi^2_1$$

Formally, the χ^2 distribution employs the gamma function (Γ) to find the density for any value (v) greater than 0 with a certain number of degrees of freedom (df):

$$f(v) = \frac{1}{2^{df/2}\Gamma(df/2)} v^{(df/2)-1} e^{-v/2} \tag{4.12}$$

Finally, what really makes this distribution useful is that we can *add* independent variables' distributions and find the joint probability distribution (again, only if independent). Thus, if $U \sim \chi^2_n$ and $V \sim \chi^2_m$, then $U + V \sim \chi^2_{n+m}$. Thus, you will find many statistics used to evaluate aspects of a regression model will follow a χ^2 distribution.

4.3.2 The t distribution

The distribution you will encounter most often is the t distribution, or sometimes called the "student's" t distribution. This is not because a student derived it, far from it, but it was instead derived by a guy named Gossett in 1908 who was working for Guinness (as in beer) at the time, and published it under the pseudonym "student" for various reasons.

The t distribution essentially works like the normal distribution, except it takes into account the small(er) sample sizes often used in research. If independent variables $Z \sim N(0, 1)$ and $U \sim \chi^2_{df}$, then the t distribution is

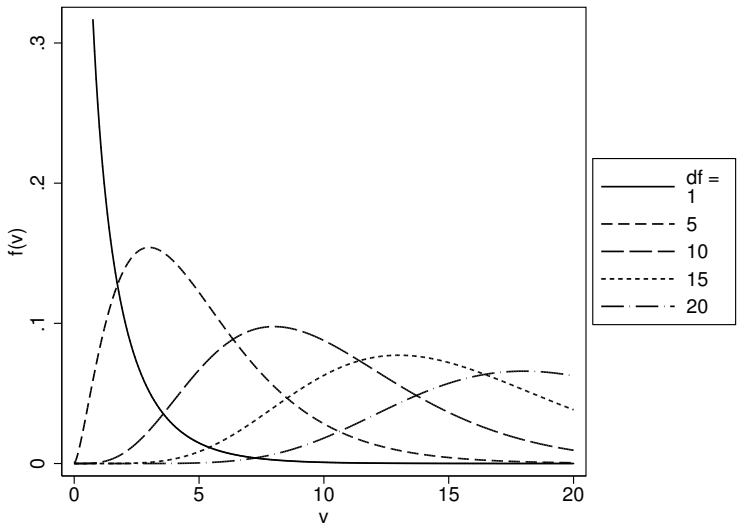

Figure 4.4: Plot of χ^2 distribution with various degrees of freedom.

the ratio of the normal Z to the square root of the ratio of $\chi^2 U$ with df degrees of freedom to df, or $t = \frac{Z}{\sqrt{U/df}}$.

Again, using the gamma function, this can we written as the following density function

$$f(t) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{df\pi}\Gamma\left(\frac{df}{2}\right)}\left(1+\frac{t^2}{df}\right)^{\frac{-(df+1)}{2}} \tag{4.13}$$

4.3.3 The F distribution

Finally, we have the F distribution. If we have two independent χ^2 variables, U and V, distributed with m and n degrees of freedom respectively, then $F = \frac{U/m}{V/n}$, or formally for all positive values of F,

$$f(F) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} F^{\frac{m}{2}-1}\left(1+\frac{m}{n}F\right)^{\frac{-(m+n)}{2}} \tag{4.14}$$

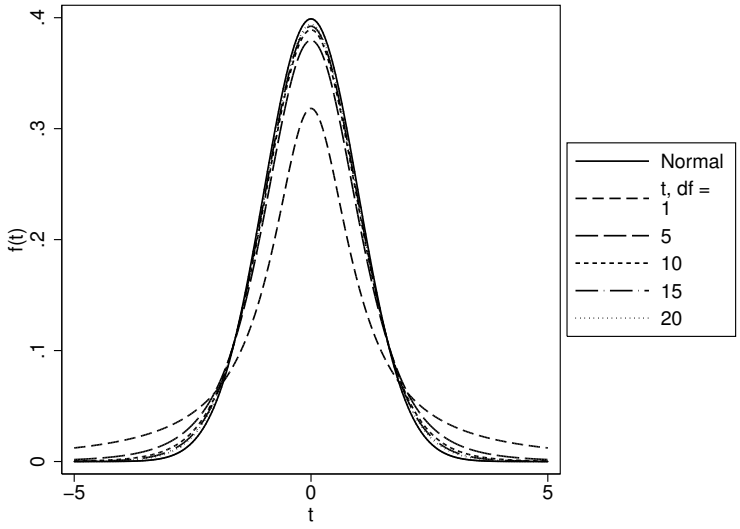

Figure 4.5: Plot of the normal distribution and two t distributions with various degrees of freedom

4.3.4 Who cares?

There are several notable aspects of statistical analysis that make all these distributions based on the normal curve important. First, means and variances are independent with normally distributed variables, making modeling them in regressions very easy. Second, variances follow χ^2 distributions, making the F-distribution appropriate for the ratio of variances.

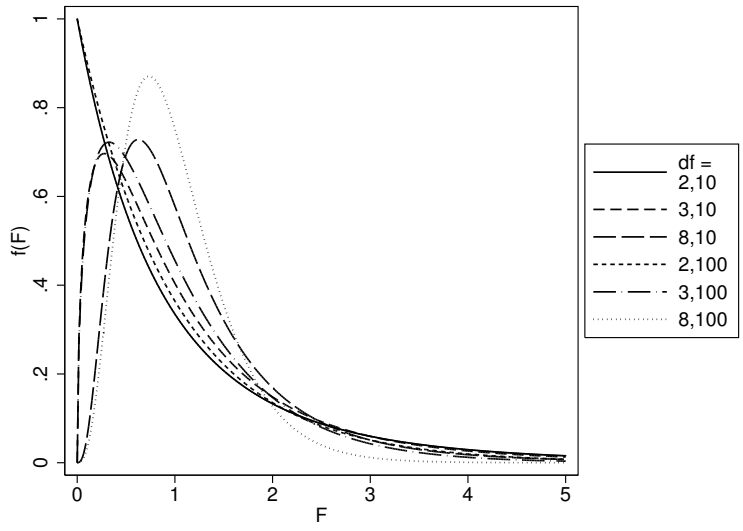

Figure 4.6: Plots of F distribution with various degrees of freedom combinations.

Chapter 5

Basic Statistical Inference

5.1 Sampling distributions

Almost all the methods in these notes assume a simple random sample. **If it is not a simple random sample, such as a two-stage cluster randomized sample from a major survey, these notes do not apply.** While the population mean, μ, is fixed, the estimation of the mean, $\bar{x}$, is random because the selection of the sample is random. It therefore makes sense that if we were to gather another sample, that new sample's estimation of the mean will be different. This is the idea of sampling distributions: given a large number of samples, how will the estimate likely vary? What would the range of our estimate be 95 percent of the time?

5.1.1 Basic principles of the central limit theorem

If repeated random samples of size N are drawn from any population with mean μ and standard deviation σ, as N becomes large the sampling distribution of sample means will approach normality, with a mean of μ and standard deviation $\frac{\sigma}{\sqrt{N}}$.

There are two important parts here: that the mean of the sampling distribution should be close to the actual mean, and the standard deviation of the sampling distribution should be close to the standard deviation of the population divided by the square root of the sample size.

Keep in mind that the population distribution doesn't have to be normal– but the sampling distribution becomes normal with large N.

It is important to note that the central limit theorem suggests that the mean of our sample can be used as our guess of the population mean and we can quantify the uncertainly of this estimate with the variance of the estimates, of which the standard error is the square-root. This standard error is an estimate of the sampling distribution's standard deviation.

5.2 The standard error of the mean

Of course, it is extremely expensive to obtain repeated samples to get a good estimate of the mean. Researchers generally only have the resources for a single sample.

The formula for the standard deviation of samples of the a mean is similar to the standard deviation of the sampling distribution, except we substitute the population standard deviation, σ, with the estimate of the standard deviation from our sample, s. We start with the variance of the statistic, $\bar{x}$:

$$V(\bar{x}) = \frac{\sigma^2}{N} \tag{5.1}$$

the square-root of which is the standard error

$$SE(\bar{x}) = \sqrt{\frac{\sigma^2}{N}} \tag{5.2}$$

or the more familiar formula

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{N}} \tag{5.3}$$

and our estimate of this replaces σ with s

$$SE(\bar{x}) = \frac{s}{\sqrt{N}} \tag{5.4}$$

This gives us the "variance" of our mean estimate. All statistics have "variances" associated with them. Much of the more advanced statistics and sample statistics that are taught in graduate schools focus on getting the right variances, and thus the right standard errors. Much the time, the formulas of the point estimates (the statistics) are pretty much the same.

5.3 The standard error of a proportion

The variance of a mean is not a function of the value of the mean, it just depends on the standard deviation. In the case of proportions, the standard deviation is a function of the estimate itself:

$$\sigma_p = \sqrt{p\left(1-p\right)} \tag{5.5}$$

therefore, the standard error of the proportion is

$$SE\left(p\right) = \sqrt{\frac{p\left(1-p\right)}{N}} \tag{5.6}$$

5.4 Hypothesis testing

So now that we have an estimate of the mean, and a sense of our uncertainty about this estimate, we can now ask a theoretical question: let's say we did draw several samples from the population, what is the range for 95 percent of the means we would get? We can do this by calculating a confidence interval of the estimated statistic. Confidence intervals generally take the form of

$$CI\left(\theta, (1-\alpha)\times 100\right) = \hat{\theta} \pm t_{df,\alpha/2} SE\left(\hat{\theta}\right) \tag{5.7}$$

Where θ is the parameter of interest, $\hat{\theta}$ is the statistic that estimates the parameter, $SE\left(\hat{\theta}\right)$ is the standard error of that statistic, α is the Type I error rate, and t is the critical value of the t-distribution associated with the df degrees of freedom of the statistic and the Type I error rate. As you can see in Figure **??**, the t distribution is similar to the Z distribution except that it has slightly different shapes for different degrees of freedom.

In this case, the parameter of interest is the population mean, μ. We are estimating that with the sample mean statistic, $\bar{y}$. Note that we are now talking about a variable y instead of x, this is because we are transitioning from any old variable to outcomes, which are typically represented with the variable y. For that statistic, we also have a standard error $SE\left(\bar{y}\right)$.

In the social sciences, the common Type I error rate is 5 percent, or $\alpha = 0.05$. What is the Type I error rate? Type I error is the chance of falsely rejecting the hypothesis that the true parameter could fall outside the confidence interval. In other words, if we are going to make the assertion that

some other parameter, θ_{null}, is different than our population parameter, θ, for which we have a sample estimate, $\hat{\theta}$, we want there to be only a $(1-\alpha)\times 100$ percent chance of being wrong.

Finally, we need to scale our standard error by something to create the confidence interval. Since the sampling distribution is normal with a standard deviation of $SE\left(\hat{\theta}\right)$, we want to form our confidence interval so that it gives us the bounds of percent of the possible sample estimates.

One option is to use the normal Z distribution, in which case, for example, a 95 percent confidence interval (meaning that $alpha = 0.05$) is bounded by -1.96 and 1.96. However, we typically use the t-distribution rather than the normal distribution because it "flattens out" as the sample size gets small, making the confidence intervals wider (see Figure **??**). Thus, in order to get the right number from the t-distribution, we need to know the degrees of freedom, which in a one sample case is $N-1$. Thus, confidence intervals for smaller samples will always be larger. They are larger for two reasons. First, if we look at the formula for a standard error,

$$SE\left(\bar{x}\right) = \frac{s}{\sqrt{N}} \tag{5.8}$$

we see that the denominator is essentially the sample size. Smaller samples, larger standard errors. The second reason is that we use the sample size to determine the degrees of freedom when selecting the value from the t distribution. Smaller samples, larger values of t for any given α.

Why do we care about this? We care because we may want to make an assertion that the estimated parameter, $\hat{\theta}$, is different from some other value, θ_{null}. If θ_{null} is outside the confidence interval, then the estimate is statistically different. Another way is to find the difference between $\hat{\theta}$ and θ_{null}, and divide by $SE\left(hat\theta\right)$. This gives you the value of t for a confidence interval, and you can compare it to the t you would use to construct a confidence interval, or the critical t. If

$$\frac{\hat{\theta}-\theta_{null}}{SE\left(\hat{\theta}\right)} > t_{critical,df,\alpha/2} \tag{5.9}$$

then the estimate is statistically different.

Can we be absolutely sure? No, because our confidence interval only covers $(1-\alpha)\times 100$ percent of the sampling distribution, another sample could have given us a different answer.

This is all backing into hypothesis testing. We could play this game all day of calculating confidence intervals and seeing if some arbitrary value falls within the bounds. A much simpler approach is to do a hypothesis test.

Most stats text books will tell you that there are 5 steps to doing a hypothesis test that include making an assertion of our Type I error we assume, stating the critical value, etc. I think these steps are silly, since no one ever does these steps in discrete order. This is how it is done in practice. We typically think of θ_{null} as the null hypothesis (or the value of the parameter or model we will assume to be true). We can write the null hypothesis like this (in terms of the population parameter θ:

$$H_0 : \theta = \theta_{null}$$

where we say that the population parameter, θ, is equal to the null value, θ_{null}. The alternative hypothesis is that these values are unequal

$$H_1 : \theta \neq \theta_{null}$$

First, we calculate our test statistic. Test statistics for parameters that use the t-distribution generally take the form of

$$t = \frac{\hat{\theta} - \theta_{null}}{SE\left(\hat{\theta}\right)} \tag{5.10}$$

for t-tests when we have a specific degrees of freedom, or

$$z = \frac{\hat{\theta} - \theta_{null}}{SE\left(\hat{\theta}\right)} \tag{5.11}$$

for z tests when there are no real degrees of freedom to consider.

Basically, what we are doing is calculating the difference between what we estimated and what we assume is true (the null hypothesis) and then we divide that difference by the standard error of our estimate.

We then look up the probability of that test on the appropriate distribution associated with the statistic and degrees of freedom. If that probability is less than 0.05, then we say we can reject the null. I do not talk about one-tail tests.

A probability less than 0.05 means that what we get from our sample is highly unlikely if we assume the null hypotheses, so we can then reject the

null hypothesis. There is still a chance that the null is correct and we got a weird sample, that's our Type I error rate, but we feel that 5 percent or less is acceptable.

5.4.1 What is a p-value?

When you have a result from one of the major distributions (Z, χ^2, t, or F) and the associated degrees of freedom (if applicable), you can generate a p-value. p-values allow us to estimate the smallest α could be for a given hypothesis test to be rejected. If the p-value is greater than 0.05, generally, you have to accept the null. We use p-values mainly because of Fisher, who thought of them as the probability under the null hypothesis of a result as, or more extreme, than the data.

5.4.2 Comparing two means

I really do not see many comparisons between a population mean and some other value. What I do see a lot of are comparisons between two groups that are sampled independently. We should be careful when we think about this, because in observational studies like the General Social Survey, we do not sample Republicans separately from Democrats. We randomly sample the population, and some are Republicans and some are Democrats. Even in experiments, we randomly sample a group then randomly assign treatment or control.

Regardless, we always talk about testing two independent means, for example between groups A and B. They are independent because the observations from group A are in no way related to observations from group B. However, we are not really dealing with a test of two statistics, we are estimating a single statistic, *the difference*. In population terms, this is

$$\theta = \mu_A - \mu_B,$$

and from our sample estimates it is

$$\hat{\theta} = \bar{y}_A - \bar{y}_B.$$

We then state the null hypothesis as this difference being 0 (i.e., $\theta_{null} = 0$) as

$$H_0 : \theta = \theta_{null}$$

$$H_0 : \mu_A - \mu_B = 0$$

and the alternative hypothesis as

$$H_1 : \mu_A - \mu_B \neq 0.$$

In the case of independent means, we can estimate the standard error of this difference using a pooled standard deviation

$$SE\left(\bar{y}_A - \bar{y}_B\right) = \frac{s_{pooled}}{\sqrt{N}} \tag{5.12}$$

where

$$s_{pooled} = \sqrt{\frac{\left(n_A - 1\right) s_A^2 + \left(n_B - 1\right) s_B^2}{n_A + n_B - 2}}. \tag{5.13}$$

The t test is then

$$t = \frac{\bar{y}_A - \bar{y}_B}{SE\left(\bar{y}_A - \bar{y}_B\right)} \tag{5.14}$$

We then evaluate the value of t against some level of α using $N - 2$ degrees of freedom.

If there is a relationship between groups A and B, like one "group" was a set of pre-tests and the second "group" was a set of post tests, each pre-test and post-test was from the same set of people, then we would be doing a paired t-test. In this case the test is

$$t = \frac{\bar{d}\sqrt{N}}{s_d} \tag{5.15}$$

where

$$\bar{d} = \frac{\sum_{i=1}^{N} \left(y_{Ai} - y_{Bi}\right)}{N} \tag{5.16}$$

and

$$s_d = \sqrt{\frac{\sum_{i=1}^{N} \left(\left(y_{Ai} - y_{Bi}\right) - \bar{d}\right)^2}{N - 1}} \tag{5.17}$$

with N pairs of values. The test is then evaluated against the t distribution using $N - 1$ degrees of freedom.

Chapter 6

Basic multivariate analyses

6.1 The analysis of variance (ANOVA)

When we are considering three or more groups, we ask whether these groups are different from each other. Another way to think about this is to think about how much of the variance occurs between the groups compared to how much variance occurs within the groups. This ratio is the idea behind ANOVA.

The building blocks of ANOVA is that we can think about each i^{th} case as part of the j^{th} group. We can then conceptualize a way to think about deviance that involves groups.

Here, we can think of the total difference between an observation and the overall mean, $y_{ij} - \bar{y}$, as the difference between that observation and its group's mean, $y_{ij} - \bar{y}_j$, plus the difference between that group's mean and the overall mean, $\bar{y}_j - \bar{y}$:

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_j) + (\bar{y}_j - \bar{y}) \tag{6.1}$$

This idea has implications for the sum of squares. We can now think of the total sum of squares (SST) as breaking down into the sum of squares within groups (SSW) and the sum of squares between groups (SSB):

$$SST = SSW + SSB \tag{6.2}$$

or

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 + \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2 \tag{6.3}$$

There is some new notation here. We now have k groups with the subscript j, and each j^{th} group has cases with the subscript i.

ANOVA's test statistic relies on the F-distribution to evaluate Type I error. The F-distribution is used to test ratios and is thus defined by two degrees of freedom, one for the numerator and one for the denominator of the ratio, and there are no negative statistics. As the number of each increases to a large number, the distribution starts to look normal.

As was just mentioned, the test statistic is a ratio. Again, we are concerned with the ratio of the variance that occurs between groups compared to that within groups. Thus, this ratio uses as the numerator the mean square between, which is the sum of squares between divided by the number of groups - 1:

$$MSB = \frac{\sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{y}\right)^2}{k-1} \tag{6.4}$$

The denominator of the ratio is the mean square within, which is the sum of squares within divided by the number of cases minus k group means

$$MSW = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y_{ij}}\right)^2}{N-k} \tag{6.5}$$

The ratio then looks like this

$$F = \frac{MSB}{MSW} \tag{6.6}$$

and we test how much of the F-distribution (with k and $N-k$ degrees of freedom) is left after this value for our Type I error rate of our test of the the null hypothesis that all group means are the same. In other words, our null hypothesis is that all the group means are the same:

$$H_0 : \mu_1 = \mu_2 \ldots = \mu_k. \tag{6.7}$$

Since this distribution has two degrees of freedom parameters, the critical value different for different combinations of k and $N-k$. See Figure **??** for various distributions.

6.1.1 ANOVA Example

As a cartoony example, let's use political views to predict vocabulary. Or, in less causal language, whether vocabulary varies by political view. These data come from the General Social Survey.

First, we have three political groups, liberal (lib), moderate (mod), and conservative (con). Summary statistics appear in Table **??**.

Table 6.1: Number correct words by political affiliation

Political affilication	N	$\bar{y}$	s
lib	733	6.32	2.21
mod	983	5.81	1.84
con	875	6.23	1.97
Total	2591	6.10	2.01

Source: General Social Survey: 2008

Table **??** is a typical ANOVA table with the total sum of squares broken down into the sum of squares within and between, with the mean squares also calculated. We can calculate the F-test using the mean squares

$$F = \frac{MSB}{MSW} = \frac{67.216}{3.984} = 16.87.$$

Which, evaluated with $(2, 2588)$ degrees of freedom on the F distribution, has a probability of 0.0000005. This indicates that at least one group is different than the others. To find out which group is different requires some regression analysis.

Table 6.2: ANOVA of word test by political views

Source	Sum of squares	*df*	Mean Squares
Between groups	134.432	2	67.216
Within groups	10311.878	2588	3.984
Total	10446.31	2590	4.03
Model Statistics			
N	2591		
F	16.87		

6.2 Covariance

The covariance is a simple idea: instead of squaring the deviance of one variable from its mean, the covariance multiplies the deviances of two variables

from their means:

$$\operatorname{cov}(y, x) = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{N - 1} \tag{6.8}$$

We can visualize this by using a scatterplot. Figure **??** is a scatterplot of

Table 6.3: Small dataset of random variables

y	x
8.20	4.51
9.21	5.36
10.22	6.23
8.56	3.84
10.66	4.90
9.93	5.00
8.29	4.80
9.18	4.34
9.22	5.78
7.29	4.06

the data with the means of x and y marked with vertical and horizontal lines and each point marked with the deviations from the mean. The data appear in Table **??**. Now, consider Figure **??**. The means of x and y have been graphed into four quadrants. Each quadrant is defined by the sign of the deviation from the means of x and y. Observations whose values of x and y are both larger than the mean of x and y respectively will fall into quadrant one $(+, +)$. Observations whose values have x and y are both lower than the respective means will fall into quadrant three $(-, -)$. The key is that the product of any deviations in either of these quadrants will be positive (a negative times a negative is positive). On the other hand, points that fall into the other quadrants (quadrant two and quadrant four) will have products that are negative (a negative times a positive is negative).

Why do we care about the products of the deviations? Consider the numerator of the covariance formula, $\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})$. All this is doing is adding up these products. Therefore, more positive products will produce a large positive number, more negative products will produce a large negative

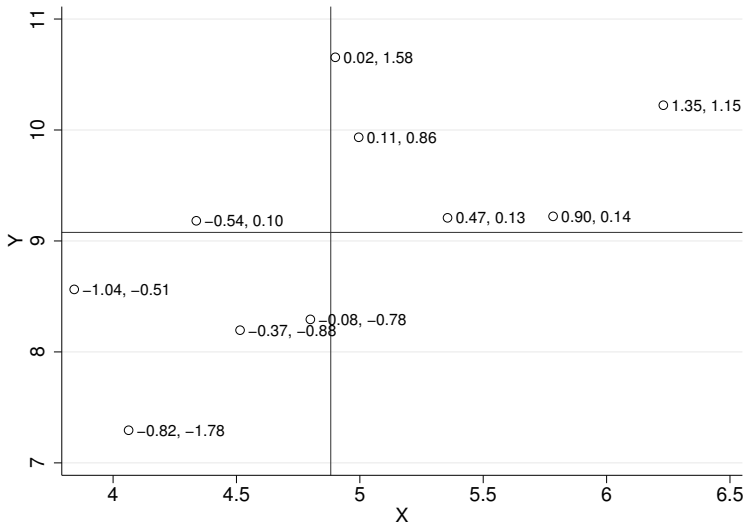

Figure 6.1: Plot of variables y and x with solid lines for means, labeled with deviations from the means

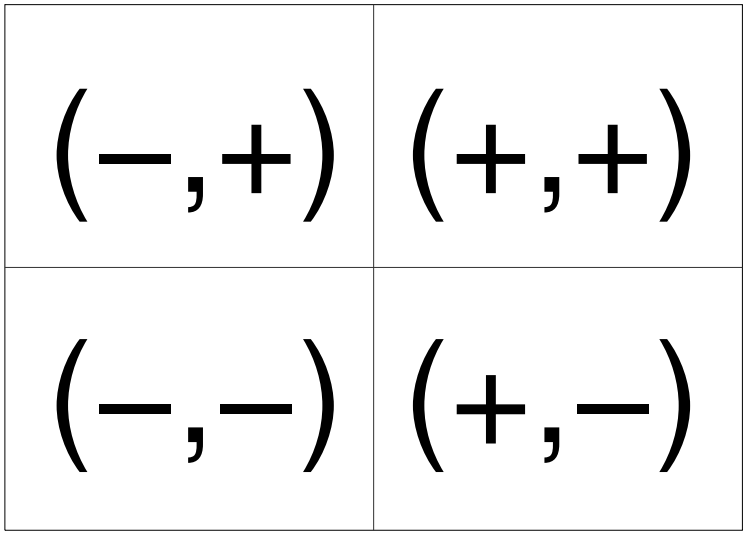

Figure 6.2: Covariance quadrants

number. In addition, larger deviations of either x or y will create larger numbers, and large deviations of both x and y will create a really big number.

If more points fall into the positive quadrants $(+,+;-,-)$ the result is a positive number. If more points fall into the negative quadrants $(-,+;+,-)$, the result is negative. If equal numbers fall into each quadrant, then they will balance out and the number will be close to 0.

Figure **??** shows the deviations marked for each point on the scatterplot: what do you think the covariance is?

The denominator of the covariance formula $(N-1)$ just turns it into an average, like the variance. Looking at Figure **??**, we can tell that the relationship between x and y is positive. The direction of a relationship is generally thought of as what happens to y as x increases. Thus, the covariance in Figure **??** is 0.47. This tells me that when x increases, so does y. If this number was negative, then as x increases y would decrease.

6.3 Correlation

One of the biggest problems with covariance is that it is really hard to interpret anything meaningful from it. Covariance identifies if the relationship is positive or negative, but it doesn't provide information about the magnitude of the relationship. In addition, the units get all messy. For instance, if the variables were wages and years of education, then the units would be wage-years. What is a wage-year? The solution to this problem is standardization. It would be helpful to standardize this quantity, and one method is to use the sum of squares of the two variables involved. This is basically what the correlation coefficient is about. It is a standardized covariance. The correlation formula is

$$r_{y,x} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(x_i - \bar{x})^2}} \tag{6.9}$$

The numerator of this equation is simply the covariance formula while the denominator is the sum of the squares equation. Since the units of the variables are the same for the numerator and denominator, the units cancel out

Correlation coefficients range from -1 to 1. If it is 0, it means that there is no covariance between x and y. If it is close to 1, it means that there is a

strong positive covariance between x and y. If it close to -1, it means that there is a strong negative covariance between x and y. All information about the units is lost. It does not matter what x and y are measured by.

6.3.1 Testing the correlation coefficient

What if we wanted to know if our correlation coefficient was *statistically* different than 0? In this case, we need to consider the sampling distribution of the correlation coefficient. I go into this in more detail below, but for now remember that we have a random *sample*, and if we started our research another day, we would get a different *sample*. Thus, the estimated correlation coefficient from our data is one of many possible estimates. We want to know the distribution of these estimates from different hypothetical samples, so we can know how likely our estimate is, assuming a world where a null hypothesis is true. In this case, our null hypothesis is that there is no correlation. For the data in Figure **??**, the correlation is 0.61. The question is, is this statistically different than 0?

Since we have a sample, we will rely on the student's t distribution. We can estimate a test statistic for a correlation with

$$t = \frac{|r|\sqrt{N-2}}{\sqrt{1-r^2}} \tag{6.10}$$

which in this case is

$$t = \frac{0.61\sqrt{10-2}}{\sqrt{1-0.61^2}}$$

$$t = 2.18$$

Now, in our case in Table **??**, we have 10 cases, and with 2 means (for two variables) we have 10-2 = 8 degrees of freedom. Figure **??** is the sampling distribution of our statistic if we assume that the null hypothesis, or the hypothesis that the correlation is 0, is true. If you look carefully, you will see that the probability of finding a correlation of 0.61 with 10 cases is about 0.06. The math to find that number is a little tedious, so most stats books tell you what test you need to get in order to have a sufficiently low probability to reject the null hypothesis. In our case, that value is about 2.31. Why 2.31? That is a number large enough to claim that if we assume the null hypothesis, the chance of finding that test in our data is less than 0.05. Another way of

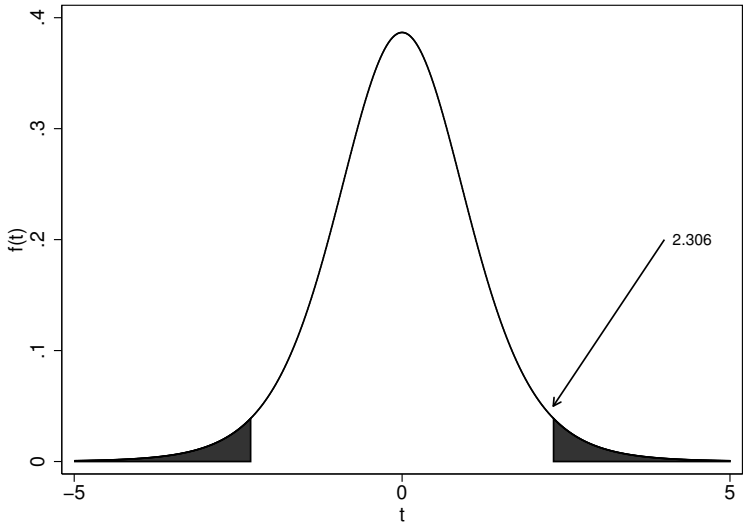

Figure 6.3: Plot of t distribution with 8 degrees of freedom and $\alpha = 0.05$ critical tails marked.

saying that is that the Type I error rate is less than 5 percent. We sometimes call the Type I error rate α and so we say that $\alpha = 0.05$.

In this case, with a test statistic of 2.18, which is smaller than the critical value of 2.31, we have to accept the null hypothesis. It was close, but not enough.

6.4 Chi-square

In the case of two categorical variables, neither covariance or correlation should be used to determine relationships. Instead, another method based on the idea of deviation is appropriate; the Chi-square (χ^2) test. The idea is to state a "null" hypothesis in terms of the researcher's expectation for the results of their analysis. The researcher states what she/he expects, and then offer an alternative hypothesis. If the data deviates in a substantial degree from what the researcher "expects," then the researcher must "fail to reject" the null hypothesis. In common practice, a null hypothesis indicates that there will be no statistically significant relationship between two variables while the alternative hypothesis indicates that there is a statistically

significant relationship between the two variables.

The trick to Chi-square tests is that the "expected" value can be anything. You may remember the Chi-square tests from your research methods or statistics class. Frequently, data is presented in table format. In such a table we have (at least) two categorical variables x and y. The variable x takes on values such as $x \in \{1 \dots i\}$ and variable y takes on values such as $y \in \{1 \dots j\}$. Each cell in the table has a number of observations associated with it, n_{ij}, for a total of $N = \sum_i \sum_j n_{ij}$ observations. In the case of a two variable table, x makes up the rows and y makes up the columns. We also have totals for each row, n_{i+}, and for each column, n_{+j}. Chi-square tests then make a comparison between the observed frequency, n_{ij} and what the researcher "expects" each cell frequency to be, m_{ij}. This test is

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}. \tag{6.11}$$

A big value for this number tells us that there is a difference between our data and our expectation. Depending on what we are doing, this expectation is different. If we are testing to see if two variables are independent, when we test against a random expectation, or we expect the cells to simply be a function of the margins. So, in this case, m_{ij} is

$$m_{ij} = \frac{n_{i+} n_{+j}}{N} \tag{6.12}$$

So, what's a big value? In this case, it depends on how many cells are in the table. Any statistical test is compared to a function, generally based on some degrees of freedom. For example, a Chi-square test is compared against a Chi-square distribution with $(rows - 1)\,(columns - 1)$ degrees of freedom. If the value of the test is far enough along that distribution to be considered unlikely if the expectation is true, then we reject the expectation. For example, a 2×2 table has a single degree of freedom and a critical value of 3.84 for a Type I error rate of 0.05. Figure **??** display some χ^2 distributions.

Chapter 7

Least squares and maximum likelihood

7.1 The mean

7.1.1 Least squares

These notes are for regression. So now is the time to know that regression produces a conditional mean (a mean that exists under certain specified conditions). Since learning regression is what we are here to do, it makes sense to show how useful means really are. If we use as our criteria for whether a summary statistic about the data is "good" or not is the deviation between the observation and our statistic, then we want a summary statistic to minimize these differences.

As mentioned earlier, adding up all of the deviances will produce 0. To circumvent this problem, the deviances were squared and then added up. Therefore, we want our mystery statistic, θ to be a *function* of these squared deviations:

$$S(\theta) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (x_i - \theta)^2 \tag{7.1}$$

A "good" estimate will have the least amount of differences. Thus, we want our statistic to minimize $\sum_{i=1}^{N} (x_i - \theta)^2$. So, how to we figure out which statistics will minimize the sum of these squared deviations? Let's break it down into known and unknown quantities. Someone smarter than me figured

out that if you rearrange the terms, you can express this function as

$$S(\theta) = \sum_{i=1}^{N} (x_i - \theta)^2$$

factor it

$$S(\theta) = \sum_{i=1}^{N} ((x_i - \theta)(x_i - \theta))$$

$$S(\theta) = \sum_{i=1}^{N} \left(x_i^2 - x_i\theta - x_i\theta + \theta^2\right)$$

$$S(\theta) = \sum_{i=1}^{N} \left(x_i^2 - 2x_i\theta + \theta^2\right)$$

then take the elements out of the summation

$$S(\theta) = \sum_{i=1}^{N} x_i^2 - 2\theta \sum_{i=1}^{N} x_i + N\theta^2 \tag{7.2}$$

Since, theoretically, we know our data, and thus know the sum of x squared, $\sum_{i=1}^{N} x_i^2$, and the sum of x, $\sum_{i=1}^{N} x_i$, and N, then this function only has one unknown, θ. We can use calculus to find the first derivative of this function with respect to θ. Relying on the fact that the derivative of a sum of functions is the sum of the derivatives of each function, we can find the derivative of each element. The derivative of $\sum_{i=1}^{N} x_i^2$ is 0, since it is a constant and doesn't involve θ, the derivative of $-2\theta \sum_{i=1}^{N} x_i$ with respect to θ is $-2\sum_{i=1}^{N} x_i$, and finally the power rule tells us that the derivative of $N\theta^2$ is $2N\theta$. Adding this all together gets us the derivative of the function with respect to θ as

$$\frac{dS(\theta)}{d(\theta)} = -2\sum_{i=1}^{N} x_i + 2N\theta \tag{7.3}$$

Then we set this first derivative to 0 (the first derivative of a function is its slope, and when the slope of a function is zero, the function is at a minimum or maximum) and solve for θ:

$$0 = -2\sum_{i=1}^{N} x_i + 2N\theta$$

$$2\sum_{i=1}^{N} x_i = 2N\theta$$

$$\sum_{i=1}^{N} x_i = N\theta$$

$$\frac{\sum_{i=1}^{N} x_i}{N} = \theta$$

Now that we have it, we name the statistic μ and also call it the mean, c.f. (**??**). For example, Table **??** presents some random numbers and the key statistics to graph the least squares function. For Table **??** this function is

$$S(\theta) = \sum_{i=1}^{N} x_i^2 - 2\theta \sum_{i=1}^{N} x_i + N\theta^2$$

$$S(\theta) = 9887.521 - 2\theta 988.596 + N\theta^2$$

and this function is visualized in Figure **??**. I show that the mean we estimate does indeed present the minimum sum of squared errors. Any other value for the mean increases error.

7.1.2 Maximum likelihood

Maximum likelihood is a method that expresses the probability of observing a particular parameter, given the observed data. Instead of attempting to minimize the error, this approach attempts to maximize this probability, or likelihood; hence maximum likelihood.

Each time we use maximum likelihood, we much work our parameter estimator assuming a particular distribution. Thus, if our variable is normally distributed we will use a different formula than the one for a binomial variable.

If we are dealing with a normally distributed variable, we can start with the probably of observing the i^{th} value

$$\Pr(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\sigma^2 (x_i - \mu)^2\right], \tag{7.4}$$

and then the probability of observing a set of xs is then

$$\Pr(x_i \ldots x_N) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\sigma^2 (x_i - \mu)^2\right], \tag{7.5}$$

Table 7.1: Random variable x

10.939	10.470	10.333	8.738	9.827
10.279	9.754	10.261	11.036	10.054
10.637	10.659	10.742	10.492	11.260
10.378	9.007	10.092	9.430	10.618
8.147	10.414	10.250	9.378	9.728
10.338	8.780	10.539	10.504	8.474
8.430	10.244	11.622	9.584	10.187
8.431	9.532	9.669	9.460	9.913
7.349	9.671	10.607	9.558	10.503
9.519	9.866	9.010	9.059	8.804
9.709	9.194	9.318	7.893	10.205
11.039	11.717	9.642	11.321	9.710
9.370	10.890	10.162	8.852	9.086
9.902	11.128	9.445	10.676	9.374
9.932	11.948	12.313	10.559	11.943
8.238	9.870	7.799	9.278	8.727
12.202	6.458	10.487	7.822	9.337
10.667	11.530	9.796	9.316	9.571
8.543	10.585	10.686	10.117	11.155
8.411	9.177	10.265	9.108	11.548

$N = 100$

$\sum_{i=1}^{N} x_i = 988.596$

$\sum_{i=1}^{N} x_i^2 = 9887.521$

$\bar{x} = 9.886$

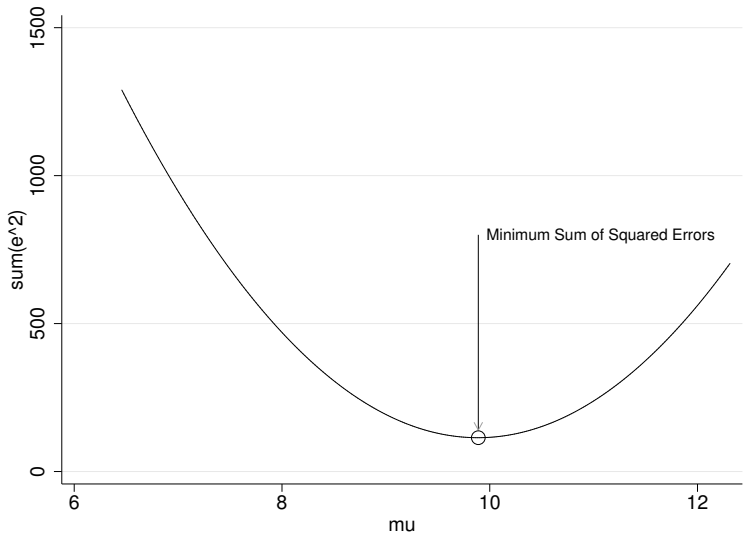

Figure 7.1: The SSE as a function of the estimate of the mean for random variable x.

or

$$\Pr\left(x_i \dots x_N\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left[-\frac{1}{2}\sigma^2 \sum_{i=1}^{N} \left(x_i - \mu\right)^2\right], \tag{7.6}$$

we can make the formula easier if we assume that $\sigma = 1$. If we consider the probability of several *independent* observations, this likelihood formula becomes a function of the mean

$$\Pr\left(\mu | x_1 \dots x_N\right) = L\left(\left(\mu | x_1 \dots x_N\right)\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^N \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \left(x_i - \mu\right)^2\right] \tag{7.7}$$

We can take the log of this function to make it more tractable

$$\ln\left(L\left(\mu | x_1 \dots x_N\right)\right) = -\frac{1}{2} \sum_{i=1}^{N} \left(x_i - \mu\right)^2 - \frac{N}{2}\left(\ln\left(\sqrt{2\pi}\right)\right) \tag{7.8}$$

which is something we can handle with basic calculus. We just need to find the derivative of this function with respect to μ. In this situation

$\frac{N}{2}\left(\ln\left(\sqrt{2\pi}\right)\right)$ is a constant that does not involve μ, and the power rule brings the square (the 2) down as a multiplayer, and $\frac{1}{2} \times 2 = 1$, so the derivative is

$$\frac{d\ln\left(L\left(\mu|x_1 \ldots x_N\right)\right)}{d\mu} = \sum_{i=1}^{N}\left(x_i - \mu\right)$$

take out the sums

$$\frac{d\ln\left(L\left(\mu|x_1 \ldots x_N\right)\right)}{d\mu} = \sum_{i=1}^{N} x_i - N\mu$$

set to 0 and solve

$$0 = \sum_{i=1}^{N} x_i - N\mu$$

$$N\mu = \sum_{i=1}^{N} x_i$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

and there we are, the mean formula again.

7.2 The proportion via maximum likelihood

While the least squares principle works well for dichotomous variables (the math doesn't care if the numbers only vary between zero and one), the next step is to introduce the idea of likelihood for non-normal distributions. This will be important once we start estimating logits and probits and other models with non-normal distributions. In this section, we figure out the maximum likelihood estimate of a proportion.

Let's start with an example. A researcher takes random poll of 100 people and asks them if they like sardines. Some people said they liked sardines and were coded as 1, while others didn't like sardines and were coded as 0. The frequency table is in Table **??**. The researcher would like to use this data to determine the probability that someone in the population will like sardines, y. Of course, we would estimate the mean of this variable to get the

Table 7.2: Frequency of dichotomous variable y

Item	Number	Percent
0=No	73	73
1=Yes	27	27
Total	100	100

$\bar{y} = 0.27$

proportion. Why is this the right thing to do? Let's start with the function that determines the chance of observing a "1" or "0" for some variable x, given a probability p

$$f\left(p|y\right) = \Pr\left(y = q\right) = p^{q}\left(1-p\right)^{1-q}, q = 0, 1 \tag{7.9}$$

The next important topic is the idea of a likelihood function. Researchers typically want to know the chance of observing the data they have if they assume some pre-specified model. In other words, if our model says that the chance of observing a 1 for any case is 50 percent, then how likely is it that we observed only 27 out of a 100? The likelihood function serves this purpose. It quantifies how likely we are to observe our data if we assume a specific model. In this case, our "model" is just some proportion we expect. Later, our models will become much more complicated multivariate functions. The likelihood function is

$$L\left(p|y_1 \ldots y_N\right) = \prod_{i=1}^{N} f\left(y_i|p\right) \tag{7.10}$$

To calculate the likelihood function, take the assumed chance of observing (which is p for $y = 1$ and $(1-p)$ for $y = 0$) for each case and multiply them together. (The symbol Π works like the summation symbol, Σ, except it means multiply everything instead of adding everything). The likelihood function reduces to

$$L\left(p|y_1 \ldots y_N\right) = p^{\sum_{i=1}^{N} y_i}\left(1-p\right)^{N-\sum_{i=1}^{N} y_i} \tag{7.11}$$

Which also looks scarier than it is. It's just the assumed model's probability, p, to the power of the total number of observed 1s times $1-p$ to the power

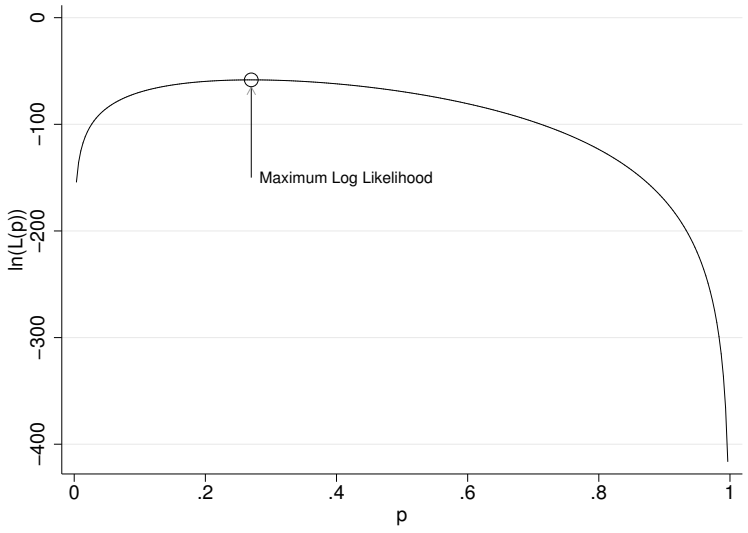

Figure 7.2: The likelihood function for random variable y with mean 0.27.

of the total number of observed 0s. Taking the natural log of this function gives us

$$\ln\left(L\left(p|y_1 \ldots y_N\right)\right) = \left(\sum_{i=1}^{N} y_i\right) \ln\left(p\right) + \left(N - \sum_{i=1}^{N} y_i\right) \ln\left(1-p\right) \quad (7.12)$$

Remember from our data that there are 27 yes answers and 100 - 27 = 73 no answers, so thus function is

$$\ln\left(L\left(p|y_1 \ldots y_N\right)\right) = (27) \ln\left(p\right) + (73) \ln\left(1-p\right)$$

Figure **??** tells a simple story: if one assumes a model where the chance of observing a 1 is the estimated mean, then the likelihood function is maximized. If any other model is assumed, the likelihood function decreases. This suggests that the optimum "model" for our data is the estimated mean. This may seem like we are going in circles, but when we have more complicated models, the idea of maximizing the likelihood of observing the data we have becomes powerful.

What process can be used to find the estimate that maximizes the likelihood? Again, the first step is to find the first derivative (which gives us an

equation for slope) of this function with respect to p:

$$\frac{d\ln\left(L\left(p|y_1 \ldots y_N\right)\right)}{dp} = \frac{\sum_{i=1}^{N} y_i}{p} - \frac{N - \sum_{i=1}^{N} y_i}{1-p} \tag{7.13}$$

The next step is to set the derivative of this function to 0 (remember, when the slope is 0, the function is either at a local minima or maxima) and solve for p

$$0 = \frac{\sum_{i=1}^{N} y_i}{p} - \frac{N - \sum_{i=1}^{N} y_i}{1-p}$$

$$0 = \left(\sum_{i=1}^{N} y_i\right)(1-p) - \left(N - \sum_{i=1}^{N} y_i\right)p$$

$$-\sum_{i=1}^{N} y_i = -pN$$

$$-\frac{\sum_{i=1}^{N} y_i}{N} = -p$$

$$\frac{\sum_{i=1}^{N} y_i}{N} = p \tag{7.14}$$

There we are, the formula that maximizes the likelihood is the mean, or proportion of 1s. In this case $p = 0.27$. It would be a simple matter to plug in the numbers (if they had been provided) to verify that $p = 0.27$.

Chapter 8

Ordinary least squares regression

8.1 A line through the data

Let's assume that a researcher wants to understand the relationship between two variables, x and y, using a geometric line. Recall that a simple two dimensional line has two parameters, and intercept and a slope. In these notes, I call the intercept β_0 and the slope β_1 to form the function

$$\hat{y}_i = \beta_0 + \beta_1 x_i. \tag{8.1}$$

Figure **??** is an example line. Note the hat above y again, this lets us know that this is a prediction, not the actual value of y. The value of $\hat{y}_i$ is a *conditional mean*. If we wanted to show an expression for the actual value of y for each case, we would need to show a residual, e, in the expression. This residual is the difference between the actual value of y and the predicted value of y:

$$e_i = y_i - \hat{y}_i = y_i - \beta_0 + \beta_1 x_i \tag{8.2}$$

It is also a deviation. Thus, the expression for the observed y that includes the elements of the line is

$$y_i = \beta_0 + \beta_1 x_i + e_i. \tag{8.3}$$

Figure **??** shows the residuals for each point on a scatterplot for the simulated data. Note how all the point below the regression line are negative, while all the points above the regression line are positive

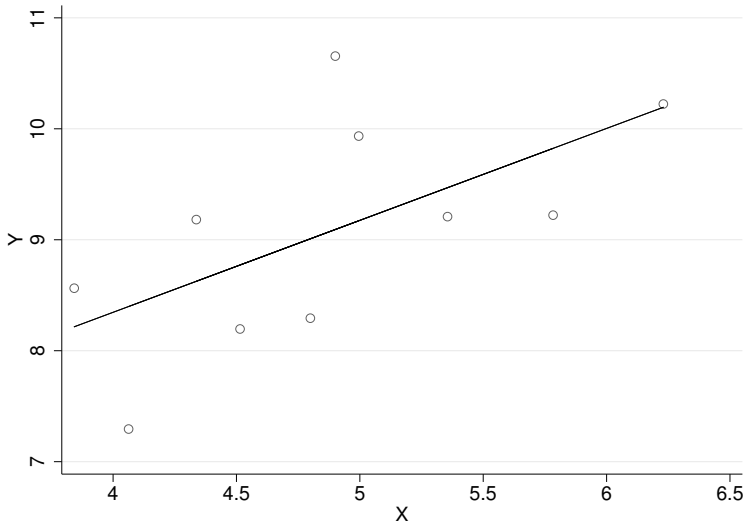

Figure 8.1: Data from Table **??** fit with $\hat{y}_i = \beta_0 + \beta_1 x_i$.

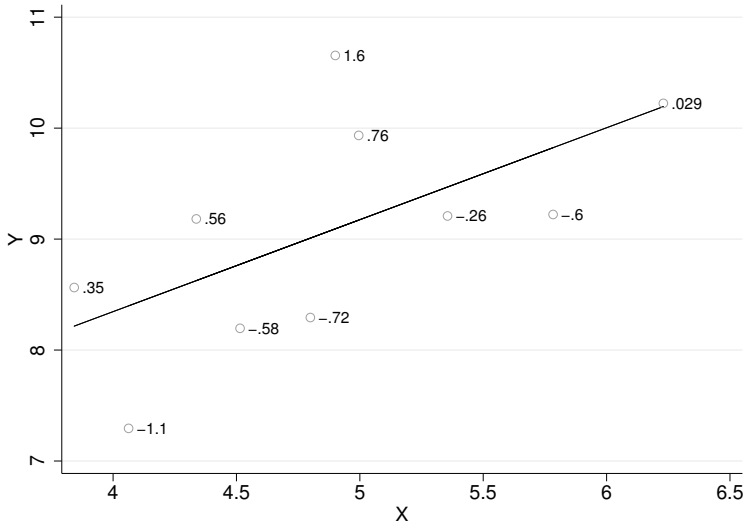

Figure 8.2: Data from Table **??** fit with $\hat{y}_i = \beta_0 + \beta_1 x_i$ marked with the value of e_i.

8.2 The regression slope

Remember that correlation removes all information about the units of x and y. Often, this is not useful to your average policy maker. They generally want to know, "if I put in one more unit of x, what happens to y?" In other words, if x increases by one unit, what is the change in y (and *in the units of* y).

We can do this with a regression slope. The formula for the regression slope is quite simple for bivaraite regression, and many parts should look familiar

$$\beta_1 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{8.4}$$

Look back at the formula for correlation. This is exactly the same with one change: we removed the sum of squares for y in the denominator. We interpret this statistic as the change in y for a single unit increase in x.

8.3 The intercept

In order to draw a line we also need a y intercept. The formula for this is also quite simple, With the slope in hand, we simply calculate the difference between the mean of y and the mean of x times the slope

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{8.5}$$

This statistic tells us the predicted value of y when x is equal to 0. This will be very important later, as we can change the implicit meaning behind $x = 0$.

8.4 Why is this the best fitting line?

Where did this formula come from? The question to motivate this is: what line best summarizes the data? To answer this question, we need to operationalize what it means to best summarize the data, just as we did for the mean.

8.4.1 Least squares formulation

One idea is that we should first quantify the difference between what we observe for y and what we predict for y. The deviation has previously been defined as e_i for each case. Now, how can we summarize these differences in a single number? To do this, we once again return to the sum of squares. We summarize the deviation of the model from the observed as the sum of the squared residuals error (SSE)

$$SSE = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{8.6}$$

When SSE is big, the model does a poor job predicting the data. When it's small, the model does a good job predicting the data. Intuitively, this should make sense. The closer the data are to the line, the smaller the SSE will be and the further the data are away from the line, the larger the SSE will be.

Now, how can we find the best line? We have a systematic way of finding the best line to fit the data. Here is the idea: we make SSE a function of the slope and intercept:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{N} e_i^2$$

First, we replace the residual with what it is, the difference between y and the regression line

$$S(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

Next, we expand

$$S(\beta_0, \beta_1) = \sum_{i=1}^{N} \left(y_i^2 - 2\beta_0 y_i - 2\beta_1 y_i x_i + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2\right)$$

and pull the summations out

$$S(\beta_0, \beta_1) = \sum_{i=1}^{N} y_i^2 - 2\beta_0 \sum_{i=1}^{N} y_i - 2\beta_1 \sum_{i=1}^{N} y_i x_i + N\beta_0^2 + 2\beta_0\beta_1 \sum_{i=1}^{N} x_i + \beta_1^2 \sum_{i=1}^{N} x_i^2$$

We then find the partial derivative of β_0 (one of the two normal equations)

$$\frac{\partial S}{\partial \beta_0} = 2N\beta_0 - 2\sum_{i=1}^{N} y_i - 2\beta_1 \sum_{i=1}^{N} x_i$$

and putting the elements back into the sum

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i) \tag{8.7}$$

and the partial derivative of β_1 (the other normal equation)

$$\frac{\partial S}{\partial \beta_1} = 2\beta_0 \sum_{i=1}^{N} x_i + 2\beta_1 \sum_{i=1}^{N} -2\sum_{i=1}^{N} y_i x_i$$

and putting the elements back into the sum

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)\, x_i \tag{8.8}$$

We then set these both to 0, pull the constants out, and divide by N. That gives us for the intercept

$$0 = \bar{y} - \beta_0 - \beta_1 \bar{x}$$

and for the slope

$$0 = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \beta_0 \bar{x} - \beta_1 \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

Two equations, two unknowns (β_0 and β_1). To get β_1, we figure that β_0 is

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

we then substitute $\bar{y} - \beta_1 \bar{x}$ for β_0 in the second equation

$$\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \beta_0 \bar{x} - \beta_1 \frac{1}{N}\sum_{i=1}^{N} x_i^2 = 0$$

$$\frac{1}{N}\sum_{i=1}^{N} x_i y_i - (\bar{y} - \beta_1 \bar{x})\,\bar{x} - \beta_1 \frac{1}{N}\sum_{i=1}^{N} x_i^2 = 0$$

and solve for β_1

$$\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{y}\bar{x} + \beta_1 \bar{x}^2 - \beta_1 \frac{1}{N}\sum_{i=1}^{N} x_i^2 = 0$$

$$\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{y}\bar{x} = -\beta_1 \bar{x}^2 + \beta_1 \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

$$\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{y}\bar{x} = \beta_1 \left(-\bar{x}^2 + \frac{1}{N}\sum_{i=1}^{N} x_i^2 \right)$$

$$\frac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{y}\bar{x}}{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2} = \beta_1$$

which is the same as equation (**??**).

The data in Table **??** produced the regression in Table **??**. This is the best fit line. To demonstrate, I attempted to draw some other lines through the data, summed up the residuals, then plotted the SSE for each line against the slope of that line in Figure **??**.

Table 8.1: Regression of y on x from data in Table **??**

Coefficients	Model
x	0.829
	(0.384)
Intercept	5.032*
	(1.892)
SEs in parentheses, $*p < 0.05$	

What we find is that no matter what slope I use, I can't do better than the formulas provided here. Compare Figure **??** to Figure **??** and you will see how regression works off the same least squares principle.

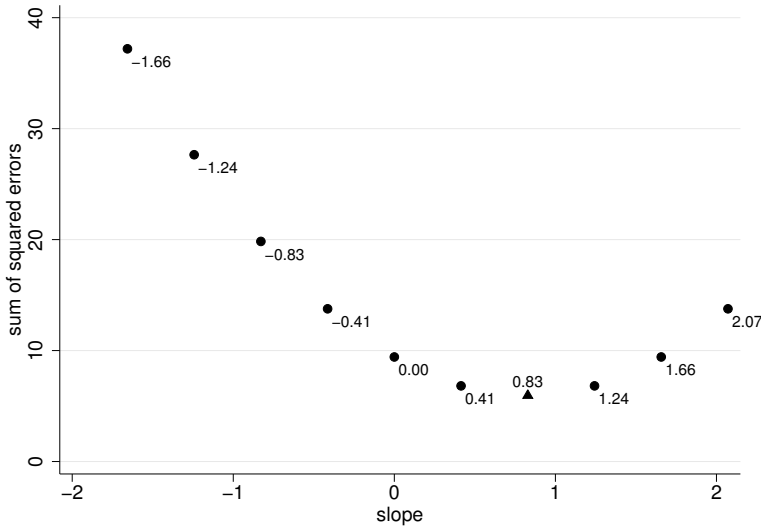

Figure 8.3: Alternate regression slopes and SSEs for data in Table **??**; real slope is 0.83.

8.4.2 Maximum likelihood formulation

Another way to get to these estimates is to consider the likelihood of observing the data given the parameters. Since the predicted value y is a conditional mean, we can start with the likelihood function of the mean

$$\ln\left(L\left(\mu|y_1 \ldots y_N\right)\right) = -\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \mu\right)^2 - \frac{N}{2}\left(\ln\left(\sqrt{2\pi}\right)\right) \tag{8.9}$$

and replace the mean with the regression formula

$$\ln\left(L\left(\beta_0, \beta_1|y_1 \ldots y_N\right)\right) = -\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \beta_1 x_i\right)^2 - \frac{N}{2}\left(\ln\left(\sqrt{2\pi}\right)\right) \tag{8.10}$$

and now we need to find the partial derivatives of this function, which look a lot like the least squares formulations. I won't do that all again, but hopefully you see that taking the derivatives of this function will produce the exact same estimators. Thus, in this case, the least squares estimators are also the maximum likelihood estimators.

However, the only reason this works is because we assume a normal distribution. That is one of the assumptions of least squares: a normally distributed set of residuals. If that is not the case, then the maximum likelihood estimator will be different altogether, and then the least squares estimators will no longer be valid.

Chapter 9

Multiple regression

Bivariate regression fits a line to a two-dimensional space. If we have a regression with two predictors we fit a plane to a three-dimensional space. For example,

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 z_i \tag{9.1}$$

is an equation for plane and provides a formula for points on a (y,x,z) coordinate system. By fitting a plane to the data, regression can independently estimate the effects of one variable, while holding the other constant. The best way to think about multiple regression is with matrix algebra

9.1 The matrix algebra formulation

Thus, one of the beauties of regression is that we can have more than one predictor. Using two variables for example, we could fit a plane to the variables x_1 and x_2:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \tag{9.2}$$

We could fit a hyperplane to three predictors, indexed as x_1, x_2, and x_3:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \tag{9.3}$$

and so on. The problem is the algebra for these solutions gets complicated fast. Instead, we have a matrix algebra equivalent of our least squares estimator

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{9.4}$$

Table 9.1: Small set of 4 cases

y	x
3	4
2	3
4	3
5	6

Matrix algebra is powerful. It deals with vectors (single rows or columns of data) and matrixes (data with multiple rows and columns). Statistics packages take the data in tables and convert them to matrixes. For instance, we can think of our outcome in Table **??** as a vector of numbers called $\mathbf{y}$, for example:

$$\mathbf{y} = \begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \end{bmatrix}$$

Here we have 4 cases ($N = 4$) where the first case of y is 3, the second case of y is 2, the third case of y is 4, and the fourth case of y is 5. We can then think of our predictors as a matrix with N rows and p columns

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 1 & 3 \\ 1 & 3 \\ 1 & 6 \end{bmatrix}$$

The first column is for the intercept. That's why some programs call it a constant, because it is for a "variable" that is a constant value of 1s. Thus, we can then think of our typical regression equation as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{9.5}$$

$$\begin{bmatrix} y_i \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p11} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & \cdots & x_{pN} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}$$

Note how easy it is to formulate any number of predictors, the regression equation always comes out to $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$. Now, the formula to obtain the slopes (or the b vector) is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}$$

This may not look familiar, but it is. For instance,

$$(\mathbf{X}'\mathbf{X}) = \sum_{i=1}^{N} (x_i - \bar{x})^2$$

which is the denominator of the slope formula, and since

$$a^{-1} = \frac{1}{a}$$

Also,

$$\mathbf{X}'\mathbf{y} = \sum_{i=1}^{N} (y_i - \bar{y})\,(x_i - \bar{x})$$

which is the numerator of the slope formula.

Matrix algebra has many rules, and I'm not going to go through all of them here. However, here is an example calculation with the small set of 4 in Table **??**. First, we need to transpose $\mathbf{X}$ to get $\mathbf{X}'$ (also known as $\mathbf{X}$ prime or the transpose of $\mathbf{X}$):

$$\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 4 & 3 & 3 & 6 \end{bmatrix}$$

Note that the transpose of $\mathbf{X}$ simply makes rows out of columns. Column 1 becomes row 1, column 2 becomes row 2, and so on. $\mathbf{X}'\mathbf{X}$ is multiplying $\mathbf{X}'$ and $\mathbf{X}$ (note: in matrix algebra the order of multiplication matter a lot). So, if

$$\mathbf{X}' = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix}$$

and

$$\mathbf{X} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \end{bmatrix}$$

then

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} (\sum_i a_{1i}b_{i1}) & (\sum_i a_{1i}b_{i2}) \\ (\sum_i a_{2i}b_{i1}) & (\sum_i a_{2i}b_{i2}) \end{bmatrix} \tag{9.6}$$

This means that we can compute X'X like so

$$\sum_i a_{1i}b_{i1} = ((1 \times 1) + ((1 \times 1) + ((1 \times 1) + ((1 \times 1))$$

$$\sum_i a_{1i}b_{i2} = ((4 \times 1) + ((3 \times 1) + ((3 \times 1) + ((6 \times 1))$$

$$\sum_i a_{2i}b_{i1} = ((1 \times 4) + ((1 \times 3) + ((1 \times 3) + ((1 \times 6))$$

$$\sum_i a_{2i}b_{i2} = ((4 \times 4) + ((3 \times 3) + ((3 \times 3) + ((6 \times 6))$$

which means

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 16 \\ 16 & 70 \end{bmatrix}$$

Next, we tackle $\mathbf{X}'\mathbf{y}$. If

$$\mathbf{y} = \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \\ b_{14} \end{bmatrix}$$

then

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} (\sum_i a_{1i}b_{i1}) \\ (\sum_i a_{2i}b_{i1}) \end{bmatrix} \tag{9.7}$$

which translates into

$$\sum_i a_{1i}b_{i1} = (1 \times 3) + (1 \times 2) + (1 \times 4) + (1 \times 5)$$

$$\sum_i a_{2i}b_{i1} = (4 \times 3) + (3 \times 2) + (3 \times 4) + (6 \times 5)$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 14 \\ 16 \end{bmatrix}$$

Now, we need to invert $\mathbf{X}'\mathbf{X}$. So, if

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

and

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \left(\frac{d}{ad-bc}\right) & \left(\frac{-b}{ad-bc}\right) \\ \left(\frac{-c}{ad-bc}\right) & \left(\frac{a}{ad-bc}\right) \end{bmatrix} \tag{9.8}$$

and also if

$$ad - bc = 24$$

then

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 2.92 & -0.67 \\ -0.67 & 0.16 \end{bmatrix}$$

Bring it all together:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} = \begin{bmatrix} (\sum_i a_{1i}b_{i1}) \\ (\sum_i a_{2i}b_{i1}) \end{bmatrix}$$

Which means we are left with

$$\mathbf{b} = \begin{bmatrix} 0.83 \\ 0.66 \end{bmatrix}$$

where 0.83 is our intercept, β_0, and 0.66 is our slope β_1. There we are. It is not important that you be able to do this, but I do want you to remember the formula $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$ because we are going to alter it from time to time for different estimators.

9.1.1 The matrix formulation of least squares

We can still show that these are least squares estimates. In matrix form, the sum of squares is $\mathbf{e}'\mathbf{e}$, so we can write the matrix of slopes as a function of the sum of squares

$$S(\mathbf{b}) = \mathbf{e}'\mathbf{e}$$

$$S(\mathbf{b}) = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$$

$$S(\mathbf{b}) = \mathbf{y}'\mathbf{y} - (\mathbf{2y}'\mathbf{X})\mathbf{b} + \mathbf{b}'(\mathbf{X}'\mathbf{X}')\mathbf{b} \tag{9.9}$$

We then again find the derivative

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{2X}'\mathbf{Xb} - \mathbf{2X}'\mathbf{y} \tag{9.10}$$

set the derivative to 0 and solve

$$0 = \mathbf{2X}'\mathbf{Xb} - \mathbf{2X}'\mathbf{y}$$

$$-\mathbf{2X}'\mathbf{Xb} = -\mathbf{2X}'\mathbf{y}$$

$$\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$$

the result matches equation (**??**).

9.1.2 The matrix formulation of maximum likelihood

We can also get to this estimate through maximum likelihood. If the mean is equal to the slopes and predictors combination, **Xb**, then the likelihood of observing our data from our parameters is

$$\ln(L(\mathbf{b}, \mathbf{V}|\mathbf{y})) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\mathbf{V}) - \frac{1}{2}(\mathbf{y} - \mathbf{Xb}))'\mathbf{V}^{-\mathbf{1}}(\mathbf{y} - \mathbf{Xb}) \tag{9.11}$$

where **V** is the variance-covariance matrix of residuals. After some work, we can find the partial derivative of this function with respect to the slopes, **b** as

$$\frac{\partial \ln(L(\mathbf{b}, \mathbf{V}|\mathbf{y}))}{\partial \mathbf{b}} = -\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{Xb} + \mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{y} \tag{9.12}$$

we then set to 0 and solve

$$0 = -\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{Xb} + \mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{y}$$

$$\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{Xb} = \mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{y}$$

$$\mathbf{b} = \left(\mathbf{X'V^{-1}X}\right)^{-1}\mathbf{X'V^{-1}y}$$

As we see in Chapter **??**, a very important set of assumptions in OLS regression is that the errors have 0 correlation and constant variance. this allows $\mathbf{V}$ to be a simple matrix where the diagonals are all σ^2 and every other number in the matrix is 0. Thus, $\mathbf{V}$ drops out and we are left with equation (**??**),

$$\mathbf{b} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'y}$$

The OLS estimator.

9.2 Interpretation

The general regression model is expressed algebracially like this

$$\hat{y}_i = \beta_0 + \sum_p \beta_p x_{ip} \tag{9.13}$$

where $\sum_p \beta_p x_{ip}$ is my shorthand for each p predictor and its associated slope. We interpret the intercept, β_0, as the predicted value of y when all predictors are equal to 0. Later we will do many neat things to control what the intercept means substantively by messing with what x_p means.

Each slope for any predictor is then interpreted as the change in the predicted value of y, in the units of y if we increase by one unit. Again, we can control what means substantively, and we will go over that later.

9.3 Standardized coefficients

One question that is often important to researchers is whether one variable is more important than another in predicting the outcome.

We can do this with the standardized regression coefficient. This is a scaled slope that takes into account the ratio of the standard deviation of the predictor to the standard deviation of the outcome

$$\beta_p^* = \beta_p \frac{s_p}{s_y} \tag{9.14}$$

This procedure makes comparisons between predictors possible because its interpretation is changed from the change in y for a one unit increase in x (in the units of x) to the *standard deviation unit* change in y for a *standard deviation increase* in x.

Be careful in doing this with dichotomous predictors.

9.3.1 Standardized coefficient example

Table **??** details the summary statistics for 200 observations from a dataset that collections wage, education, age and gender.

Table 9.2: Summary Statistics for Wage Model

	mean	sd	min	max
wage	16.03	8.32	3.13	45.36
edu	13.48	2.91	5	20
age	36.73	12.31	16	64

Table 9.3: Model predicting wages by education and age

Coefficients	Model
edu	0.882***
	(0.184)
age	0.218***
	(0.043)
Intercept	-3.866
	(3.061)
Model Statistics	
N	200.000
F	22.947
R^2	0.189
df Regression	2.000
Sum of Squares Regression	2603.472
df Error	197.000
Sum of Squares Error	11175.277
SEs in parentheses, $***p < 0.001$	

Table **??** presents the regression model for wages as a function of education (edu) and age. We can calculate the standardized coefficients using the information in Tables **??** and **??**. For example, the standardized coefficient for edu is

$$\beta^*_{edu} = \beta_{edu}\frac{s_{edu}}{s_{wage}} = 0.882\frac{2.91}{8.32} = 0.308$$

and for age is

$$\beta^*_{age} = \beta_{age}\frac{s_{age}}{s_{wage}} = 0.218\frac{12.31}{8.32} = 0.322$$

You will notice that even though the coefficient for education was larger, the standardized coefficient for age was larger because age had a larger standard deviation.

9.4 Introduction to coding and interpreting nominal predictors

So far we have only worked with regression predictors that are continuous in nature. However, in social science we often have nominal variables of interest that we think would make good predictors of continuous outcomes (for nominal outcomes, see the logit and probit chapter). This is just a brief introduction, we will have a chapter devoted to other things you can do with nominal variables.

9.4.1 Two groups

We can consider a variable dichotomous when it takes on only two values, 0 and 1. This can indicate a trait of some sort, or may simply be a way to organize a variable with two values.

In the bivariate case, the slope, standard error, and t-test of some outcome y regressed on a dichotomous predictor d will give the same answer as the independent groups' t-test. That is to say that in the model

$$y_i = \beta_0 + \beta_1 d_i + e_i \tag{9.15}$$

the value of β_1 equals the difference between the average of y when $d = 1$ and $d = 0$

$$\beta_1 = (\bar{y}|d = 1) - (\bar{y}|d = 0) \tag{9.16}$$

Also, the standard error of β_1 is the same standard error from the independent test. As a result, the t-test and result are also the same.

9.4.2 More than two groups

We could use ANOVA to test generally whether these means are different. Table **??** told us with an F test of 16.87, and with (2,2588) degrees of freedom, that groups were different. Unfortunately, this doesn't tell us about who is different, just that at least one group is.

What we need is to enter dichotomous (or "dummy") variables into a regression model. This works by creating for each category in x (except one reference group) new variables, d_j where $j \in \{1 \dots k\}$ that is equal to 1 if that case is a member of that group, and 0 otherwise:

$$d_j = \begin{cases} 1 & \text{if } x = j; \\ 0 & \text{otherwise.} \end{cases} \tag{9.17}$$

Picking moderate as our reference group, and creating variables *lib* and *con*, we can fit the following model

$$\widehat{words}_i = \beta_0 + \beta_1 lib_i + \beta_2 con_i$$

This model is presented in Table **??**. Note that the F-test is the same as what was reported by the ANOVA model. A few things to note in this model. First, the intercept is the predicted value of y when all predictors are equal to 0. Thus, this is the predicted value for the reference group (moderate).

Compare the constant with the mean of the moderate group. It's the same. Now, also compare the effect of *lib*, it is the difference in the means of the liberal group and the moderates. The effect of *con* is the difference in the means of the conservative group and the moderates.

Sometimes researchers do not want to know the difference between a particular group and a reference group. Perhaps the research question is about the difference between a particular group and the population average, or specifically the average of group averages. In this case we use deviance coding. In deviance coding, instead of just a simple "1 for the group and 0 otherwise," we instead code each group as 1 for the group, -1 for the reference (m), and 0 otherwise:

$$d_j = \begin{cases} 1 & \text{if } x = j; \\ -1 & \text{if } x = m; \\ 0 & \text{otherwise.} \end{cases} \tag{9.18}$$

we then fit the model with this coding Now the intercept is an estimate

Table 9.4: Model predicting words correct by political affiliation

Coefficients	Model
lib	0.513***
	(0.097)
con	0.422***
	(0.093)
Intercept	5.812***
	(0.064)
Model Statistics	
N	2591.000
F	16.869
R^2	0.013
df Regression	2.000
Sum of Squares Regression	134.432
df Error	2588.000
Sum of Squares Error	10311.878
SEs in parentheses, $***p < 0.001$	

Table 9.5: Model predicting words correct by political affiliation–deviance coding

Coefficients	Model
lib	0.201***
	(0.058)
con	0.111*
	(0.055)
Intercept	6.124***
	(0.040)
Model Statistics	
N	2591.000
F	16.869
R^2	0.013
df Regression	2.000
Sum of Squares Regression	134.432
df Error	2588.000
Sum of Squares Error	10311.878
SEs in parentheses, $***p < 0.001$	

of the average of group averages $((6.32 + 5.81 + 6.23)/3) = 6.12$, and each slope is the difference between that group and the average of group averages. For example, we see a difference of 0.20 words between liberals and and the average of group averages. Note, also, that all the model statistics are the same.

Table 9.6: Model predicting words correct by political affiliation–linear contrasts

Coefficients	Model
$constrast_1$	-0.312***
	(0.054)
$constrast_2$	0.045
	(0.050)
Intercept	6.124***
	(0.040)
Model Statistics	
N	2591.000
F	16.869
R^2	0.013
df Regression	2.000
Sum of Squares Regression	134.432
df Error	2588.000
Sum of Squares Error	10311.878
SEs in parentheses, $***p < 0.001$	

A third method of coding variables is linear contrasts. Suppose, after our investigation, that we wanted to test other hypotheses such as how two groups compared to another. Suppose we wanted to test the null hypotheses that moderates were less smart than the average of liberals and conservatives. Such a hypothesis would be

$$H_0 : \mu_{mod} = \frac{\mu_{lib} + \mu_{con}}{2}$$

which can be rewritten as

$$H_0 : 0 = 1 \times \mu_{mod} - 0.5 \times \mu_{lib} - 0.5 \times \mu_{con}$$

A second hypothesis would test specifically whether liberals could be smarter

than conservatives, ignoring moderates.

$$H_0 : \mu_{lib} = \mu_{con}$$

which can be rewritten as

$$H_0 : 0 = 0 \times \mu_{mod} + 1 \times \mu_{lib} - 1 \times \mu_{con}$$

and all could be symbolized generally with contrast coefficients

$$H_0 : 0 = c_1 \times \mu_{mod} + c_2 \times \mu_{lib} + c_3 \times \mu_{con}$$

where c_1, c_2, and c_3, are the contrast coefficients. We create new variables for each set of contrast coefficients. For example, we create a new variable $contrast_1$ that is equal to 1 (c_1) for moderates, -0.5 for both liberals (c_2) and conservatives (c_3). We then create another variable, $contrast_2$ that is equal to 0 (c_1) for moderates, 1 for liberals (c_2), and

$$\widehat{words}_i = \beta_0 + \beta_1 contrast_1 + \beta_2 contrast_2$$

The results of this model are in Table **??**. The effect of the first contrast shows that the moderate average is less than the average of liberals and conservatives. We see that the first contrast is statistically signifiant. However, the second contrast is not: liberals are not smarter than conservatives. Again, the model fit statistics are the same. For any set of contrasts, there are three sets of rules:

1. The coefficients must add to 0 for each contrast
2. The sum of the products of all contrasts must equal 0
3. You must have m-1 contrasts, where m is the number of categories

Chapter 10

OLS model inference and evaluation

So now we know how to get slopes and the intercept. Now we need to get a handle on whether we can conclude anything substantive about them. There are two aspects to model evaluation: hypothesis tests of coefficients and evaluating model fit. These topics are of course connected and we see that when we test blocks of coefficients.

Let us consider, again, a model where wages are a function of education and other variables. In Tables **??** and **??** I referenced a dataset of 200 observations with three variables: hourly wages ("wage"), years of education ("edu"), and age. Table **??** displays three different models. The first model is simply

$$\hat{y}_i = \beta_0 + \beta_1 edu_i \tag{10.1}$$

the second model adds age to the model

$$\hat{y}_i = \beta_0 + \beta_1 edu_i + \beta_2 age_i \tag{10.2}$$

and the final model adds a dichotomous female indicator

$$\hat{y}_i = \beta_0 + \beta_1 edu_i + \beta_2 age_i + \beta_3 female_i \tag{10.3}$$

Table **??** gives a lot information. First we get the slopes and intercept on the upper table. Second, we get model fit statistics on the lower table. In this section, I describe these measures.

Table 10.1: Models predicting wages

Coefficients	Model 1	Model 2	Model 3
edu	0.835***	0.882***	0.847***
	(0.195)	(0.184)	(0.184)
age		0.218***	0.229***
		(0.043)	(0.044)
female			-2.033
			(1.074)
intercept	4.780	-3.866	-2.817
	(2.682)	(3.061)	(3.091)
Model Statistics			
N	200.000	200.000	200.000
F	18.410	22.947	16.692
R^2	0.085	0.189	0.204
df Regression	1.000	2.000	3.000
Sum of Squares Regression	1172.179	2603.472	2804.011
df Error	198.000	197.000	196.000
Sum of Squares Error	12606.570	11175.277	10974.738
SEs in parentheses, $***p < 0.001$			

10.1 The variance and covariance of regression coefficients

In statistics, we are obsessed with the variance of our estimators. The variance of an estimator is just the square of the standard error. Thus, once we know the variance of the estimator, we can just get the square root and use that for hypothesis testing. In matrix notation, the variance of the regression slopes is

$$\mathbf{V}\left(\hat{\mathbf{b}}\right) = \left(\frac{\mathbf{1}}{\mathbf{N}}\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\sigma^{\mathbf{2}} \tag{10.4}$$

where σ^2 is the variance of the regression residuals, or

$$\sigma^2 = \frac{\sum_{i=1}^{N}\left(e_i - \bar{e}\right)^2}{N-p} \tag{10.5}$$

and $\mathbf{X}'\mathbf{X}$ is the sum of squares of the matrix of predictors. We can express this with conventional algebra in the bivariate regression model as

$$\text{var}\left(\beta_1\right) = \frac{\sigma^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \tag{10.6}$$

and we can express the standard error of the slope as the square root of this

$$SE\left(\beta_1\right) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}} \tag{10.7}$$

Here we see that the variance, and thus the standard error, is a function of the variance in the residuals and the variance in the predictor. It's the variance in the predictor that is the big player here. The larger the variance in x, or the more x is "spread out," the smaller the standard error. This is related to this is the variance of predicted point, $\hat{y}_0$, where again the crucial ingredient is the variance of x

$$\text{var}\left(\hat{y}_0\right) = \left(\frac{\sigma^2}{N} + \frac{\left(x_0 - \bar{x}\right)^2\sigma^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}\right)\sigma^2 \tag{10.8}$$

Here, $\hat{y}_0$ is the predicted value of y associated with a hypothetical value of x, x_0. We see that as the hypothetical value of x moves away from the mean

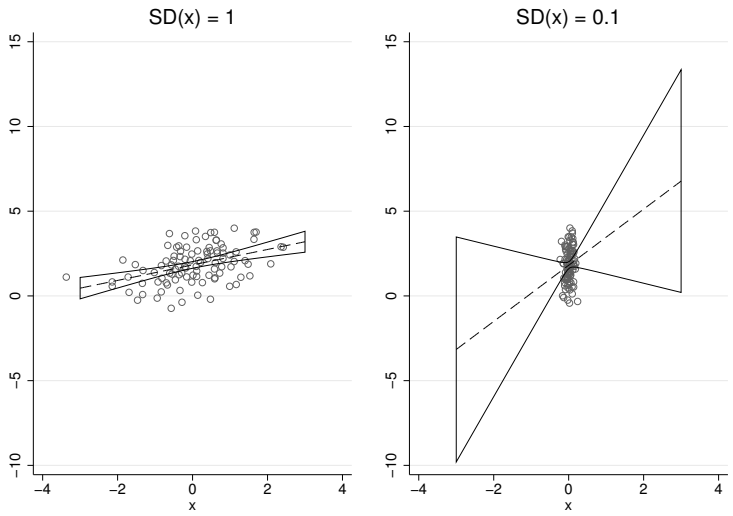

Figure 10.1: Variance of regression slopes as a function of the variance of x

of x, that the variance will increase. This formula allows us to construct a confidence interval for each prediction

$$CI\left(\hat{y}_0, (1-\alpha)\times 100\right) = \hat{y}_0 \pm t_{df,\alpha/2}\sqrt{\left(\frac{\sigma^2}{N} + \frac{(x_0-\bar{x})^2\,\sigma^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)\sigma^2} \quad (10.9)$$

Thus, the greater the variance in x, the smaller the confidence interval in any predicted value of y. A smaller confidence interval means that the range of possible slopes for y is smaller. Figure **??** features this difference. The panel on the left is a situation in which x has the larger standard deviation. We see that the slope is accurate and the 95 percent confidence interval (the fan like shape around the prediction line) of the prediction is pretty tight, implying a small standard error of the slope. The panel on the right, however, x has a small standard deviation and we see that the range of possible slopes is large because the 95 percent confidence interval of values is quite high.

10.2 Hypothesis testing with regression slopes

Recall our hypothesis testing for the difference between means. We were testing the difference between two mean values of y, $\bar{y}_A$ and $\bar{y}_B$. Testing

regression slopes is quite similar. Here we are testing the difference between a predicted value of y when x is equal to some arbitrary value, A and the predicted value of y when that value of x increases by a single unit to B. In other words, we are comparing $\hat{y}_A$ and $\hat{y}_B$.

Obviously, since regression slopes reflect the difference in the predicted value of y reflected in a one unit change in x, our regression slope reflects this difference. Here the null hypothesis is that this slope is 0, or that there is no relationship

$$H_0 : \beta_p = 0 \tag{10.10}$$

and the alternative hypothesis is that this slope is something other than 0

$$H_1 : \beta_p \neq 0 \tag{10.11}$$

Note that I use the term β_p here because this applies for any coefficient, including the intercept. With the null hypothesis equal to 0, the form of our t-test is the slope divided by the standard error; c.f., (**??**)

$$t = \frac{\beta_p - 0}{SE\left(\beta_p\right)} = \frac{\beta_p}{SE\left(\beta_p\right)} \tag{10.12}$$

This t-test has $N - p - 1$ degrees of freedom, where p is the number of predictors and we subtract 1 for the intercept as well. Returning to our education and wages example in Table **??** (Model 1), we see that for every year of education, the average wages increase by about 83.5 cents. The standard error of this estimate is 0.195. The t-test for this coefficient is

$$t = \frac{0.835}{0.195} = 4.28$$

The *df* Error in the table tells us our degrees of freedom for this test. Using a computer or table, we can find that the probability of finding this slope if the null hypothesis were true is 0.00003, far below the $\alpha = 0.05$ threshold. Therefore, we say the effect is statistically significant. Most computer programs will tell you this probability, known as a p-value. Thus, we think that the slope is significantly different from 0 and thus we make the conclusion that an effect of education on wages is plausible.

We also test whether the intercept is equal to 0:

$$t = \frac{4.780}{2.682} = 1.78$$

In this case, since the p-value is 0.077, and we cannot reject the null hypothesis, and we have to conclude that if someone with 0 years of education, there wages are practically 0, which of course makes intuitive sense.

Finally, since the intercept and slope are jointly estimated, we also estimate a covariance of these parameters

$$\operatorname{cov}(\beta_0, \beta_1) = \sigma^2 \left(\frac{-\bar{x}}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right) \tag{10.13}$$

which is actually estimated, along with the variances of the slopes themselves, in a matrix form as

$$\mathbf{V}\left(\hat{\mathbf{b}}\right) = \sigma^{\mathbf{2}} \left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}} \tag{10.14}$$

which in in this case is

$$\mathbf{V}\left(\hat{\mathbf{b}}\right) = \begin{bmatrix} \operatorname{var}(\beta_{edu}) & \operatorname{cov}(\beta_{edu}, \beta_{intercept}) \\ \operatorname{cov}(\beta_{intercept}, \beta_{edu}) & \operatorname{var}(\beta_{intercept}) \end{bmatrix} = \begin{bmatrix} 0.038 & -0.510 \\ -0.510 & 7.193 \end{bmatrix}$$

Note that $\operatorname{var}(\beta_{intercept})$ is the square of the standard error for the intercept and that $\operatorname{var}(\beta_{edu})$ is the square of the standard error for the slope. This doesn't have an immediate use in bi-variate regression, but in multiple regression it is a useful quantity for accurate testing differences between coefficients of predictors with formulas like

$$t = \frac{\beta_p - \beta_q}{\sqrt{\operatorname{var}(\beta_p) + \operatorname{var}(\beta_q) - 2\operatorname{cov}(\beta_p, \beta_q)}} \tag{10.15}$$

10.3 Model fit by way of ANOVA

Folks often ignore the ANOVA table in regression output because they do not know why it's there. Here is a cool fact for bivariate regression: if you square the t-test of the slope, you find that it is equal to the F-test in the output. Least squares regression and ANOVA are very much related. Recall how in ANOVA we consider how the total sum of squares is decomposed into the sum of squares between groups and the sum of squares within groups

$$SST = SSW + SSB \tag{10.16}$$

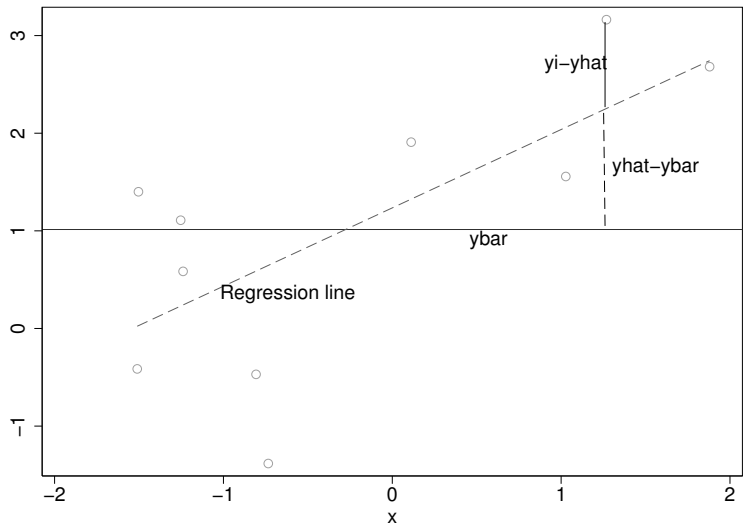

Figure 10.2: The deviance of y_i to $\hat{y}_i$ to $\bar{y}$

or

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2 + \sum_{j=1}^{k} n_j\,(\bar{y}_j-\bar{y})^2 \tag{10.17}$$

The analogue in regression is that instead of group means, $\bar{y}_j$, we have predicted values of y for a given value of x, $\hat{y}$. Instead of the sum of squares within groups (SSW), we have the sum of squares error (SSE), which is how each case deviates from it's predicted value. Instead of the sub of squares between (SSB), we have the sum of squares regression (SSR), which is how the regression prediction deviates from the overall mean of y,

$$SST = SSE + SSR \tag{10.18}$$

or

$$\sum_{i=1}^{N}(y_i-\bar{y})^2 = \sum_{i=1}^{N}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2 \tag{10.19}$$

This is visualized in Figure **??**, which is data from Table **??**. The solid line on the scatter plot is the overall mean of y, the dotted line with the slope is

the regression fit line. We then take a single point and draw a solid vertical line from that point and the regression line. This represents $y_i - \hat{y}_i$. We then draw a dotted line from the predicted value of the regression line to the overall mean. This represents $\hat{y}_i - \bar{y}$. The total difference between the observation and and the overall mean, $y_i - \bar{y}$, is the sum of these quantities.

With these quantities we can do an F-test to test whether the sum of squares residual is less than the sum of squares total

$$F = \frac{MSR}{MSE} \tag{10.20}$$

where

$$MSR = \frac{SSR}{p} \tag{10.21}$$

and

$$MSE = \frac{SSE}{N - p - 1} \tag{10.22}$$

This test is then evaluated against the F distribution with $(p, N - p - 1)$ degrees of freedom. The degrees of freedom for the sum of squares error is $N - p - 1$, just like the t-test, and the degrees of freedom for the sum of squares regression is p, again where p is the number of predictors excluding the intercept.

Examining the fit statistics from our wage and education example, Model 1 in Table **??**, the sum of squares from the model is 1,172.179 and the sum of squares error is 12,606.570. We then see the degrees of freedom (df) for each quantity. We can test the fit of the model with the F-test,

$$F = \frac{\left(\frac{1172.179}{1}\right)}{\left(\frac{12606.570}{198}\right)} = 18.410$$

which is associated with a probability (p-value) of 0.00003. This should be familiar, that's because the unrounded t-test was 4.29, the square of which is 18.410. This means that our model contributes something to explaining the variation in wages.

10.4 Model fit by way of R^2

Another statistic reported in our regression output is R^2. Substantively, this is a measure of how much variation is "explained" in the data. The calculation for R^2 is simply the ratio of the sum of square regression to the sum of squares total ($SST = SSR + SSE$)

$$R^2 = \frac{SSR}{SSR + SSE} \tag{10.23}$$

Since it is a ratio and SSR is always smaller than SST, it ranges from 0 to 1, with better model fit as we approach 1. In the first model in Table **??**, education explains about 8.5 percent of the variation in wages. What is also interesting is that in bivariate regression R^2 is literally the square of the correlation coefficient. This means that education and wages have a correlation of $\sqrt{0.085} = 0.292$. In multiple regression, the meaning is a little bit vague, but can be thought of as the square of the correlation of all predictors and the outcome.

The adjusted R-square is a somewhat different formula that includes a penalty for adding several variables that are not correlated with the outcome. This is sometimes useful since R-squares will always increase with new variables and so we can fool ourselves with large R-squares by putting a lot of junk in the model. The formula is

$$R^2_{adjusted} = 1 - \left(\left(1 - R^2\right) \frac{N-1}{N-p-1} \right) \tag{10.24}$$

where p is the number of predictors.

10.5 Testing blocks of coefficients

The t-tests of coefficients test each coefficient individually. This test is a test of whether that parameter is equal to 0. Another way to think of it is that it is a test of whether that variable adds to the explanatory power of the model compared to a model without that variable. This test compares the sum of squares error in the full model (or *unrestricted*) to the sum of squares error in the *restricted* model. For a single variable, this ratio is an F-test with $(1, N - p_U - 1)$ degrees of freedom, for more than 1 variable, the degrees of freedom are $(J_R, N - p_U - 1)$, where J_R is the number of variables being

tested and p_U is the number of predictors (including the intercept) in the full model. The test is

$$F = \frac{\left(\frac{SSE_R - SSE_U}{J_R}\right)}{\left(\frac{SSE_U}{N - p_U - 1}\right)} \tag{10.25}$$

For example, Model 2 in Table **??** adds age as a predictor. The t-test indicates it is a significant predictor according to the three stars ($***$), but we can also find out with an F test. Looking at the model statistics, we can call Model 1 the restricted model, so SSE_R is 12606.570, and Model 2 the unrestricted model, so SSR_U is 11175.277. Since the model adds a single variable, $J_R = 1$, and $N - 3 - 1 = 200 - 2 - 1 = 196$. Thus, our test is

$$F = \frac{\left(\frac{12606.570 - 11175.277}{1}\right)}{\left(\frac{11175.277}{196}\right)}$$

$$F = \frac{1431.293}{57.017}$$

$$F = 25.103$$

With $(1, 196)$ degrees of freedom, the probability of observing this test is 0.00001, which is very significant. We can check our work by estimating the t-test on the age effect:

$$t = \frac{0.218}{0.043}$$

$$t = 5.070$$

the square of which is almost the same number as the F test (differences due to rounding).

Model 3 in Table **??** adds the effect of gender. However, this effect is not significant.

Looks like gender ($female$) isn't significant, but does it add to the explanatory power of the model?

10.6 Model Likelihood

Since the least squares estimator is the same as the maximum likelihood estimator for normally distributed outcomes, OLS models have log-likelihood

functions that get maximized, see section **??**. Computers generally store the model likelihood along with other model statistics. With these we can perform likelihood ratio tests that compare different models on the same data. The ratio-test is

$$\chi^2 = 2\left(\ln\left(L\left(\theta_a\right)\right) - \ln\left(L\left(\theta_{null}\right)\right)\right) \tag{10.26}$$

when comparing the results of two likelihood functions (the value is evaluated on the χ^2 distribution of a single degree of freedom). There is an example of this test in **??**.

10.7 Important assumptions of OLS regression

The formulas discussed in the previous sections make several assumptions. You'll find that a lot of statistical theory starts with "assume..." then "we can think of this relationship as..." Assumptions form the basis of any statistical analysis. Often, the new methods that come out are a result of data breaking some assumption so a new method needs to be created. In fact, most of these notes are about what to do when you break these assumptions. Thus, it is important to understand these assumptions.

10.7.1 y is a linear function of the predictors

The first assumption is important for interpretation. We assume that y is a linear function function of the predictors. When this is true, graphs tend to look like Figure **??**. In many cases relationships are not linear and if you do not transform the data the estimated relationships are not valid. A classic violation of this is a quadratic relationship. Figure **??** displays this situation. When the relationship is quadratic and you fit a linear model without transformations, the slope may not reflect the data. In this case, the fitted slope is close to 0 even though it is obvious there is a relationship in the data. Unfortunately, there is no easy way to check this assumption outside of a examining the data. This is why we have theory. The best regression diagnostic is theory. Your model should be governed by theory. Trust your theory.

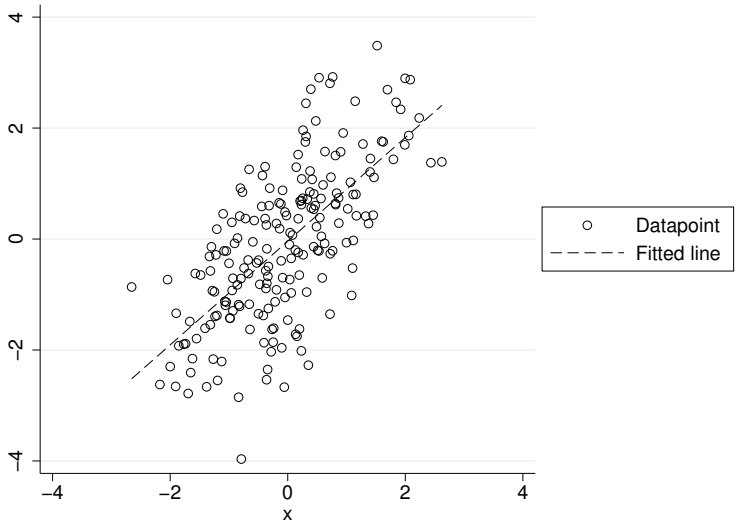

Figure 10.3: Data where y is a linear function of x

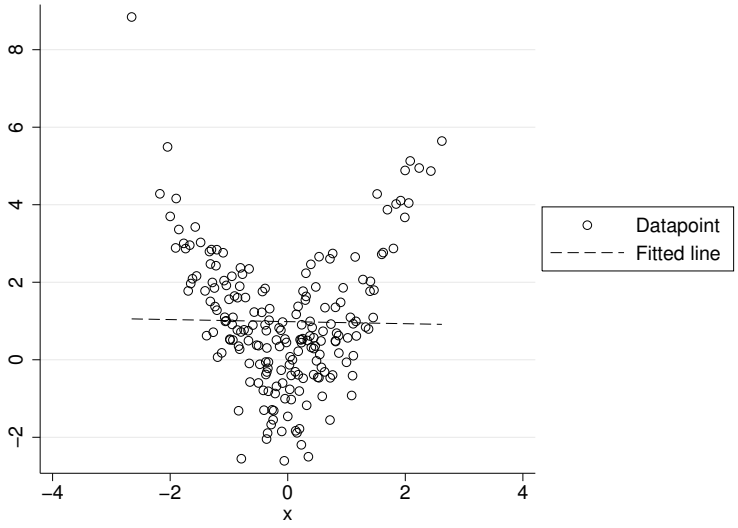

Figure 10.4: Data where y is not a linear function of x

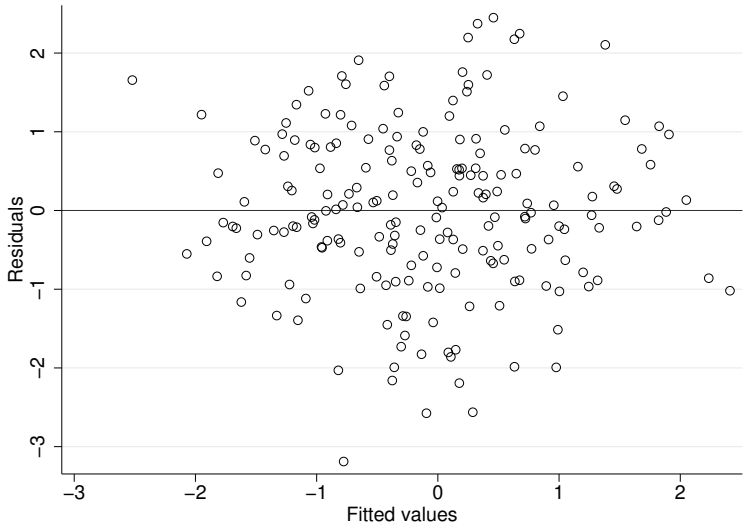

Figure 10.5: Data where e is zero on average

10.7.2 The expected value of any residual is zero

Related to the previous assumption is that we assume that the expected value of a residual is 0. "Expected" is fancy statistics language for "average." Thus, we expect that the residuals to be 0, on average, for any level of predictor. The math forces this to be true overall. However, I find more nuanced way of thinking of this assumption is to assume that for every value of x, the expected value of the residuals is 0. If we have a multivariate model, then instead of x we can use the fitted value of y. The reason we can use a fitted value of y is that the fitted value of y is just a linear combination of all the predictors. We can inspect this visually by making a scatterplot of the residuals by the fitted values, see Figure **??** where we meet the assumption that residuals average to 0. If, on the other hand, there is some pattern to the residuals, like in a un-modeled quadratic relationship, the plot may look like Figure **??**.

10.7.3 The variance of the residuals is constant

As we will see throughout the book, many of the formulas for standard errors and other measures make the assumption that there is a single variance for

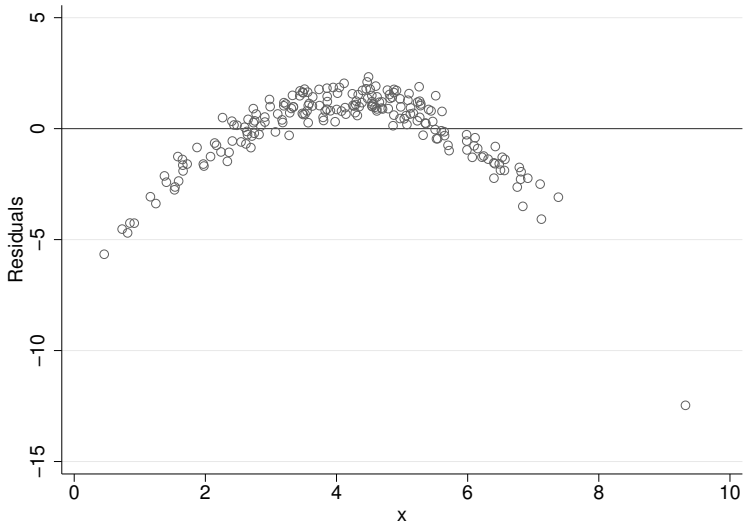

Figure 10.6: Data where e is not zero on average

all the residuals, σ^2.

However, in many situations this is not the case. When this assumption is violated, standard error become biased, making hypothesis tests difficult. Often this happens for practical reasons. For example, we will look at an example where charitable giving is a function of income. When income is low, there is little variance in giving because those who do not have a lot of money cannot give anything. At the other end, there is quite a bit of variance: you have people like Bill Gates who give a lot, and people like Steve Jobs who did not give much (at least publicly). When this is violated, you may see a "fan" pattern in your residuals like in Figure **??**

10.7.4 The covariance of all residuals is zero

One of the most commonly violated assumptions is that all the residuals are independent. We can express this as

$$\text{cov}\left(e_i, e_j\right) = 0 \tag{10.27}$$

There are several ways in which observations of y can be correlated:

1. Observations are adjacent to each other across time

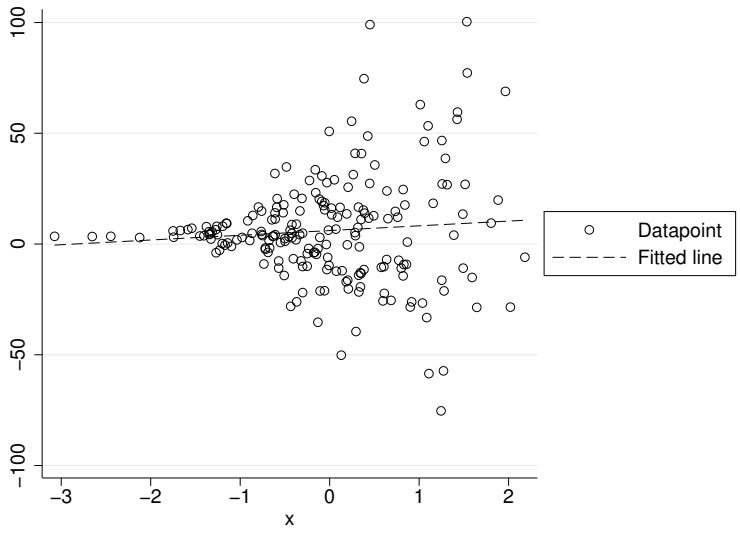

Figure 10.7: Data where e does not have constant variance

2. Observations are members of common geographical regions

3. Observations are members of other common meaningful clusters

4. Any combination of the above, and more

This is an old problem in social science and survey research. There are many solutions to this issue and we cover many of them in these notes.

10.7.5 The values of the predictors are not random

In normal regression we assume that the predictors, especially the intercept, are not random variables. When we cluster sample, that is sample groups then units within those groups, the intercept can be argued to be a random variable and thus violating this assumption. Another example is time within units. Many of the multilevel models are designed to compensate for this issue.

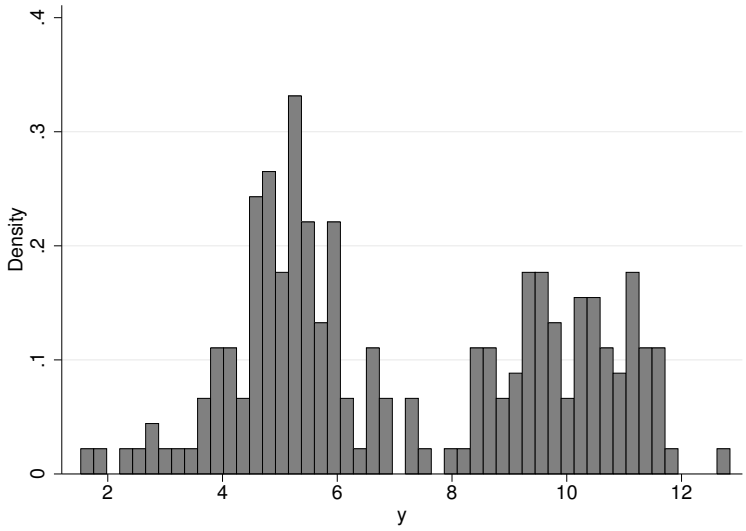

Figure 10.8: Data y as a function of a dummy variable

10.7.6 The values of the predictors are not exact linear combinations of each other

This assumption is pretty easy to confirm. If it is violated, and one variable is exactly correlated with another, it is impossible in invert . Many software packages with automatically remove offending variables from models.

10.7.7 The errors are distributed normally

Many students misunderstand this assumption to mean that the outcome needs to be distributed normally. This is not the case. The residuals, the outcome net of the model, must be normally distributed. For example, consider a model a continuous predictor x and another dichotomous predictor z. A histogram of the data appears in Figure **??** and you can see the bi-modal distribution. We see in Figure **??** that net of the model the errors are almost normally distributed. We need to make this assumption to prove that the least squares estimator is the maximum likelihood estimator.

However, sometimes there is no way to achieve normal errors. Situations in which the outcome is dichotomous or a count variable with a large proportion of zeros will never produce normally distributed outcomes. In those

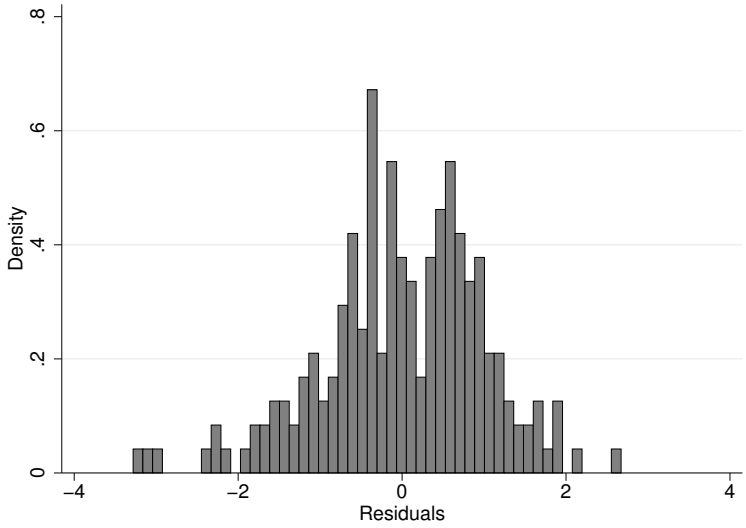

Figure 10.9: Residuals of a model for y that included a dummy variable

situations we need to use generalized linear models (GLMs).

Chapter 11

Colinearity

Colinearity among predictors is an interesting problem in regression. From one point of view, it is a scourge that reduces the precision of estimates and can make our coefficients unstable. From another point of view, it is the reason we perform multiple regression in the first place. The following offers an explicit example of the problem. Note, however, the actual issue is often less obvious.

I will first describe how it creates problem for precision, as it is an interesting discussion. From there, there really is not much to write about, since most "solutions" are pretty cheap and don't work that well.

11.1 Perfect colinearity

If two predictors in the model are perfectly correlated, they must be measuring the exact same thing. Therefore, it does not make sense to put the same information into the model twice. Don't worry about inadvertently making this mistake, as the **X'X** matrix will not be able to be inverted in this case and your software will produce an error.

11.2 Non-perfect colinearity

This leaves us with the more plausible issue. In any case, two variable in your model are likely to be correlated *at least a little bit*. This follows simple logic that if x and z are correlated with y, then there must be at least some relationship between x and z, even if it is just a little bit.

Again, there is nothing wrong with this, since this is why we want to perform multiple regression in the first place. However, there are limits to how much correlation we can have among our predictors.

11.2.1 The variance inflation factor

Recall from section **??** that in the bivariate case the variance of the regression coefficient is equation (**??**)

$$\text{var}(\beta_1) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

this can be extended to the multiple regression case for slope j of predictor j by adding the variance inflation factor (VIF), $\frac{1}{1-R_j^2}$, to the formula

$$\text{var}(\beta_j) = \frac{1}{1-R_j^2}\frac{\sigma^2}{(N-1)s_j^2} \tag{11.1}$$

The VIF is literally what it sounds like, its the extent to which (or factor of) the variance of a predictor is increased, given the multiple correlation between it and the other predictors.

Here, R_j^2 is the key ingredient. R_j^2 is like any other R^2 statistic, except here it is how much of predictor j is explained by the *other predictors*. The larger R_j^2, the larger the variance (and standard error) of β_j.[1]

But how much is too much correlation among the predictors? Consider Figure **??**, where we plot the square root of the VIF against the value of R_j^2. Note that the standard error doubles when R_j^2 is 0.75. That is, our standard errors only double when three fourths of the variation in one variable is explained by the other predictors. After that, things get worse pretty fast.

11.2.2 Example with city data

The example that follows is pretty silly, but it highlights the issues that you can encounter with high levels of multicolinearity. Suppose the data in Table , at the end of this chapter. It's census data from the 1980s and

[1]Note, also that since s^2 is the variance, we can return to the sum of squares my multiplying the variance by $N-1$.

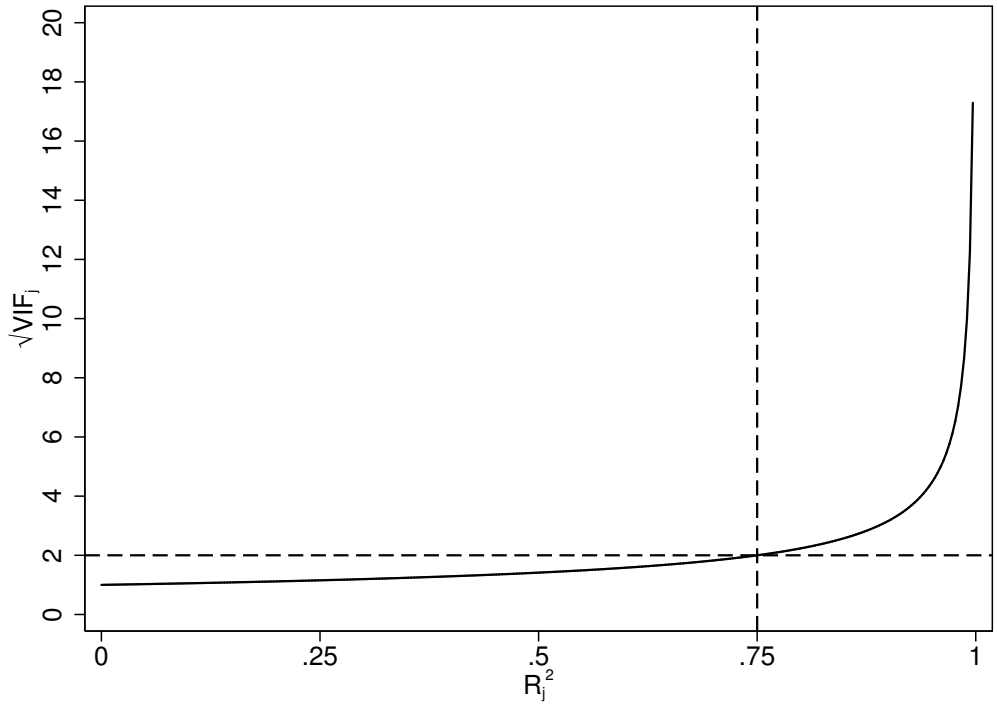

Figure 11.1: Relationship between the variance inflation factor and R_j^2

includes the average rent, average family income, median household income, and percent of the population with a college degree for various rural places.

Table **??** provides the correlations and Figure **??** is a scatter plot matrix of these variables. As you can see, and expectedly, mean family income and median household income are highly correlated.

Table 11.1: Correlations of variables in Table

	Mean rent	Mean family income	Median household income	Percent with BA
Mean rent	1.00			
Mean family income	0.63	1.00		
Median household income	0.66	0.96	1.00	
Percent with BA	0.58	0.81	0.72	1.00

An analysis of these data has a simple hypothesis, that higher the family income, holding education constant, the higher rent should be. Now, examine model 1 in Table **??**, mean family income has a negative (and significant) slope. This is the other thing that can happen with highly correlated predictors, the sign can flip.

This isn't to say that the model is *wrong*, but we have to take it literally: holding the median constant, if the mean increases (i.e. the data become more skewed), rent decreases.

This, however, is not a useful analysis. The better option is to select either the mean family income or the median household income as we do in Models 2 and 3.

The discussion of VIFs can inform the difference in the significance of the effect of education between Model 2 and 3. Notice how, in Table **??** that the percent with a college degree is more correlated with mean family income. That correlation is 0.8132, which translates into a VIF of

$$VIF = \frac{1}{1 - R_j^2} = \frac{1}{1 - 0.8132^2} = 2.95$$

making the effect of education not significant.

However, the correlation with the median income is lower, 0.7238, with a VIF of only 2.10, and the effect of education is now significant.

$$VIF = \frac{1}{1 - R_j^2} = \frac{1}{1 - 0.7238^2} = 2.10$$

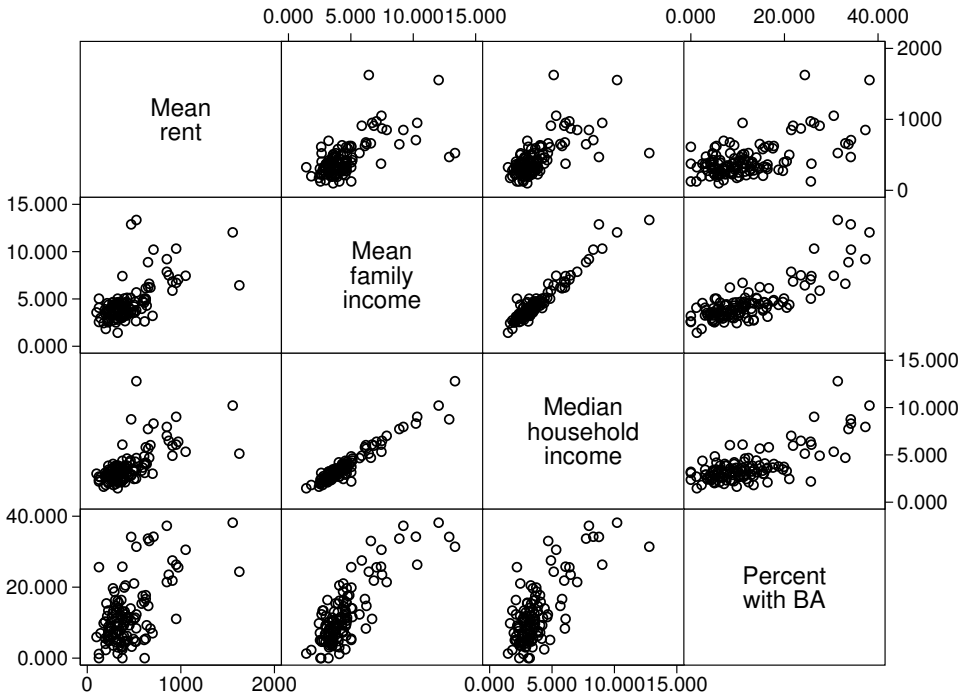

Figure 11.2: Scatter matrix of variables in Table

11.2.3 What to do?

There are essentially two things you can do to solve colinearity:

- drop variables
- combine variables

Drop variables

This is essentially what we did in Models 2 and 3 in Table **??**. Since we can make the argument that both income variables are essentially measuring the same thing, we just pick one and be done with it.

Table 11.2: Models predicting mean rent in census places Table

Variable	Model 1	Model 2	Model 3
Mean family income	-79.636*	56.360***	
	(38.517)	(14.952)	
Median household income	140.822***		69.508***
	(37.090)		(13.827)
Percent with BA	0.111**	0.060	0.064*
	(0.037)	(0.036)	(0.029)
Intercept	133.748**	106.433*	96.588*
	(43.034)	(44.795)	(39.645)
Model Statistics			
N	120.000	120.000	120.000
F	34.683	40.206	48.528
R^2	0.473	0.407	0.453
df Regression	3.000	2.000	2.000
Sum of Squares Regression	3548574.288	3056921.326	3402777.993
df Error	116.000	117.000	117.000
Sum of Squares Error	3956217.179	4447870.140	4102013.473
Variance Inflation Factors			
Mean Income	21.81	2.95	
Median Income	15.55		2.10
Percent with BA	3.40	2.95	2.10
SEs in parentheses, $*p < 0.05, **p < 0.01, ***p < 0.001$			

Combine variables

The other option is to use factor analysis (which is beyond the scope of these notes for now) to create a common factor score that combines the information from the predictors which are highly correlated.

Chapter 12

Outliers

Data are not perfect. At times, we run into the problem of one case having more influence over the least squares line than other cases. In regression, we assume that each case exerts equal influence on the regression line. In some cases, a single extreme value of a predictor can change the entire regression line. For example, consider Figure **??**. The graph on the left is for well behaving data, and the graph on the right shows the effect of a single extreme x variable. Things are not as bad if we have an extreme value on the outcome, however, as we see in Figure **??**.

Life isn't so clear as these figures, however. How would we be able to detect outlier variables when the number of cases are several hundred? Turns out there are several different measures and clues to examine. Let's go over the favorites.

After you fit a model, there are about 6 big methods to detect outliers. These methods create a value for each case, and so these must be plotted for visual inspection. Generally by the case number, or "index."

As a reference, each measure plotted on normal data is displayed in Figure **??**.

12.1 Leverage

If we examine the plots above, we can notice that outlier x tend to drag the regression line around. This is called leverage. Leverage can be modeled by saying that each j^{th} *fitted* value is a function of all the *observed* values of y,

$$\hat{y}_j = h_{1j}y_1 + \ldots + h_{nj}y_n \tag{12.1}$$

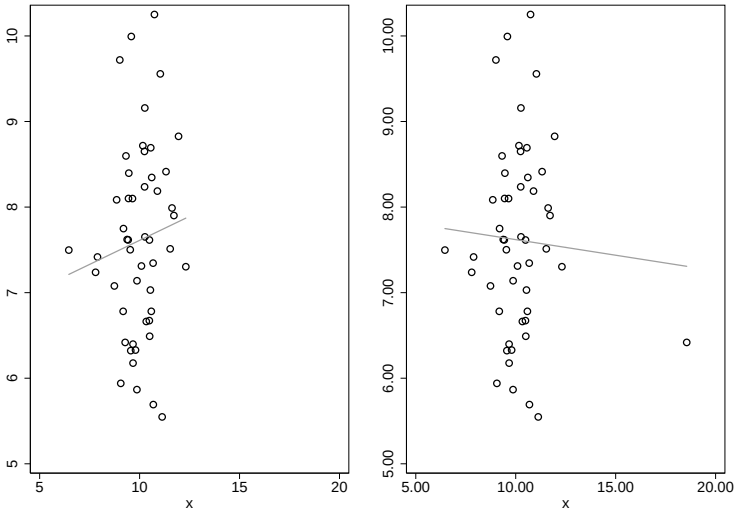

Figure 12.1: Effect of extreme value of x

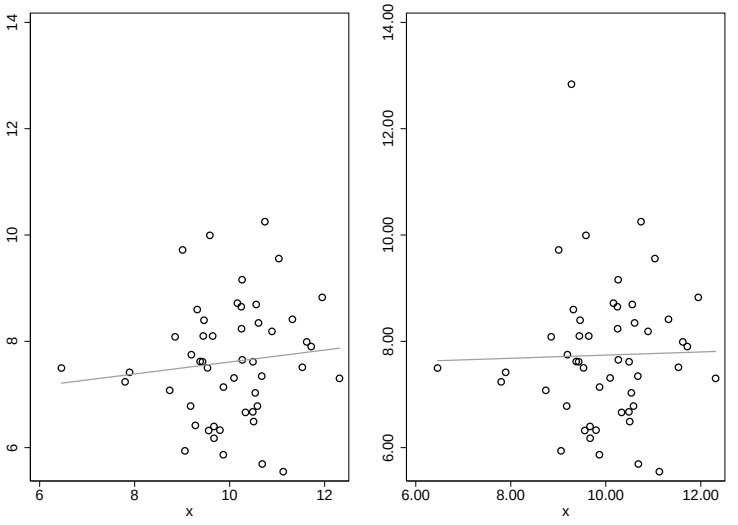

Figure 12.2: Effect of extreme value of y

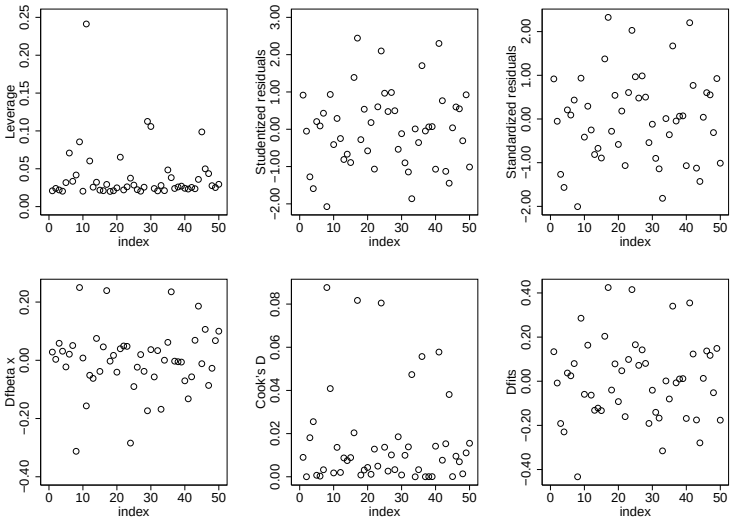

Figure 12.3: Outlier values by case number of typical data in Table

$$= \sum_{i=1}^{N} h_{ij} y_i \tag{12.2}$$

In a simple bivariate regression, h_j can be calculated with

$$h_j = \frac{1}{N} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{12.3}$$

For example, the calculations are shown in Table **??** for a small set of 5 cases.

If we examine the data in Table and plot the leverage values for the regression of y on the x column with the extreme case, you can see that case has a leverage value of 0.62 in Figure reffig:xleverlab.

However, a more useful method to detect which case(s) has (have) the extreme values is to plot the leverage value by case number, or *: index* as in Figure **??**.

As you can see in Figure **??**, the leverage formulas have less utility in the cases of extreme y values.

Table 12.1: Example leverage calculations

	x	$(x-\bar{x})^2$	h
	-0.960	0.006	0.201
	-0.260	0.605	0.312
	-3.020	3.928	0.930
	-0.150	0.789	0.346
	-0.800	0.057	0.211
Sum		5.385	
Mean	-1.038		

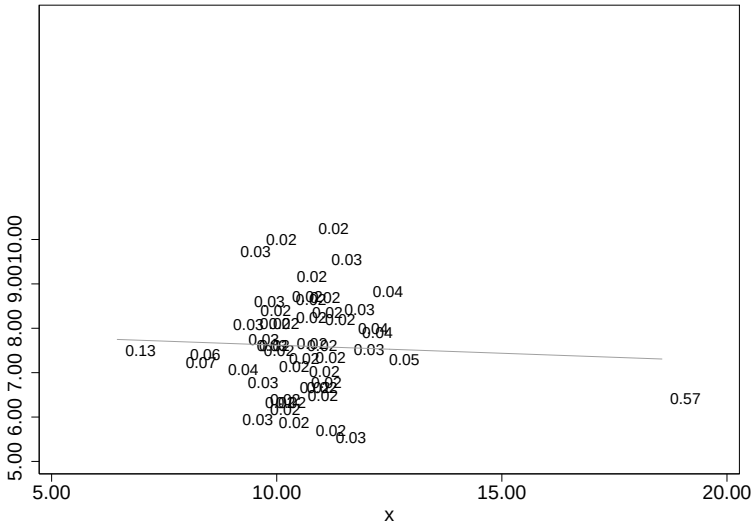

Figure 12.4: Scatter plot of data in Table (y with extreme x) leverage values

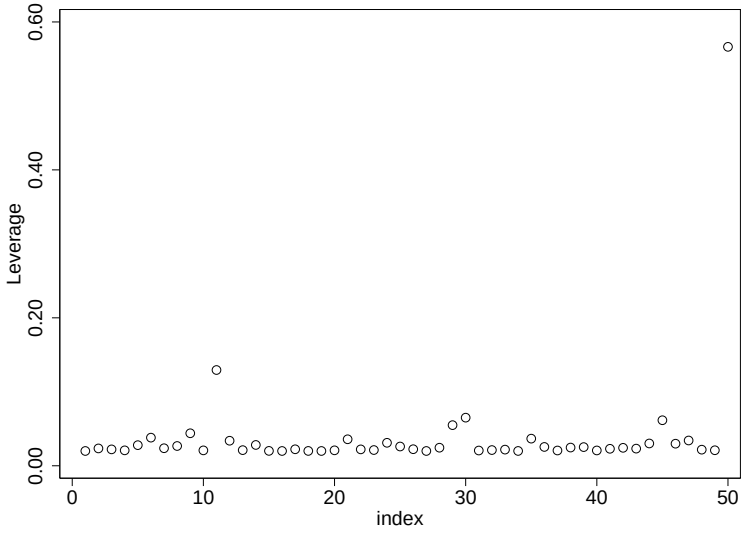

Figure 12.5: Leverage values for extreme value of x by case number of data in Table

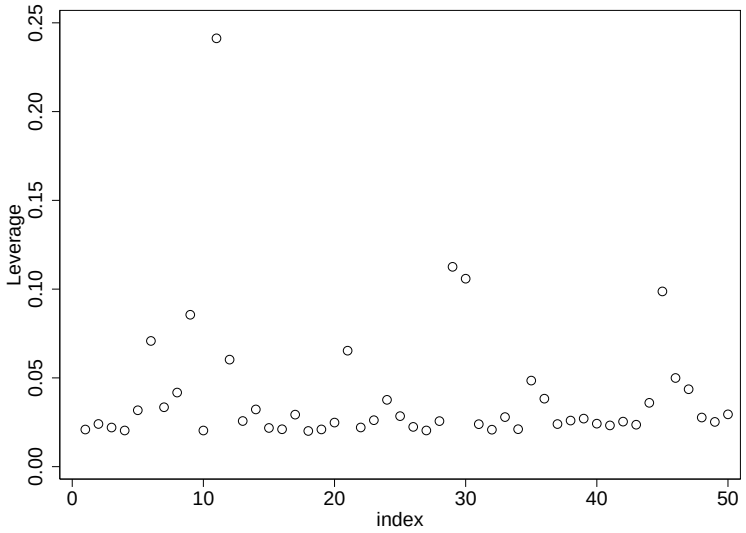

Figure 12.6: Leverage values for extreme value of y by case number of data in Table

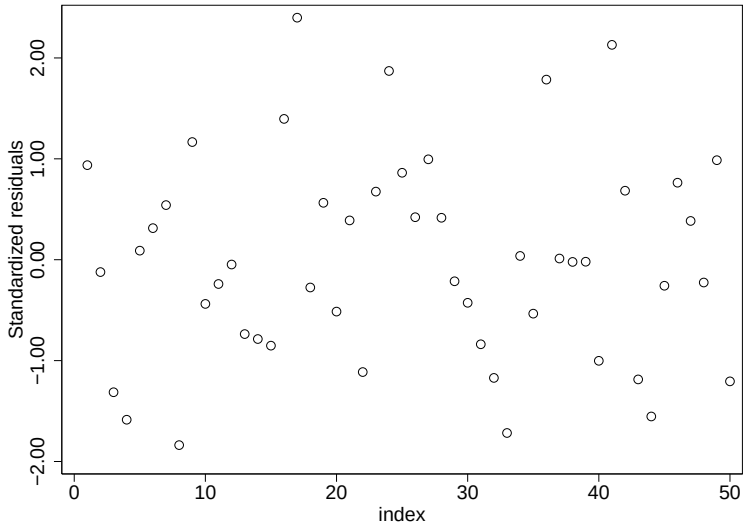

Figure 12.7: Standardized residual values for extreme value of x by case number of data in Table

12.2 Standardized residuals

Another plausible measure is the standardized residual from the model relative to the leverage, h_i

$$e_i^{(standardized)} = \frac{e_i}{s_e\sqrt{1-h_i}} \tag{12.4}$$

where s_e is the standard deviation of the residuals from the model. Figures with extraordinary values of x and y are displayed in Figures **??** and **??**, respectively.

12.3 Studentized residuals

Another approach to residuals is to employ a standard deviation of the results from a model that excludes a particular case. Thus, for a given case i, this residual is

$$e_i^{(studentized)} = \frac{e_i}{s_e^{(-i)}\sqrt{1-h_i}} \tag{12.5}$$

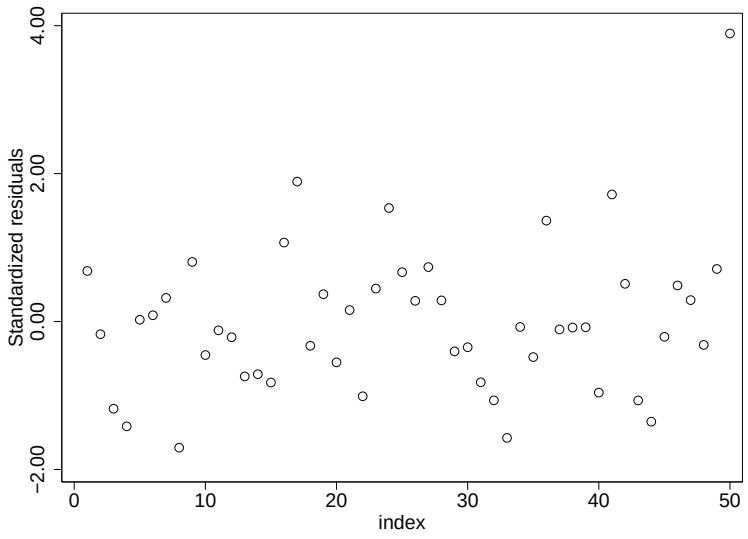

Figure 12.8: Standardized residual values for extreme value of y by case number of data in Table

Figures with extraordinary values of x and y are displayed in Figures **??** and **??**, respectively.

12.4 DFBETA

Yet another way to think about outliers is not how they influence the errors, but instead how they influence the regression slopes. Thus, the DFBETA. This is the difference in a regression slope for predictor p if a particular case i was removed

$$D_{pi} = \beta_p - \beta_p^{-i} \tag{12.6}$$

Figures with extraordinary values of x and y are displayed in Figures **??** and **??**, respectively.

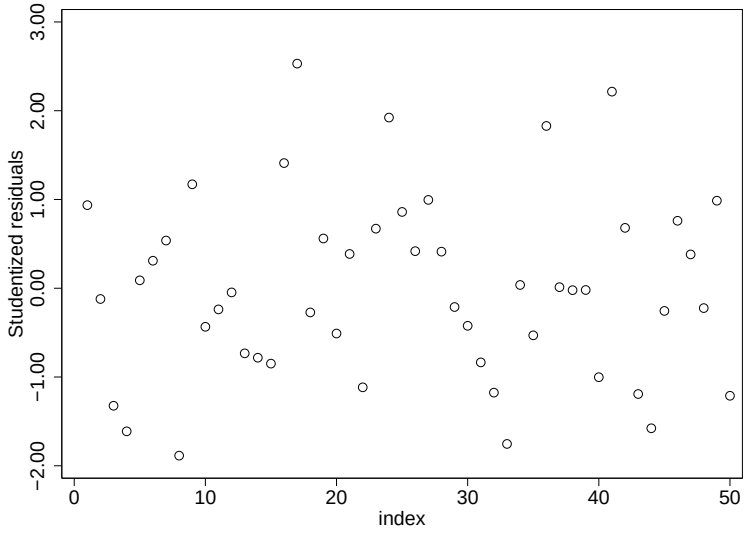

Figure 12.9: Studentized residual values for extreme value of x by case number of data in Table

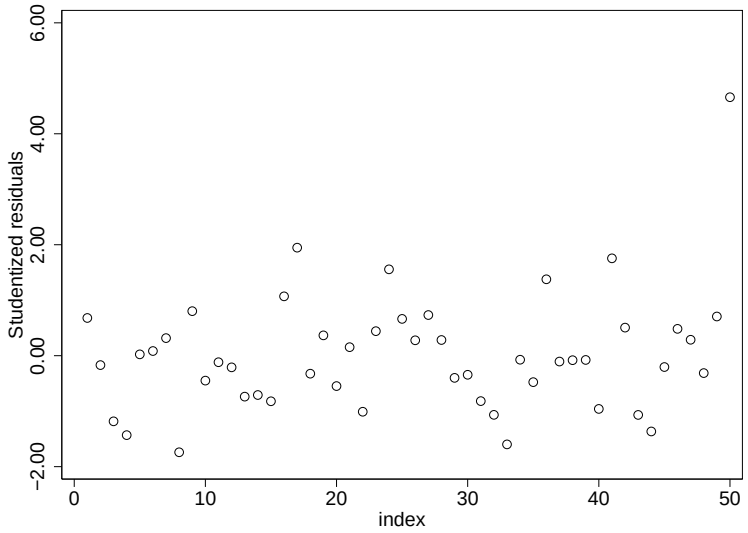

Figure 12.10: Studentized residual values for extreme value of y by case number of data in Table

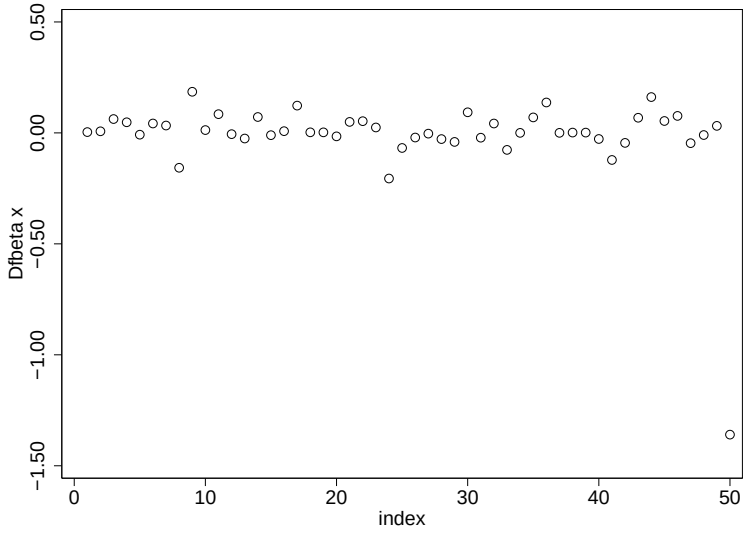

Figure 12.11: Df-Beta values for extreme value of x by case number of data in Table

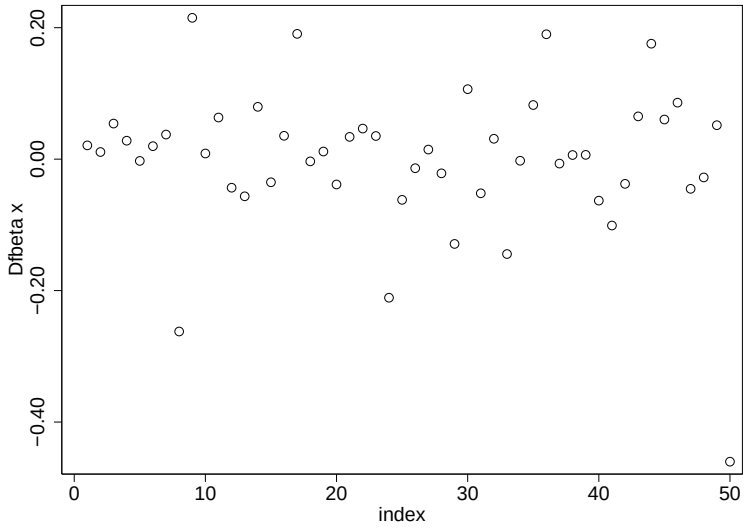

Figure 12.12: Df-Beta values for extreme value of y by case number of data in Table

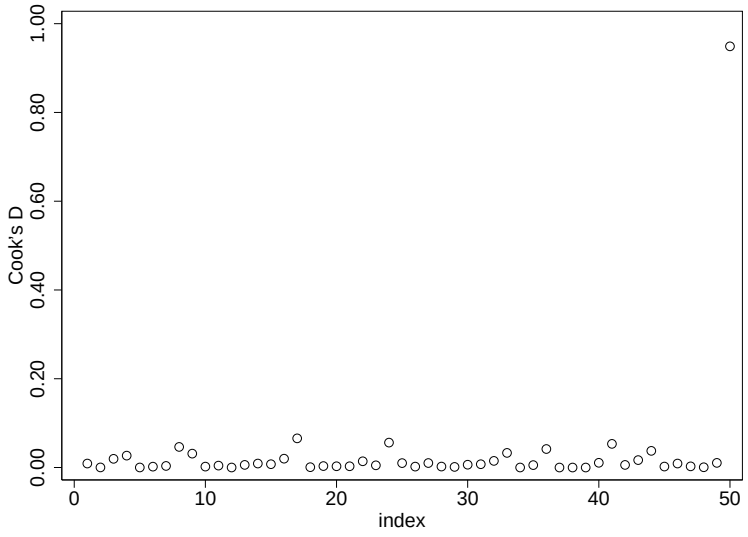

Figure 12.13: Cook's distance values for extreme value of x by case number of data in Table

12.5 Cook's distance

The Cook's distance measure combines the leverage values with standardized residuals

$$D_i = \frac{e_i^{(standardized)}}{k+1} \times \frac{h_i}{1-h_i} \tag{12.7}$$

Figures with extraordinary values of x and y are displayed in Figures **??** and **??**, respectively.

12.6 Dfits

Similar to Cook's distance is the DFITS measure.

$$D_i = e_i \times \sqrt{\frac{h_i}{1-h_i}} \tag{12.8}$$

Figures with extraordinary values of x and y are displayed in Figures **??** and **??**, respectively.

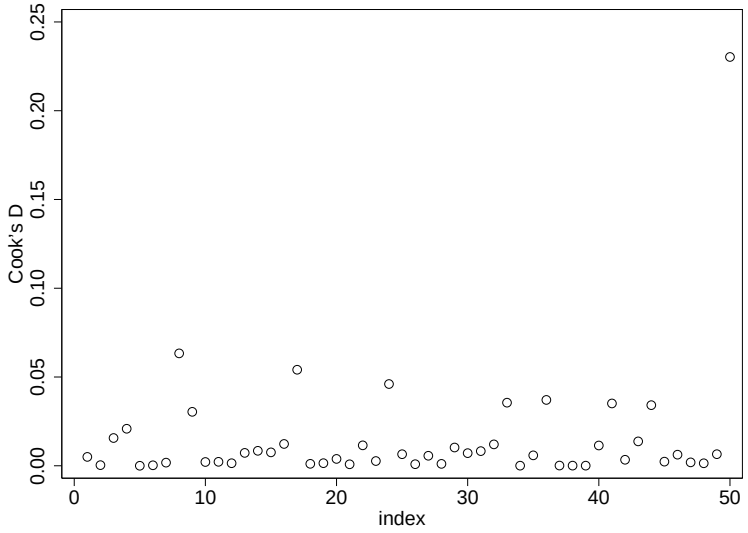

Figure 12.14: Cook's distance values for extreme value of y by case number of data in Table

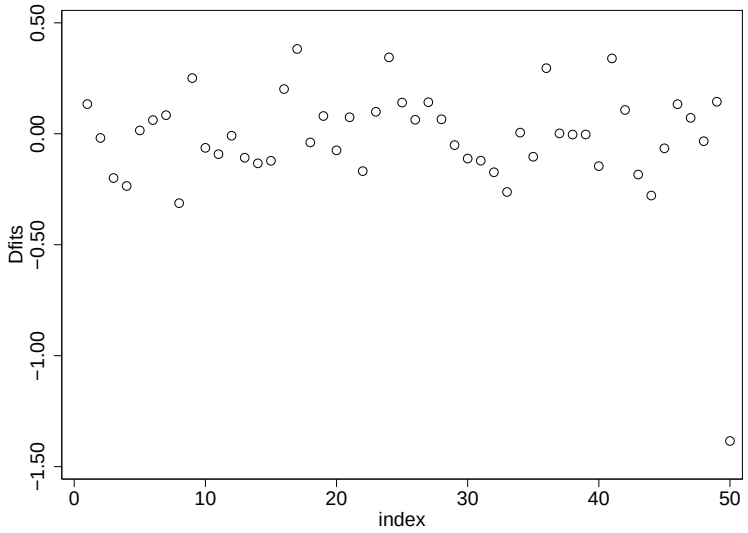

Figure 12.15: D-fit values for extreme value of x by case number of data in Table

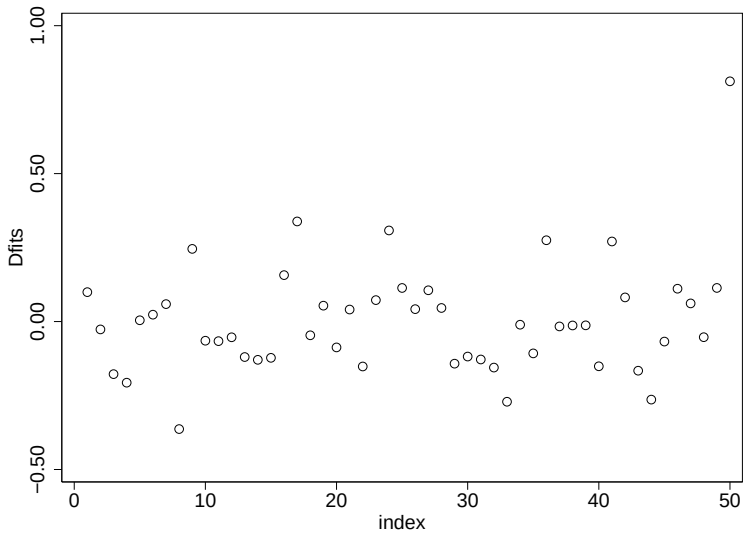

Figure 12.16: D-fit values for extreme value of y by case number of data in Table

12.7 Summary

As you can see, there are several measures to detect outliers. In most cases, it will be difficult to spot them initially. Table **??** gives a general summary of the different methods and how well they work to discover extreme x or y values. In the end, I recommend the use of the Dfits measure as it combines the leverage values with the residuals (thus involves both the outcome and predictors).

Table 12.2: Summary of outlier measures for extreme x and y

Measure	Extreme x	Extreme y
Leverage	Yes	No
Standardized residuals	No	Yes
Studentized residuals	Maybe	Yes
DFBETA	Yes	No
Cook's distance	Yes	Maybe
DFits	Yes	Yes

Chapter 13

Coding strategies

This chapter is all about how to alter our predictors and outcomes to make regression models give us the information we need to know.

13.1 Centering

The most basic transformation of a variable is to add or subtract something from it. An important example of this is to center a variable. This is also called de-meaning a variable (but I prefer "centering"). Centering a variable is the simple procedure of taking a variable and subtracting its mean value.

$$x_i^* = x_i - \bar{x} \tag{13.1}$$

This does powerful things for regressions and it makes your life easier. The benefits include making your intercept more interpretable and understanding complex coding schemes such as interactions much easier. It should be noted that centering will not change your slopes or the standard errors of your slopes. It will change your intercept and the standard error of your intercept.

It works by moving the y-intercept to the mean of your predictor. Since the intercept is always when your predictor(s) are 0, you make 0 the mean when you subtract it, see Figure **??**. Centering can be done in models with a single variable or with many variables.

To understand what centering does, consider Model 3 from Table **??**:

$$\hat{y}_i = \beta_0 + \beta_1 edu_i + \beta_2 age_i + \beta_3 female_i$$

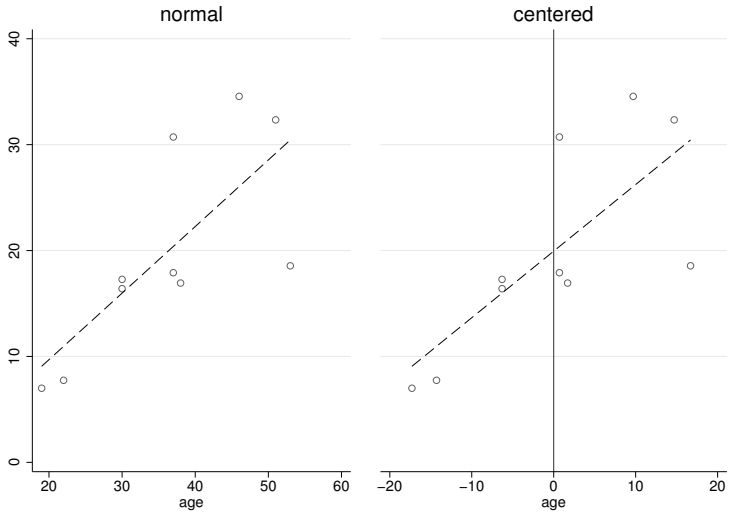

Figure 13.1: Illustration of centering

We know that β_1 is the effect of a single year of education on wages, β_2 is the effect of a single year of age on wages, and β_3 is the difference between females and males. What does β_0 mean again? It's the (average) value of $wage$ when edu, age, and $female$, all equal zero. Consider the models in Table **??**, which use more data than the wage models in the other tables. In Model 1, none of the variables are centered. So, according to this model, the intercept tells us that uneducated babies who are male make -4.651 dollars an hour. This doesn't make sense on many levels. We now move to Model 2, where we have centered age. Now the intercept tells us that an *averaged aged* male with no education can expect to make about 5 dollars an hour. This is plausible. Next, in Model 3, we also center education. Now, an averaged aged male with an average number of years of education makes 17.30 an hour. Notice that in each case, while the intercept changed, the standard error of the intercept kept getting smaller and smaller. No other numbers in the table changed.

The standard error of the intercept gets smaller because the variance of any predicted value is always lowest at the mean. The reason for this is because the variance of any predicted value of y based on some value of x,

Table 13.1: Models predicting age with different centering strategies

Coefficients	Model 1	Model 2	Model 3
edu	0.930***	0.930***	0.930***
	(0.034)	(0.034)	(0.034)
age	0.261***	0.261***	0.261***
	(0.009)	(0.009)	(0.009)
female	-3.474***	-3.474***	-3.474***
	(0.207)	(0.207)	(0.207)
Intercept	-4.651***	5.006***	17.289***
	(0.597)	(0.474)	(0.147)
Model Statistics			
N	3997.000	3997.000	3997.000
F	590.668	590.668	590.668
R^2	0.307	0.307	0.307
df Regression	3.000	3.000	3.000
Sum of Squares Regression	75828.174	75828.174	75828.174
df Error	3993.000	3993.000	3993.000
Sum of Squares Error	170869.757	170869.757	170869.757

Model 2: age centered, Model 3: edu and age centered.
SEs in parentheses, $***p < 0.001$

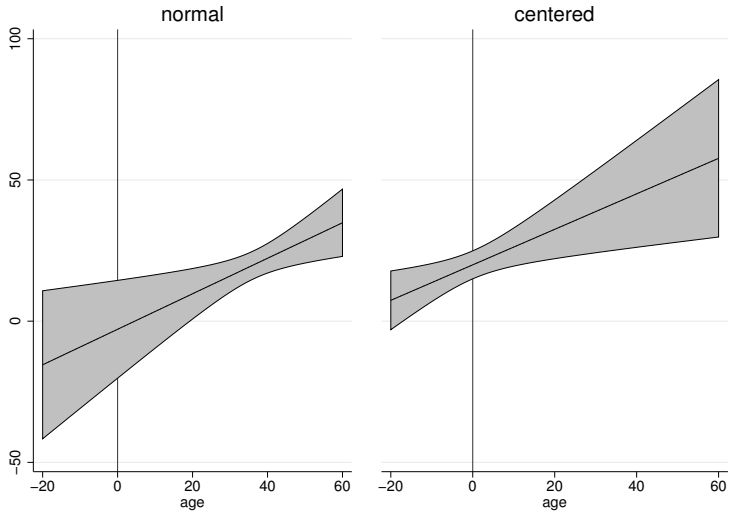

Figure 13.2: Confidence interval of regression line

say x_0, is equation (**??**)

$$\text{var}\left(\hat{y}_0\right) = \left(\frac{\sigma^2}{N} + \frac{\left(x_0 - \bar{x}\right)^2 \sigma^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}\right)\sigma^2$$

and that fraction to the right gets smaller the closer the value is to the mean of x. This gets reflected in the smaller standard error of the intercept. We can see this in the bivariate case by looking at Figure **??**. Note how the confidence intervals are small around the mean of the distribution of the predictor. Centering moves the intercept to this point of small(er) variance. Centering is a good practice for building a good "reference" statistic.

13.2 Changing the unit

A regression slope always tells us the difference in the outcome based on a single-unit increase in the predictor. Sometimes, a single unit increase isn't meaningful. For example, take age. Your wage probably won't increase that much from year to year. According to Table **??**, Model 1, it will increase by 26.1 cents a year. A more meaningful distinction would be 10 years of age.

Table 13.2: Wage models with different age and education units

Coefficients	Model 1	Model 2	Model 3
edu	0.930***	0.930***	2.823***
	(0.034)	(0.034)	(0.104)
age	0.261***	2.613***	2.613***
	(0.009)	(0.087)	(0.087)
female	-3.474***	-3.474***	-3.474***
	(0.207)	(0.207)	(0.207)
Intercept	-4.651***	-4.651***	7.632***
	(0.597)	(0.597)	(0.353)
Model Statistics			
N	3997.000	3997.000	3997.000
F	590.668	590.668	590.668
R^2	0.307	0.307	0.307
df Regression	3.000	3.000	3.000
Sum of Squares Regression	75828.174	75828.174	75828.174
df Error	3993.000	3993.000	3993.000
Sum of Squares Error	170869.757	170869.757	170869.757

Model 2: age in 10 year units.
Model 3: age in 10 year units and edu is standardized.
SEs in parentheses, $***p < 0.001$

To capture this, we could divide age by 10 and rerun the model. This would make $\beta_2 = 2.613$. By moving the decimal of the predictor left, we move the decimal of the slope (and standard error) to the right. Compare the effect of age in Models 1 and 2 in Table **??**. Nothing else changes. It is good practice to keep your predictors in units to maximize the digits available in your tables.

13.3 Standardization

Another transformation is z-scoring. It is simply centering the predictor and dividing by the standard deviation of the predictor

$$x_i^{(z)} = \frac{x_i - \bar{x}}{s_x} \tag{13.2}$$

Sometimes this is called standardization. This changes the units of the variable into "standard deviation" units. This also centers variables on their mean. The slope's interpretation is now that a one-unit increase in the predictor is "a standard deviation" increase. Compare the effect of education in Models 2 and 3 in Table **??**. In Model 2, for every year of education, wages increase by 93 cents. In Model 3, we see that an increase of a standard deviation in education increases wages by about 2.82 dollars. Nothing else changes.

13.4 Power Transformations

Sometimes the data are not linear. They need to be linear for a regression to count. There are several ways to transform data. One is the Box-Cox family of transformations

$$x \to x^p \equiv \frac{x^p - 1}{p} \forall p > 0 \tag{13.3}$$

This gives us the various effects of squaring, etc. This is nice when we think that the effect of x on y is not constant (like it accelerates or decelerates). **Be careful with squares, always graph out our predicted values to know what you are dealing with**.

Many people make the mistake that just because a square term is significant, it must mean a u-shaped relationship. That depends on the strength

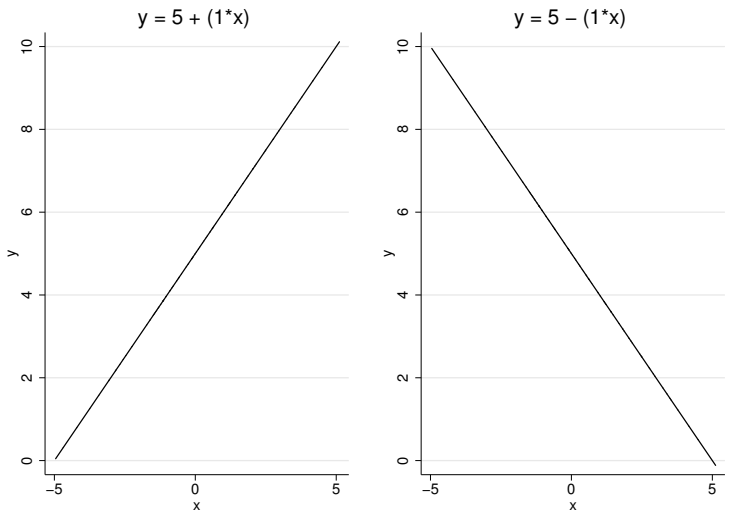

Figure 13.3: Models without a quadratic relationships

of the main effect (i.e., the effect of x) and the effect of the square (i.e., the effect of x^2). Figures **??** through **??** illustrate the differences. What can also be important is to know what the slope of the line is at any given point along the line. Let's take the function

$$y = \beta_0 - \beta_1 x - \beta_2 x^2$$

$$y = 5 - 1x - 0.2x^2$$

What is the slope (or the slope of the tangent line) of this line at $x = 0.50$? Calculus gives us an easy formula for this

$$\frac{\partial y}{\partial x}|x = \beta_1 + 2\beta_2 x \tag{13.4}$$

for example, when x = 0.50,

$$\frac{\partial y}{\partial x}|x = -1 + 2\,(-.2)\,x$$

$$\frac{\partial y}{\partial x}|x = -1 + 2\,(-.2)\,0.50$$

$$\frac{\partial y}{\partial x}|x = -1.2$$

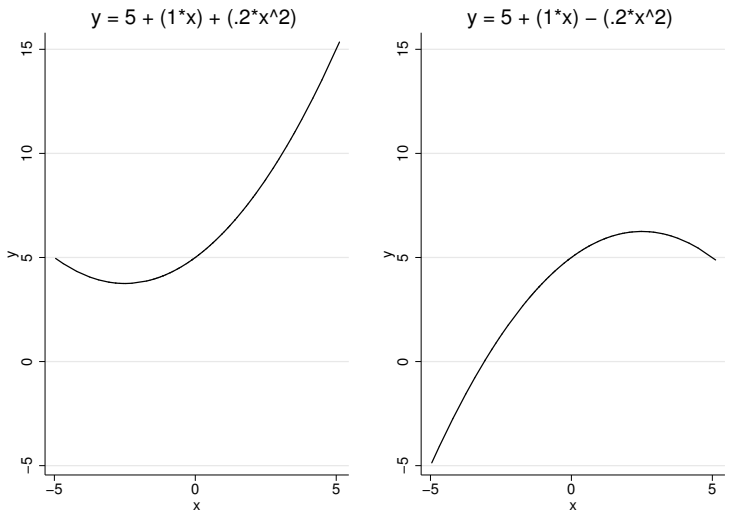

Figure 13.4: Models with a positive relationship that accelerates (left) and one that decelerates (right)

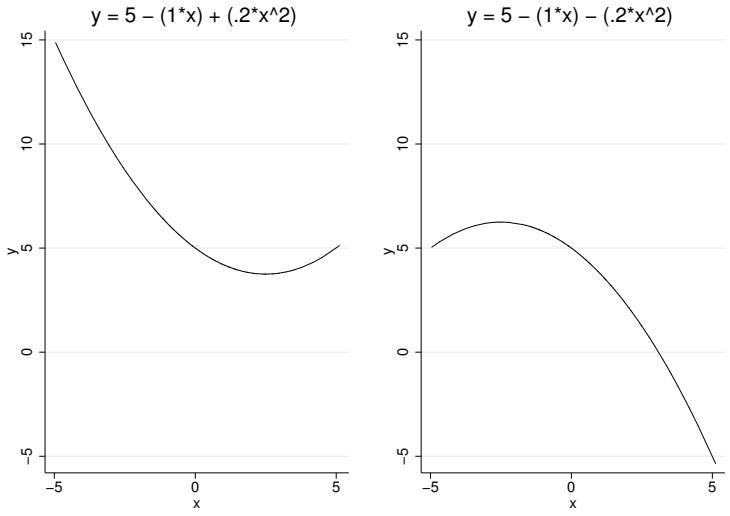

Figure 13.5: Models with a negative relationship that decelerates (left) and one that accelerates (right)

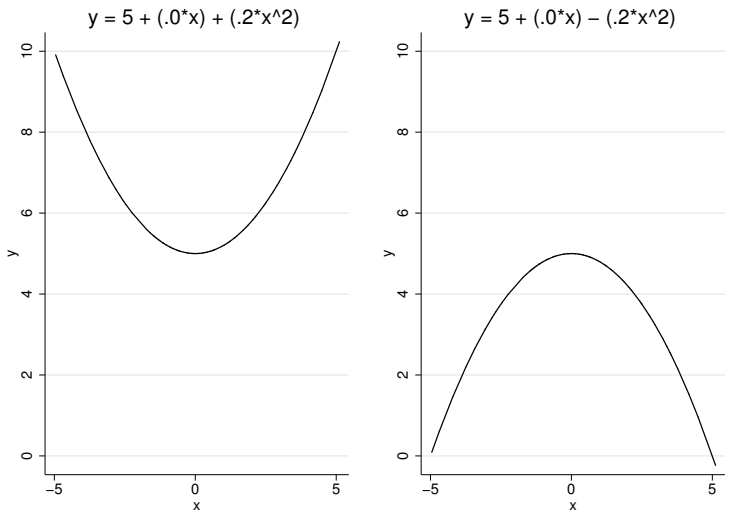

Figure 13.6: Models without a main effect but a positive quadratic relationship (left) and a negative relationship (right)

This is visualized in Figure **??**; notice the intercept for the tangent line is the same as the regression line.

13.4.1 Logs

A special case of a power transformation is the log. Figure **??** shows the natural log, $\ln(x)$. This helps tone down predictors, and even outcomes, that are highly skewed. Once transformed, we can then run our regression on them. Even more fun is logging both x and y, which produces an elasticity (i.e., a 1 percent increase in x gives a β percent change in y).

Examples with logs as outcome, predictor, or both

I have a sample of people who give to charities. We want to know the relationship between what they give and their income in thousand dollar units. Table **??** give the summary statistics of the variables we will use: the amount given to charity in the last year, household income in thousand

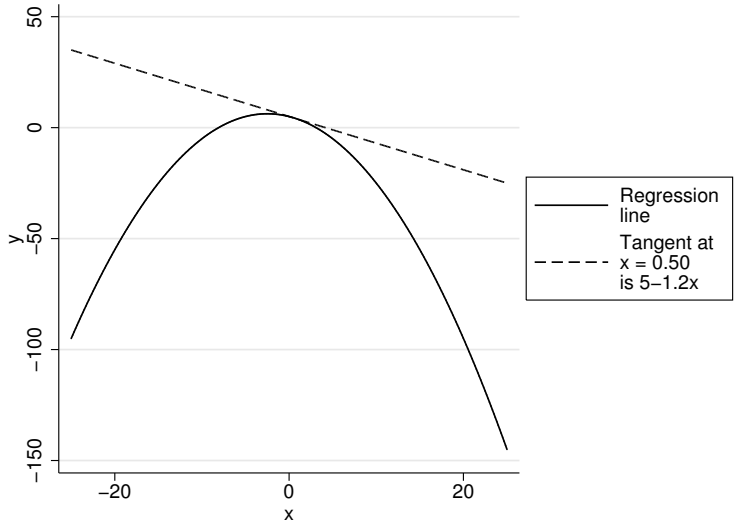

Figure 13.7: Quadratic relationship with tangent

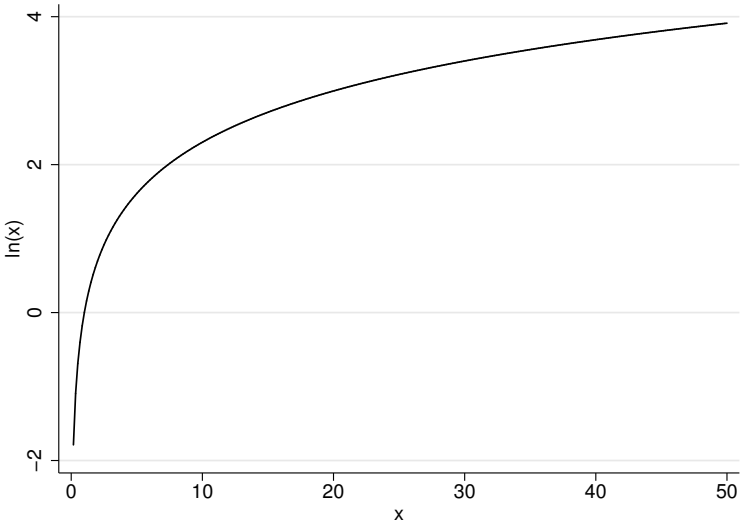

Figure 13.8: $\ln(x)$

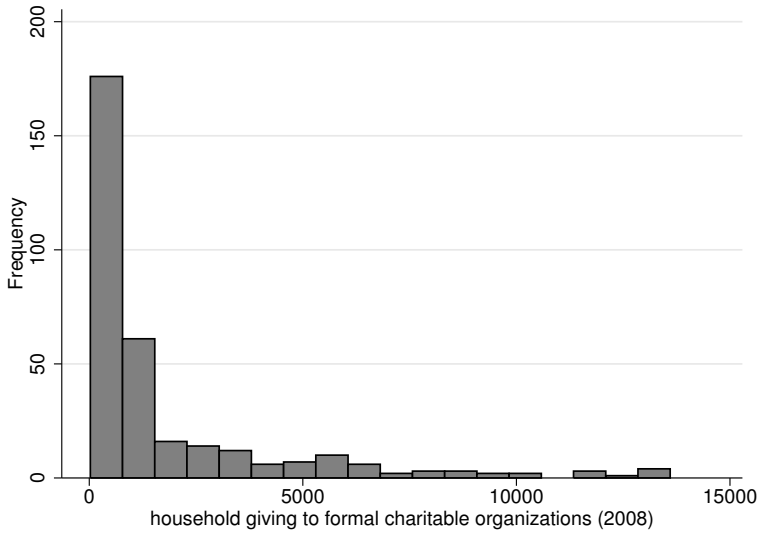

Figure 13.9: Distribution of charitable giving

dollar units[1], the natural log of the amount given, and the natural log of the household income in thousands. When we examine our outcome, amount

Table 13.3: Summary statistics for charity models

Variable	Description	$mean$	$SE\,(mean)$
gaveamt	Amount given	1796.99	148.64
hhincome	Income in k	60.36	1.20
lngaveamt	ln(Amount given)	6.54	0.08
lnhhincome	ln(Income in k)	4.02	0.02

given, we see that it is highly skewed, see Figure **??**. This will most likely produce several problems:

1. The variance will not be constant

2. The mean of the residuals will not consistently be zero

[1]In this section we present data where income was measured as several categories and total amount given was assessed numerically. We transform the income categories into the midpoint dollar amount, then divide by 1000.

3. We may have outliers in the tail of the distribution

Table 13.4: Models predicting charitable giving in dollars

Coefficients	Model 1	Model 2	Model 3	Model 4
hhincome centered	22.614***	0.014***		
	(6.777)	(0.004)		
ln(hhincome) centered			1156.648**	0.730***
			(356.840)	(0.193)
Intercept	1796.988***	6.543***	1796.988***	6.543***
	(146.392)	(0.079)	(146.528)	(0.079)
Model Statistics				
N	328.000	328.000	328.000	328.000
F	11.134	14.822	10.506	14.296
R^2	0.033	0.043	0.031	0.042
df Regression	1.000	1.000	1.000	1.000
df Error	326.000	326.000	326.000	326.000
Log-likelihood	-3049.966	-582.776	-3050.271	-583.030
Models 2 and 4 have a logged outcome. hhincome in thousand dollar units. SEs in parentheses, $***p < 0.001$				

Table **??** presents four models. The first model predicts giving with income using the unaltered versions of both variables (note: income is centered):

$$gave_i = \beta_0 + \beta_1 \left(hhincome_i - \overline{hhincome_i}\right) + e_i.$$

The second model predicts the natural log of giving with the unaltered version of income:

$$\ln\left(gave_i\right) = \beta_0 + \beta_1 \left(hhincome_i - \overline{hhincome_i}\right) + e_1.$$

The third model predicts unaltered giving with the natural log of income, centered on the mean natural log of income:

$$gave_i = \beta_0 + \beta_1 \left(\ln\left(hhincome\right)_i - \overline{\ln\left(hhincome\right)_i}\right) + e_i.$$

Finally, the fourth model predicts the natural log of giving with the natural log of income, centered as well:

$$\ln\left(gave_i\right) = \beta_0 + \beta_1 \left(\ln\left(hhincome\right)_i - \overline{\ln\left(hhincome\right)_i}\right) + e_i.$$

Interpretation of Model 1 Since we centered income, the intercept indicates that a household with average income gives about 1,797 dollars. For each thousand dollars increase in income, giving increases by about 23 dollars. Overall, this model explains about 3.3 percent of the variation.

Interpretation of Model 2 In this model, we have transformed giving into the natural log of giving. Income is still centered, so the intercept is still the average natural log of giving. We can take the exponent of this quantity to estimate the geometric mean of giving (not the arithmetic mean)

$$e^{\beta_0} = (\bar{y}_{geometric}|x = 0) \tag{13.5}$$

$$e^{6.543} = 694.37$$

which tells us that the average household gives about 694 dollars. This estimate is much smaller than Model 1 (1,797 dollars).

Estimating the effect of income requires us to take the exponent of β_1. Thus, $\left(e^{\beta_1} - 1\right) \times 100$ will tell us the percent increase in the outcome for a one unit change in the predictor if $\beta_1 > 0$. When $\beta_1 < 0$, $\left(1 - e^{\beta_1}\right) \times 100$ will tell us the percent decrease in the outcome for a one unit change in the predictor. For example, when income is at the mean, we have already established that the amount given is 694.37 dollars. If we increase income by a thousand dollars (a one unit increase), the dollar amount given is

$$e^{6.543+0.014} = 704.16$$

The ratio of these two amounts is

$$\frac{y|x = 1}{y|x = 0} = \frac{704.16}{694.37} = 1.014$$

which is equal to the exponent of the slope

$$e^{\beta_1} = \frac{y|x = 1}{y|x = 0}$$

$$e^{0.014} = 1.014$$

which indicates a

$$\left(e^{\beta_1} - 1\right) \times 100 = (1.014 - 1) \times 100 = 1.4 \text{ percent}$$

increase. Note that the fact that the coefficient was 0.014 and the percent increase was 1.4 is a coincidence. When the coefficient is larger, the numbers will not match.

If we log these variables, their distributions behave better (but what we really care about is the distribution of giving)

Interpretation of Model 3 We know that the average donation for the log of the average household income is 1,796.99. What does the coefficient mean? To understand that we need to realize that since we logged our predictor, our comparison is between the average log and an additional log. This would be a lot of income. Let's instead say that income increases 10 percent (or by a factor of 1.1). This makes the difference in our outcome as

$$1156.648 \times \ln(1.1) = 110.24 \text{ dollars}$$

The final use of log-transformations in OLS regression is when both dependent and predictor variable are logged.

Interpretation of Model 4 In this model, both the outcome and predictor is transformed to the natural log. This produces an elasticity. Thus, coefficient tells us the percent (*not proportion*) change in the outcome based on a 1 *percent* change in the predictor. Here, we see that giving changes by 0.730 percent, for each percentage change in the predictor. Here is a walk-through on how this works.

Table **??** gives the means for our variables, both the unaltered version and the natural log version. From Table **??**, we can find that the dollar amount for the mean natural log of giving is $e^{6.543} = 694.37$ dollars. The mean natural log of income is 4.02, which translates into an amount of $e^{4.02} = 55.70$ thousand dollars. A 1 percent increase of 55.70 thousand dollars is $1.01 \times 55.70 = 56.26$ thousand dollars, of which the natural log is 4.03. Thus, a 1 percent increase in the natural log of income is 0.01. Plugging this into the model, our predicted natural log of giving is $6.543 + 0.730\,(0.01) = 6.550$, the exponent of which is 699.45. $(699.45 - 694.37)/694.37 = 0.01$, or a 1 percent increase.

Testing models with log likelihoods

As a bonus, Table **??** includes the log likelihood of each model. We can only compare log likelihoods for models with the same outcome and exact same

set of data. That means we can compare Models 1 and 3, and Models 2 and 4.

Recall that the log likelihood of a model is $\ln(L(\theta))$. Also remember that the test of two log likelihoods for models A and B is equation (**??**)

$$\chi^2 = 2\left(\ln\left(L\left(\theta_a\right)\right) - \ln\left(L\left(\theta_{null}\right)\right)\right)$$

We see the log likelihood of Model 1 in Table **??**, the model with unaltered predictors with the unaltered outcome, is -3029.966, and the likelihood of Model 3, the model with the natural log of income with the unaltered outcome is -3050.271. The test comparing whether Model 3 is a better model is

$$\chi^2 = 2\left(-3029.966 - -3050.271\right) = 40.61$$

We can then evaluate this against the χ^2 distribution with 1 degree of freedom, which produces a probability of 0.000000002. This is clear evidence that using the unaltered version of the predictor, Model 1, is a much better model. Looking at Models 2 and 4, this test results in 0.508, the probability of which is 0.476, telling us that there is no difference, statistically, in the model fit between Models 2 and 4.

Chapter 14

Analysis of covariance (ANCOVA)

Analysis of covariance (ANCOVA) is simply a regression with both continuous and categorical predictors. It has, however, a special application in situations in which we have "pre" and "post" scores.

14.1 The meaning of the intercept

Before we enter a categorical predictor, let us first examine some properties of the intercept in models with paired y and x variables. Consider the hypothetical data in Table . This table has three columns, the first column is our post-test scores, the second column is our pretest scores, and the third column is the difference for reference. Since the "pre" and "post" tests are from the same units, we can compare the average change with a *paired* t-test

$$t = \frac{\bar{d}\sqrt{N}}{s_d} \tag{14.1}$$

where $\bar{d}$ is the average difference (that is, the average of a new variable measuring the difference) and s_d is the standard deviation of those differences. The result for the data in Table is $\bar{d} = 9.4791$ and $s_d = 10.98465$ with 100 observations (which means the standard error of the difference is $\frac{s_d}{\sqrt{100}} = 1.098465$). This leads to a t-value of 8.6294, which has a very small p-value.

How can we replicate this result with regression?

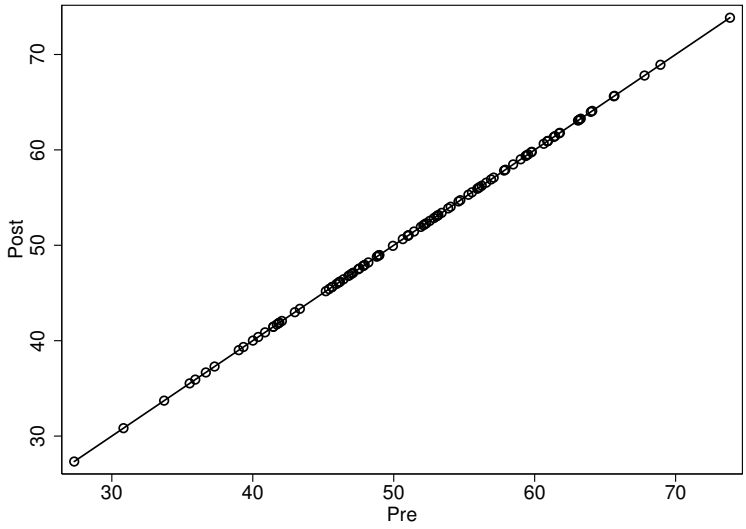

Figure 14.1: Scatter plot of data in which pre and post scores are the same with solid regression line

14.1.1 As-is regression

The idea is simple, suppose that there was no change, a scatter plot would simply be a diagonal line, as in Figure **??**. If, however, on average respondents did change, then the line should rise or fall and this would be reflected in the intercept.

Thus, we may think we could fit the model

$$post_i = \beta_0 + \beta_1 pre_i + e_i \tag{14.2}$$

and look at the intercept.

This thinking is adequate when the regression slope (β_1) for pre is approximately 1. However, the intercept is less telling about the change for the average pre score if the slope is greater or less than 1, as can be seen in Figure **??**

A regression of post on pre, however, gives the difference for a low-scoring pre. This is reflected in Model 1 in Table **??**, where the intercept is -12.498. This result is very different than the result of the paired t-test. The problem is that we are generally interested in the change for the *average* pretest, and the intercept reflects the change for a pre of 0.

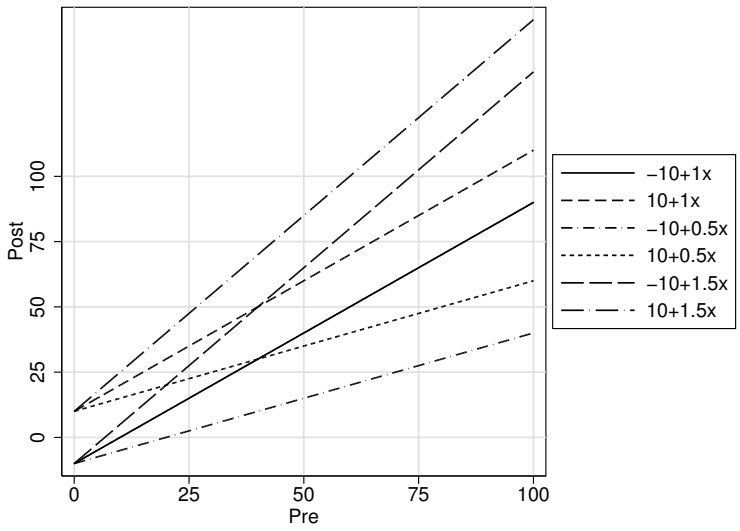

Figure 14.2: Different functions for pre-post regressions

14.1.2 Regression on pre-centered variables

If we center post and pre, on the *pre* mean,

$$post_i - p\bar{r}e = \beta_0 + \beta_1 \left(pre_i - p\bar{r}e\right) \tag{14.3}$$

the intercept now reflects the average change for the average pre-test. This is reflected in Model 2 in Table **??**. Note that in this case, β_0 is the result of the paired t-test. We can see the difference in Figure **??**. When the data are not centered, the intercept is negative. However, by centering both the post and pre on the pre mean, we move the axis for x and y and now the intercept is positive. In general, this gives the *average* change for the *average* pre, which reflects the paired t-test difference.

14.1.3 Regression on the difference

The first is to create a new variable that is the difference of post-pre and run an intercept only regression

$$post_i - pre_i = \beta_0 + e_i \tag{14.4}$$

This is represented in Model 3 in Table **??**, which reflects the result of the paired t-test.

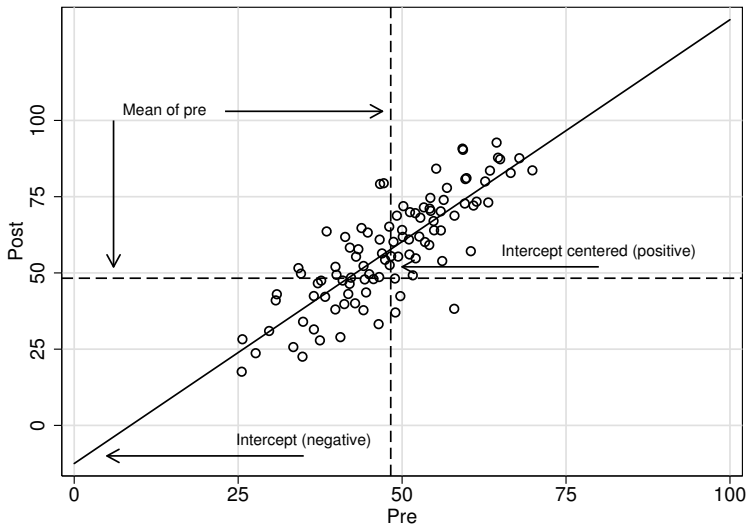

Figure 14.3: Scatter plot of data in Table with solid regression line and dashed reference lines

14.1.4 Why different standard errors for centered and difference models?

Comparing Models 2 and 3 in Table **??** you will notice that the standard errors for the intercept are different. To understand why, we need to remember where the standard error of intercepts comes from. The intercept is a predicted value (when x is 0), and so like any other predicted value, the variance of the predicted value is equation (**??**)

$$\text{var}\left(\hat{y}_0\right) = \left(\frac{\sigma^2}{N} + \frac{\left(x_0 - \bar{x}\right)^2 \sigma^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}\right)\sigma^2$$

and the standard error is

$$\text{SE}\left(\hat{y}_0\right) = \sigma\left(\frac{1}{N} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}\right)^{1/2} \tag{14.5}$$

which is governed by the residual variation σ and the deviation from the mean. In the case of a centered model, the term $\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$ drops out, which

leaves

$$\mathrm{SE}\left(\hat{y}_0\right)^{centered} = \sigma\left(\frac{1}{N}\right)^{1/2} \tag{14.6}$$

$$\mathrm{SE}\left(\hat{y}_0\right)^{centered} = \frac{\sigma}{\sqrt{N}} \tag{14.7}$$

which is the same as the variance of the mean, c.f. equation (**??**), and since σ is smaller in Model 2 than in Model 3 (because we use the covariate), the standard error is smaller reflecting the use of more information.

Table 14.1: Models to analyze differences between pre and post data in Table

Coefficients	Model 1	Model 2	Model 3
Pre	1.455***	1.455***	
	(0.102)	(0.102)	
Intercept	-12.498*	9.479***	9.479***
	(5.035)	(1.007)	(1.098)
Model Statistics			
N	100.000	100.000	100.000
F	202.700	202.700	0.000
R^2	0.674	0.674	0.000
df Regression	1.000	1.000	0.000
Sum of Squares Regression	20547.835	20547.835	0.000
df Error	98.000	98.000	99.000
Sum of Squares Error	9934.307	9934.307	11945.602
σ	10.068	10.068	10.985
Model 2 pre and post both centered on pretest mean			
Model 3 predicts post-pre difference			
SEs in parentheses, $*p < 0.05, **p < 0.001$			

14.1.5 Gains in power using regression technique

A natural question is to ask under what circumstances using the regression technique is favorable. That is, when does it decrease standard errors by a measurable amount?

This can be answered with a quantity I call the PTIF for paired test inflation factor, which is

$$PTIF = \frac{1 + v - 2\rho\sqrt{v}}{(1 - \rho^2)\, v} \times \frac{N - 2}{N - 1} \tag{14.8}$$

where ρ is the correlation between y and x v is the ratio of the variance of y to x. This quantity tells you how much larger the variance of the test is using the paired approach compaired to the the regression approach.

14.2 ANCOVA: Bush election opinion example

100 randomly sampled cases from the American National Election Study (2000 Pre- and Post-Election Survey) are presented in Table . The cases are evenly divided among respondents who believed the outcome of the 2000 election was fair and those who believed it was not fair. Prior to the election and after things were settled, respondents were asked to give a temperature score (0-100) for several public figures, including G. W. Bush. The "before" and "after" columns are those scores that correspond the before and after the election, respectively.

Table **??** presents the averages on the pre and post scores, and the average difference between post and pre (post-pre) by whether the respondents thought the election was fair.

Table 14.2: Mean scores about G. W. Bush before and after 2000 election by whether election was fair

Was election fair?	Score before election	Score after election	After - before
No	55.84	51.50	-4.34
Yes	65.40	66.36	0.96
Total	60.62	58.93	-1.69

14.2.1 Was there change, on average?

We could of course perform a paired t-test

In this case the paired t-test is

$$t = \frac{\bar{d}\sqrt{N}}{s_d} \tag{14.9}$$

$$t = \frac{-1.69\sqrt{100}}{18.44} \tag{14.10}$$

$$t = -0.92 \tag{14.11}$$

Which has a standard error of the difference of 1.844 and results in a p value of 0.36 with $N - 1 = 99$ degrees of freedom.

14.2.2 As-is regression

As before we begin with a regression without centering. This is presented as Model 1 in Table **??**. If we take this model literally, we would believe that on average, the approval for Bush increased nearly 12 points from before the election to after. In fact, this is reflecting the change for persons who scored him at 0 before the election, and thus represents only regression towards the mean.

14.2.3 Regression on pre-centered data

Looking to Model 3 in Table **??**, our intercept again reflects the marginal difference in Table **??** of -1.690, which in this case is not significant; but the standard error is smaller than the t-test. We can see the difference between the as-is regression and the pre-centered data in Figure **??** and Figure **??**.

14.2.4 ANCOVA model: adding a categorical co-variate

Now things get interesting. We can now ask ourselves if believing the election was fair influenced the impact on respondents from before to after the election. Adding in the simple covariate, a dummy variable indicating whether the respondent felt the election was fair, leads to Models 2 and 4 in Table **??**. In both models, the effect of believing the election was fair yields the same

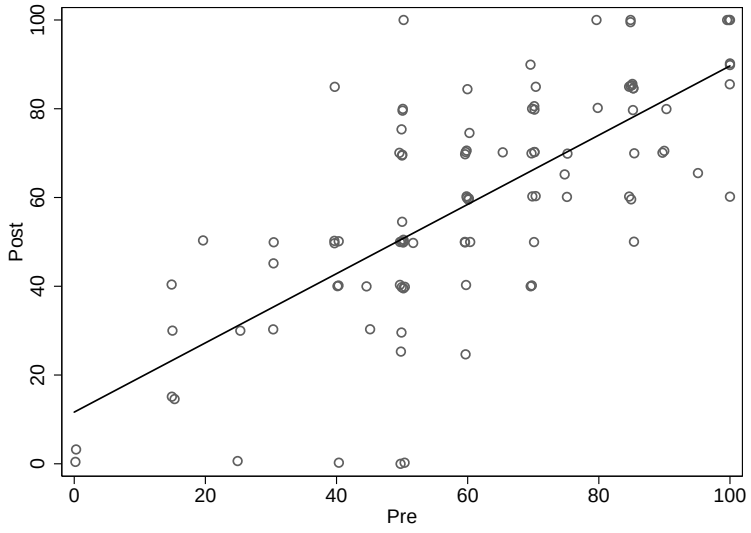

Figure 14.4: Scatter plot of data in Table with regression line (note: small jitter added to points)

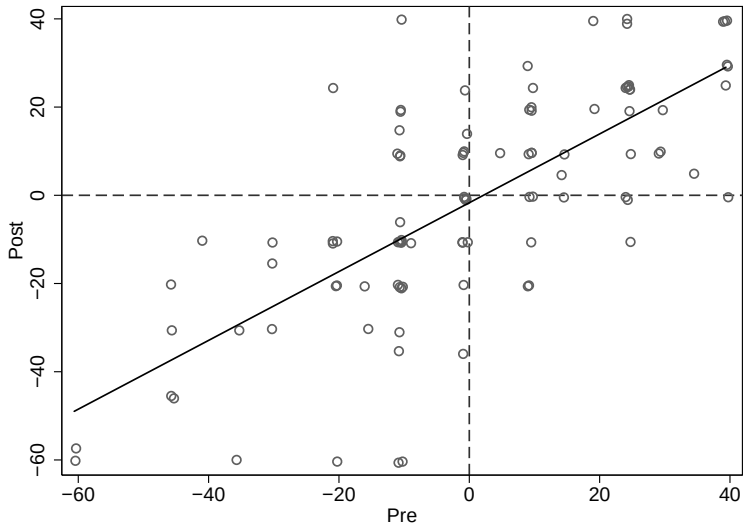

Figure 14.5: Scatter plot of data in Table (centered on pre mean) with regression line (note: small jitter added to points)

result, model fit, etc. However, the intercept in Model 2 would lead us to believe that there was no change for those who felt the election was not fair. However, this again reflects individuals who scored Bush as 0 on the pre score. Using the centering technique, we instead find that for a person who thought the election was unfair, but had the global average on the pretest, their opinion of Bush decreased 5.556 points. Those who believed the election was fair increased -5.556+7.733=2.177 points.

In order to replicate the results in Table **??**, we would need to center pre and post on the pre means specific to the categorical variable. We do this in Model 5 in Table **??**. In that case, the intercept is -4.34, and the effect of fair is 5.3, for a difference specific to those who thought the election was fair of -4.34+5.3 = 0.96, which is reflected in the means table. Unfortunately, by removing the mean differences due to the fairness indicator, we are no longer comparing the same pretest and our results are no longer significant.

Lord's paradox

Which is the correct method? If we centered on the within-group means, we find no effect. If we center on a global mean, we find an effect. This is the crux behind the famous Lord's paradox, named after Frederic M. Lord who called attention to it in 1967. There is no apparent answer, except to be explicit on the research question and use the correct means when centering.

Table 14.3: Models to analyze differences between pre and post election scores for G. W. Bush

Coefficients	Model 1	Model 2	Model 3	Model 4	Model 5
Pre	0.780***	0.746***	0.780***	0.746***	0.746***
	(0.077)	(0.077)	(0.077)	(0.077)	(0.077)
Fair		7.733*		7.733*	5.300
		(3.575)		(3.575)	(3.497)
Intercept	11.640*	9.869	-1.690	-5.556*	-4.340
	(4.997)	(4.974)	(1.781)	(2.500)	(2.472)
Model Statistics					
N	100.000	100.000	100.000	100.000	100.00
F	102.575	55.553	102.575	55.553	47.674
R^2	0.511	0.534	0.511	0.534	0.496
df Regression	1.000	2.000	1.000	2.000	2.000
Sum of Squares Reg.	32542.960	33973.928	32542.960	33973.927	29155.688
df Error	98.000	97.000	98.000	97.000	97.000
Sum of Squares Error	31091.550	29660.582	31091.549	29660.582	29660.582

Model 3 pre and post centered on pretest mean
Model 4 pre and post centered on pretest mean
Model 5 pre and post centered on pretest mean within group
*SE*s in parentheses, $*p < 0.05, * * *p < 0.001$

Chapter 15

Interactions

In many analyses we may consider that the effect of x on y is moderated by a third variable, z. That is to say, the slope of x may be different depending on the level or value of z. These ideas can be tested in a regression framework through interactions. An interaction between two variables, a and b is simply the product of the variables, which can be noted as $a \times b$.

15.1 Interactions between categorical predictors

First, consider a model in which y is a function of two categorical predictors, x, and z. Each predictor takes on two values (0 and 1). Suppose some variable y was a function of both of these variables, and that these variables interacted in some way to produce effects greater than the combination of their marginal effects. That is to say, x has an effect when $x = 1$, z has an effect when $z = 1$, but when x and z both are 1, there is an additional effect.

Consider Table **??**, in which I report the mean of y for different values of x and z, as well as the marginals. We can fit a model to estimate the difference between $x = 0$ and $x = 1$ regardless of the level of z:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Where β_0 is the average of y when $x = 0$ and β_1 is the difference between $\hat{y}|x = 1$ and $\hat{y}|x = 0$, or $\beta_1 = (\hat{y}|x = 1) - (\hat{y}|x = 0)$. Model 1 in Table **??** gives this effect (with some rounding error in the third decimal).

Table 15.1: Means of y across two categorical predictors

	$z = 0$	$z = 1$	All z
$x = 0$	0.449	-1.158	-0.355
$x = 1$	-0.977	0.290	-0.343
All x	-0.264	-0.434	-0.349

$N = 400$, balanced data

Likewise, we can fit a model to estimate the difference between $z = 0$ and $z = 1$ regardless of the level of x:

$$y_i = \beta_0 + \beta_1 z_i + e_i$$

Where β_0 is the average of y when $z = 0$ and β_1 is the difference between $\hat{y}|z = 1$ and $\hat{y}|z = 0$, or $\beta_1 = (\hat{y}|z = 1) - (\hat{y}|z = 0)$. Model 2 in Table **??** gives this effect (with some rounding error in the third decimal).

Note that in both cases, the intercept is the mean of y when the predictors are zero. In this case, x and z are independent, and so the effects in Model 3

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i$$

do not change. Unfortuneatly, the intercept, β_0, does not equal any of the cells.

However, we can compare the mean of y when $x = 0$ and $z = 0$, with the mean of y when $x = 1$ and $z = 0$, with the mean of y when $x = 0$ and $z = 1$, and the mean of y when $x = 1$ and $z = 1$, using an interaction. Thus, we fit Model 4 in Table **??**:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 (x_i \times z_i)$$

In Models 1, 2, and 3, none of the effects were significant. In Model 4, we see that we have an effect for x, and effect for z, and an interaction effect for $x_i \times z_i$. With this model, we are able to reproduce each cell in the table.[1]

$$(\hat{y}|x = 0, z = 0) = \beta_0 = 0.449$$

$$(\hat{y}|x = 1, z = 0) = \beta_0 + \beta_1 = 0.449 + -1.429 = 0.977$$

[1] note: there is some rounding error

$$(\hat{y}|x=0,z=1)=\beta_0+\beta_2=0.449+-1.606=-1.158$$

$$(\hat{y}|x=1,z=1)=\beta_0+\beta_1+\beta_2+\beta_3=0.449+-1.429+-1.606+2.873=0.290$$

Note that interactions can be interpreted in the number of ways that we have variables interacting. For example, β_3 can be interpreted as the way the effect of x changes when z is equal to 1. That is to say, we can think of β_3 as

$$\left(\frac{\Delta y}{\Delta x}|z=0\right)=\beta_1$$

and

$$\left(\frac{\Delta y}{\Delta x}|z=1\right)=\beta_1+\beta_3$$

Similarly, β_3 can be interpreted as the way the effect of z changes when x is equal to 1.

$$\left(\frac{\Delta y}{\Delta z}|x=0\right)=\beta_2$$

and

$$\left(\frac{\Delta y}{\Delta z}|x=1\right)=\beta_2+\beta_3$$

In both cases, it is β_3 that is making the difference.

15.1.1 Example of categorical interactions on math scores

I used the Early Childhood Longitudinal Study, Kindergarten class of 1998-1999 to estimate the effects of gender and race on math scores for kindergartners. The math scores were standardized to allow the slope coefficients to represent changes in standard deviations. Model 1 in Table **??** presents the following model where math is a function of gender:

$$math_i=\beta_0+\beta_1 female_i+e_i$$

Model 2 presents the following model where math is a function of race:

$$math_i=\beta_0+\beta_1 black_i+e_i$$

Table 15.2: Models predicting y as a function of categorical predictors in for data in Table **??**

Coefficients	Model 1	Model 2	Model 3	Model 4
x	0.011		0.011	-1.425***
	(0.129)		(0.129)	(0.151)
z		-0.170	-0.170	-1.606***
		(0.129)	(0.129)	(0.151)
$x \times z$				2.873***
				(0.214)
Intercept	-0.355***	-0.264**	-0.270*	0.449***
	(0.091)	(0.091)	(0.111)	(0.107)
SEs in parentheses, $*p < 0.05, **p < 0.01, ***p < 0.001$				

Model 3 presents the following model where math is a function of gender and race:

$$math_i = \beta_0 + \beta_1 female_i + \beta_2 black_i + e_i$$

and Model 4 presents the following model where math is a function of gender, race, and the interaction of gender and race:

$$math_i = \beta_0 + \beta_1 female_i + \beta_2 black_i + \beta_3 \left(female_i \times black_i\right) + e_i.$$

We first interpret Model 1. Recalling that the intercept, β_0, is always the value of y when all predictors are zero, we interpret that standardized math scores are equal to about 0.014 for *male* students, and that *female* students' math scores are 0.027 standard deviations lower, on average, than *male* students. However, this finding is not statistically signifiant, with a t-value of

$$t = \frac{\beta_1}{SE\left(\beta_1\right)} = \frac{-0.027}{0.019} = -1.42$$

which has a probability, or p-value, of 0.078, which is greater than the threshold of $\alpha = 0.05$. Thus, the effect is not significant. Turning to Model 2, we interpret that standardized math scores are equal to about 0.050 for *non-Black* students, and that *Black* students' math scores are 0.501 standard deviations lower, on average. This finding is statistically signifiant with p-value of less than 0.001, and is marked as such in the table with three stars $(***)$.

Table 15.3: Model predicting kindergartner math scores as a function of gender and race

Coefficients	Model 1	Model 2	Model 3	Model 4
female	-0.027		-0.021	-0.035
	(0.019)		(0.019)	(0.020)
black		-0.501***	-0.500***	-0.570***
		(0.032)	(0.032)	(0.046)
female×black				0.134*
				(0.064)
Intercept	0.014	0.050***	0.060***	0.067***
	(0.014)	(0.010)	(0.014)	(0.014)
Model Statistics				
F	2.018	244.643	122.955	83.456
R^2	0.000	0.022	0.022	0.023
df Regression	1.000	1.000	2.000	3.000
df Error	10694.000	10694.000	10693.000	10692.000
Math score standardized, N= 10696				
SEs in parentheses, $***p < 0.001$				*Source: ECLS-K*

Model 3 presents both variables, $female$ and $black$ as predictors, and we find again that the effect of being female is not significant, but the effect of being Black is, again estimating a gap of about half a standard deviation.

Model 4 is more interesting, introducing the interaction between race and gender. We now have four groups of students and their means: Non-Black males, Non-Black females, Black males, and Black females. The estimated mean of Non-Black males is calculated as

$$\hat{y} = \beta_0 + \beta_1 (female) + \beta_2 (black) + \beta_3 (female \times black)$$

$$\hat{y} = \beta_0 + \beta_1 (0) + \beta_2 (0) + \beta_3 (0 \times 0)$$

$$\hat{y} = \beta_0$$

$$\hat{y} = 0.067$$

The intercept is now the mean of non-Black males, with an average score of 0.067. The mean of Non-Black females is

$$\hat{y} = \beta_0 + \beta_1 (female) + \beta_2 (black) + \beta_3 (female \times black)$$

$$\hat{y} = \beta_0 + \beta_1 (1) + \beta_2 (0) + \beta_3 (1 \times 0)$$
$$\hat{y} = \beta_0 + \beta_1$$
$$\hat{y} = 0.067 + -0.035 = 0.032$$

However, since β_1 is not significant, the difference between non-Black males and females is not statistically significant. Turning to Black males, we find that the predicted mean is

$$\hat{y} = \beta_0 + \beta_1 (female) + \beta_2 (black) + \beta_3 (female \times black)$$
$$\hat{y} = \beta_0 + \beta_1 (0) + \beta_2 (1) + \beta_3 (0 \times 1)$$
$$\hat{y} = \beta_0 + \beta_2$$
$$\hat{y} = 0.067 + -0.570 = -0.503$$

Finally, the mean for Black females is

$$\hat{y} = \beta_0 + \beta_1 (female) + \beta_2 (black) + \beta_3 (female \times black)$$
$$\hat{y} = \beta_0 + \beta_1 (1) + \beta_2 (1) + \beta_3 (1 \times 1)$$
$$\hat{y} = \beta_0 + \beta_1 + \beta_2 + \beta_3$$
$$\hat{y} = 0.067 + -0.035 + -0.57 + 0.134 = -0.404$$

Thus, because of the interaction between $female$ and $black$, we find that among Black students, the gap in math scores is less for females than it is for males. That is to say, gender moderates the effect of race on math scores.

15.2 Interactions between a categorical and a continuous predictor

More often than interactions between categorical predictors is the wish to see if the effect of a continuous predictor is moderated by membership in a particular group. For example, suppose we had three groups of observations, Group 1, Group 2, and Group 3, and that we had some outcome variable y and predictor of interest, x. Such a data is presented in Table **??**.

We then plot each group onto the same scatterplot in Figure **??**, with Group 1 as circles, Group 2 as triangles, and Group 3 as squares. Overall, it appears that there is a negative relationship between x and y, but closer

Table 15.4: Small dataset of random variables from three groups

Group 1		Group 2		Group 3	
y	x	y	x	y	x
30.97	8.14	70.69	8.98	16.81	18.98
18.97	9.02	74.53	8.91	21.07	18.91
28.00	8.16	79.28	11.20	12.09	21.20
32.02	10.58	66.17	8.22	16.84	18.22
42.92	10.82	69.19	10.47	6.37	20.47
30.26	7.78	61.24	6.11	24.59	16.11
37.75	10.79	82.44	11.37	14.20	21.37
28.87	8.34	72.89	10.09	12.37	20.09
33.99	10.70	79.51	11.89	8.18	21.89
41.90	13.50	80.01	12.35	5.93	22.35

inspection reveals that Groups 1 and 2 appear to have a positive relationship between x and y. How can we decompose this information? First, we create dichotomous indicators d_2 and d_3, where

$$d_2 = \begin{cases} 1 & \text{if } Group = 2; \\ 0 & \text{otherwise.} \end{cases}$$

and

$$d_3 = \begin{cases} 1 & \text{if } Group = 3; \\ 0 & \text{otherwise.} \end{cases}$$

using Group 1 as the reference group. We then explore the data with a series of models in Table **??**. The first model is simply a bivariate model of y on x,

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

The second model adjusts the intercept of the bivariate relationship by entering the dichotomous indicators d_2 and d_3,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_{2i} + \beta_3 d_{3i} + e_i$$

and finally, the third model enters interactions of x with d_2 and d_3,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_{2i} + \beta_3 d_{3i} + \beta_4 \left(d_{2i} \times x_i\right) + \beta_5 \left(d_{3i} \times x_i\right) + e_i$$

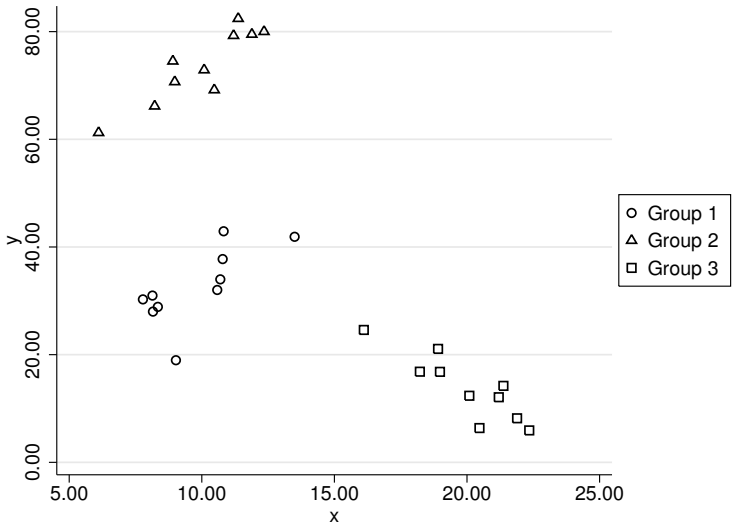

Figure 15.1: Scatter plot of data in Table **??**

Table 15.5: Models from data in Table **??**

Coefficients	Model 1	Model 2	Model 3
x	-3.238***	0.998	2.784***
	(0.737)	(0.671)	(0.742)
d_2		40.855***	36.338**
		(2.938)	(10.220)
d_3		-28.875***	64.276***
		(7.434)	(15.833)
$d_2 \times x$			0.422
			(1.020)
$d_3 \times x$			-5.578***
			(1.020)
Intercept	82.854***	22.799**	5.328
	(10.447)	(6.887)	(7.373)
SEs in parentheses, $**p < 0.01$, $***p < 0.001$			

Model 1 in Table **??** reports a negative relationship between x and y. This is confirmed in the first panel of Figure **??**, where we see the fit line sloping downward. While this describes the bivariate relationship, our inspection of the data revealed a positive relationship for Groups 1 and 2.

We then fit Model 2, and work to understand exactly what the effects of d_2 and d_3 are doing. Starting with Group 1. We can find the group specific regression model for Group 1 by plugging values

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 d_2 + \beta_3 d_3$$

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 0 + \beta_3 0$$

$$\hat{y} = \beta_0 + \beta_1 x$$

and we find that the terms for d_2 and d_3 drop off.

We then move to Group 2. We can again eliminate some coefficients by simply plugging in numbers to the regression model

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 d_2 + \beta_3 d_3$$

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 1 + \beta_3 0$$

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2$$

Since β_0 and β_2 are constants, we can gather terms to create an adjusted intercept

$$\hat{y} = (\beta_0 + \beta_2) + \beta_1 x$$

Therefore, the intercept for Group 2 in Model 2 in Table **??** is $\beta_0 + \beta_2$. We can do the same for Group 3:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 d_2 + \beta_3 d_3$$

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 0 + \beta_3 1$$

$$\hat{y} = \beta_0 + \beta_1 x + \beta_3$$

$$\hat{y} = (\beta_0 + \beta_3) + \beta_1 x$$

and we find that the intercept for Group 3 in Model 2 in Table **??** is $\beta_0 + \beta_3$. Yet, in each case, the slope is β_1. That means that each group should have its own intercept, but the same slope. That is, the model for Group 1 is

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\hat{y} = 22.799 + 0.998x$$

and Group 2 is

$$\hat{y} = (\beta_0 + \beta_2) + \beta_1 x$$

$$\hat{y} = (22.799 + 40.855) + 0.998x$$

$$\hat{y} = 63.564 + 0.998x$$

and Group 3 is

$$\hat{y} = (\beta_0 + \beta_3) + \beta_1 x$$

$$\hat{y} = (22.799 + -28.875) + 0.998x$$

$$\hat{y} = -6.076 + 0.998x$$

This draws three parallel lines, and we can see that in the second panel of Figure **??**. The problem with Model 2, however, is that the slope is not significant. We then move to Model 3 where we interact x with d_2 and d_3.

Understanding models with such interactions is tractable when we perform the same "plug and chug" procedure as in the non-interaction models. First, we can discover the regression line for Group 1:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 d_2 + \beta_3 d_3 + \beta_4 (d_2 \times x) + \beta_5 (d_3 \times x)$$

Remember for Group 1, d_2 and d_3 both equal 0, so

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 0 + \beta_3 0 + \beta_4 (0 \times x) + \beta_5 (0 \times x)$$

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\hat{y} = 5.328 + 2.784x$$

Simple as that, all other slopes get zeroed out. Next, we move to Group 2

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 1 + \beta_3 0 + \beta_4 (1 \times x) + \beta_5 (0 \times x)$$

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 + \beta_4 x$$

Now, we have two constants, β_0 and β_2, and two slopes for x, β_1 and β_4. So, we can collect terms

$$\hat{y} = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x$$

Thus, we can view β_2 as the Group 2 adjustment for the intercept, and β_4 as the adjustment to the slope of x for Group 2. That means that the Group 2 specific regression model is

$$\hat{y} = (5.328 + 36.338) + (2.784 + 0.422)\, x$$

$$\hat{y} = 41.666 + 3.206x$$

We do the same steps for Group 3

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 0 + \beta_3 1 + \beta_4 (0 \times x) + \beta_5 (1 \times x)$$

$$\hat{y} = \beta_0 + \beta_1 x + \beta_3 + \beta_5 x$$

$$\hat{y} = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)\, x$$

$$\hat{y} = (5.238 + 64.276) + (2.784 + -5.578)\, x$$

$$\hat{y} = 69.514 + -2.794x$$

This means that Group 2 does indeed have a negative slope. We can see these lines in the third panel of Figure **??**.

15.2.1 Example of a continuous-categorical interaction on math scores

We return to the ECLS-K data on math scores for kindergartners. In these models we predict math scores based on two variables. The first variable is a dichotomous variable for school sector, *private*, coded 0 for public and 1 for private school. The next variable is *pared*, which is a standardized (z) score of parental education with a mean 0 and standard deviation of 1.

Model 1 in Table **??** models math as a function of school sector

$$y_i = \beta_0 + \beta_1 private_i + e_i$$

Model 2 models math as a function of parental education

$$y_i = \beta_0 + \beta_1 pared_i + e_i$$

Model 3 models math as a function of school sector and parental education

$$y_i = \beta_0 + \beta_1 private_i + \beta_2 pared_i + e_i$$

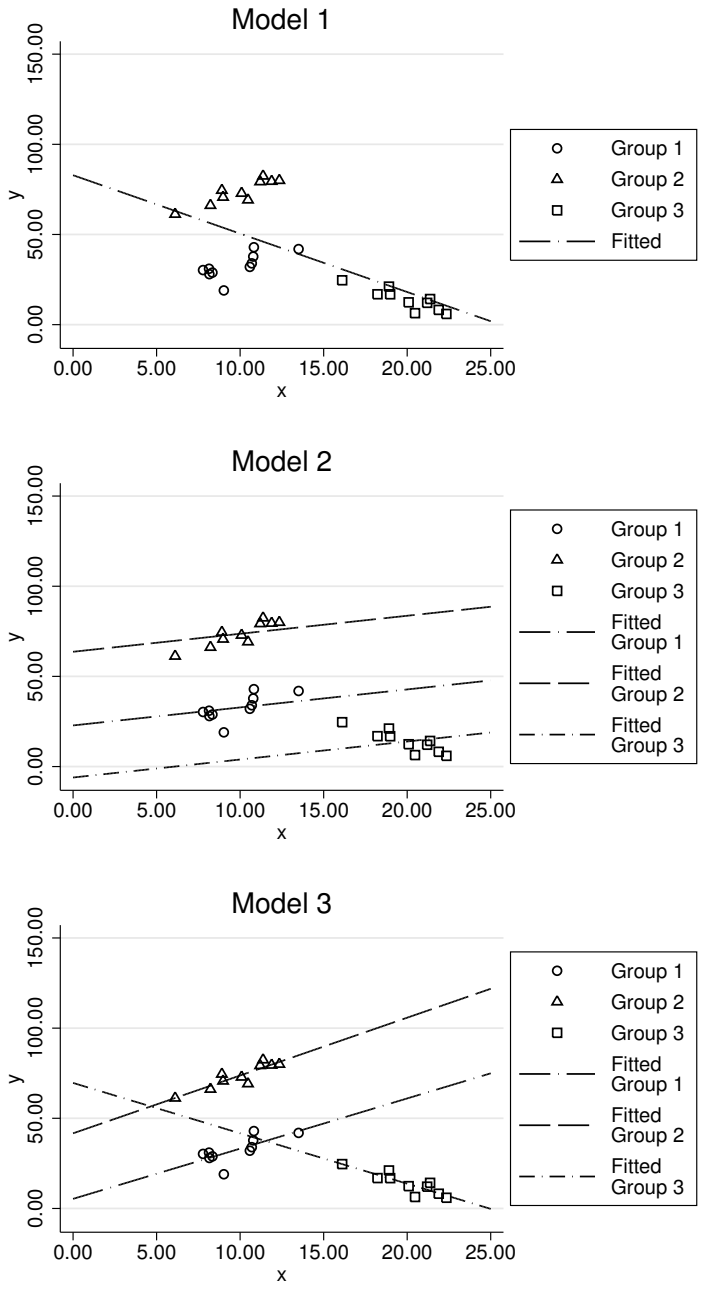

Figure 15.2: Scatter and fit plots of Models from data in Table **??**

Table 15.6: Models predicting math scores as a function of school sector and parental education

Coefficients	Model 1	Model 2	Model 3	Model 4
private	0.348***		0.174***	0.209***
	(0.023)		(0.023)	(0.024)
pared		0.349***	0.328***	0.349***
		(0.011)	(0.011)	(0.012)
private × pared				-0.113***
				(0.028)
Intercept	-0.077***	0.000	-0.039***	-0.036***
	(0.011)	(0.009)	(0.010)	(0.010)
Model Statistics				
F	228.079	1100.807	582.391	394.440
R^2	0.021	0.093	0.098	0.100
df Regression	1.000	1.000	2.000	3.000
df Error	10694.000	10694.000	10693.000	10692.000
Math score and *pared* standardized, N= 10696				
SEs in parentheses, $***p < 0.001$				*Source: ECLS-K*

and finally, Model 4 tests whether school sector moderates the relationship between parental education and math scores

$$y_i = \beta_0 + \beta_1 private_i + \beta_2 pared_i + \beta_3 \left(private_i \times pared_i\right) + e_i.$$

We see in Model 1 that students who attend private schools have, on average, 0.348 standard deviations higher math scores than students who attend public schools. In Model 2, the data report that for each standard deviation increase in parental education, that math scores also increase about 0.349 standard deviations.

However, these effects are somewhat spurious, since when we combine the effects in Model 3, we find the effect of school sector reduced to almost half, while the effect of parental education remains. Model 4 offers some interesting information, however.

If we look at the model, we can calculate the "public" school relationship between parental education and math

$$\hat{y} = \beta_0 + \beta_1 private + \beta_2 pared + \beta_3 \left(private \times pared\right)$$

$$\hat{y} = \beta_0 + \beta_1 0 + \beta_2 pared + \beta_3 (0 \times pared)$$
$$\hat{y} = \beta_0 + \beta_2 pared$$
$$\hat{y} = -0.036 + 0.349 pared$$

which shows that in *public* schools, parental education is a powerful predictor. In private schools, the effect of parental education is

$$\hat{y} = \beta_0 + \beta_1 1 + \beta_2 pared + \beta_3 (1 \times pared)$$
$$\hat{y} = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) pared$$
$$\hat{y} = (-0.036 + 0.209) + (0.349 + -0.113) pared$$
$$\hat{y} = 0.173 + 0.236 pared$$

which means that in *private* schools, the effect of parental education is less, 0.236, than in *public* schools, 0.349. However, the intercept is also larger in private schools.

15.3 Interactions between continuous predictors

Interactions between continuous predictors are often tricky to understand. My strategy is to fit the model with continuous predictors, but then convert one variable into a quasi-categorical variable during interpretation. My method to make this easier is to convert one of the variables in used in the interaction into a z score; usually the proposed moderating variable.

The model for a continuous interaction is like any other model with interactions. The outcome y is predicted with (along with controls) two variables x and z, and the product of x and z, $x \times z$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 (x_i \times z_i) + e_i$$

The trick to visualizing how the moderation affects the regression lines is to pick three values for z, and if z is standardized, natural candidates are (a) one standard deviation below the mean, -1, (b) the mean, 0, and (c) one standard deviation above the mean, 1. Thus, the regression line for one standard deviation below the mean is

$$\hat{y} = \beta_0 + \beta_1 + \beta_2 (-1) + \beta_3 (x \times -1)$$

$$\hat{y} = \beta_0 + \beta_1 x - \beta_2 - \beta_3 x$$

$$\hat{y} = (\beta_0 - \beta_2) + (\beta_1 - \beta_3)\, x$$

the regression line for the mean is

$$\hat{y} = \beta_0 + \beta_1 + \beta_2\,(0) + \beta_3\,(x \times 0)$$

$$\hat{y} = \beta_0 + \beta_1 x$$

and the regression line for one standard deviation above the mean is

$$\hat{y} = \beta_0 + \beta_1 + \beta_2\,(1) + \beta_3\,(x \times 1)$$

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 + \beta_3 x$$

$$\hat{y} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x.$$

15.3.1 Example of a continuous-continuous interaction on income

In this example we will use data from the 2008 GSS to estimate how income is affected by the prestige of someone's job verses their level of education. Our running hypothesis is that the prestige moderates the impact of education on income.

Income in the GSS is measured categorically, with ≥ 25 thousand dollars as the top category. Just as we did in **??** (see footnote), I mid-pointed the categories to create a continuous variable. Table **??** presents four models. The first model predicts income as a function of the standardized prestige, z

$$y_i = \beta_0 + \beta_1 prestige_i + e_i$$

Model 2 predicts income as a function of years of education

$$y_i = \beta_0 + \beta_1 educ_i + e_i$$

Model 3 predicts income as a function of both prestige and years of education

$$y_i = \beta_0 + \beta_1 prestige_i + \beta_2 educ_i + e_i$$

and finally, Model 4 includes the interaction of prestige and years of education

$$y_i = \beta_0 + \beta_1 prestige_i + \beta_2 educ_i + \beta_3\,(prestige_i \times educ_i) + e_i.$$

Table 15.7: Models predicting income in thousands as a function of occupational prestige and education

Coefficients	Model 1	Model 2	Model 3	Model 4
prestige	1.529***		0.813***	4.767***
	(0.120)		(0.138)	(0.594)
educ		0.629***	0.460***	0.448***
		(0.041)	(0.048)	(0.047)
prestige × educ				-0.273***
				(0.040)
Intercept	22.250***	13.515***	15.940***	16.532***
	(0.120)	(0.570)	(0.663)	(0.662)
Model Statistics				
F	163.625	238.865	131.906	105.361
R^2	0.068	0.093	0.106	0.124
df Regression	1.000	1.000	2.000	3.000
df Error	2234.000	2332.000	2232.000	2231.000
prestige standardized, $N = 2236$				
SEs in parentheses, $***p < 0.001$				*Source: GSS 2008*

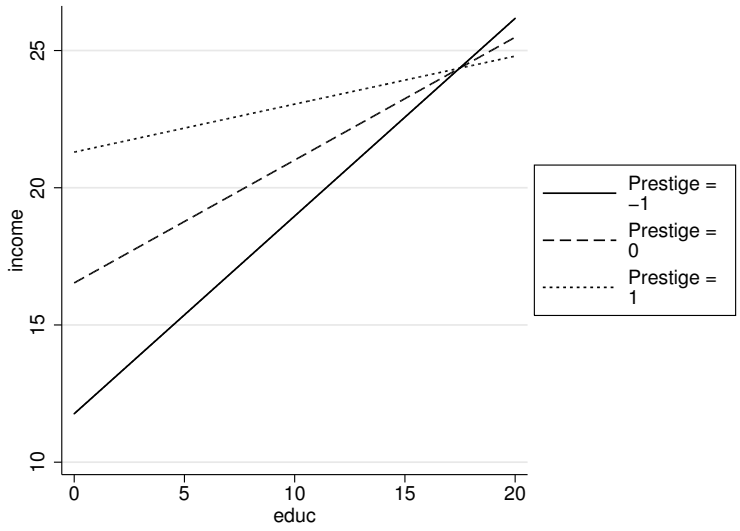

Figure 15.3: Income as a function of education by prestige of occupation from Model 4 in Table **??**

Looking at Model 1, and noting the prestige is a standardized variable and that income is measured in thousands, we see that for each standard deviation increase in prestige, annual income increases by about 1,529 dollars, with an average income of about 22,250 dollars. An examination of Model 2 indicates that for each year of education, annual income increases by about 629 dollars.

Model 3 estimates the effect of education, holding constant the effects of occupational prestige. We see the effect of prestige is smaller when we control for education, this time showing only a 813 dollar increase for each standard deviation increase in prestige. The effect is also reduced, showing only a 460 dollar increase for each year of education.

We now turn to Model 4, which includes the interaction of $(prestige \times educ)$, we can focus on the effect of education for different levels of prestige. Since prestige is standardized, we can pick three levels of prestige: (a) one standard deviation below the mean, -1, (b) the mean, 0, and (c) one standard deviation above the mean, 1.

First, we figure out the model for education where prestige = -1

$$\hat{y} = \beta_0 + \beta_1 prestige + \beta_2 educ + \beta_3 (prestige \times educ)$$

$$\hat{y} = \beta_0 + \beta_1 - 1 + \beta_2 educ + \beta_3 (-1 \times educ)$$

$$\hat{y} = \beta_0 - \beta_1 + \beta_2 educ - \beta_3 educ$$

$$\hat{y} = (\beta_0 - \beta_1) + (\beta_2 - \beta_3) educ$$

$$\hat{y} = (16.532 - 4.767) + (0.448 - -0.273) educ$$

$$\hat{y} = 11.765 + 0.721 educ$$

Next, the model for education where prestige = 0 (the mean)

$$\hat{y} = \beta_0 + \beta_1 prestige + \beta_2 educ + \beta_3 (prestige \times educ)$$

$$\hat{y} = \beta_0 + \beta_1 0 + \beta_2 educ + \beta_3 (0 \times educ)$$

$$\hat{y} = \beta_0 + \beta_2 educ$$

$$\hat{y} = 16.532 + 0.448 educ$$

Finally, the model for education where prestige = 1

$$\hat{y} = \beta_0 + \beta_1 prestige + \beta_2 educ + \beta_3 (prestige \times educ)$$

$$\hat{y} = \beta_0 + \beta_1 1 + \beta_2 educ + \beta_3 (1 \times educ)$$

$$\hat{y} = \beta_0 + \beta_1 + \beta_2 educ + \beta_3 educ$$

$$\hat{y} = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) educ$$

$$\hat{y} = (16.532 + 4.767) + (0.448 + -0.273) educ$$

$$\hat{y} = 21.299 + 0.175 educ$$

Note that, substantively, when occupational prestige is low (-1), the intercept is also low, but the effect of education is strong. When prestige is average (0), the intercept is higher, signaling that pay is somewhat better, but the returns on education are lower (lower slope). Finally, when prestige is high (1), the intercept is even higher and the effects of education are lower still. Of course, this all makes sense. These patterns are visualized in Figure **??**.

Chapter 16

Heteroskedasticity

Recall the typical regression model

$$y_i = \beta_0 + \sum_p \beta_p x_{ip} + e_i$$

We make the assumption that

$$\text{var}\,(e_i) = \sigma^2$$

Note how there is no subscript on sigma squared. We assume that the variance is constant for each case i. That is, we assume *homoskedasticity*. We can visualize homoskedasticity in Figure **??**, where the residuals of each prediction have a relatively constant distribution.

Heteroskedasticity, where σ^2 is not constant across values of our predictions, can cause standard errors to be biased, making hypothesis tests difficult to interpret. We can visualize heteroskedasticity in Figure **??**, where the residuals of each prediction are not consistent.

16.1 Why non-constant variance is a problem

In bivariate regression, we can represent the variance (the square of the standard error) of the slope as equation (**??**)

$$\text{var}\,(\beta_1) = \frac{\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

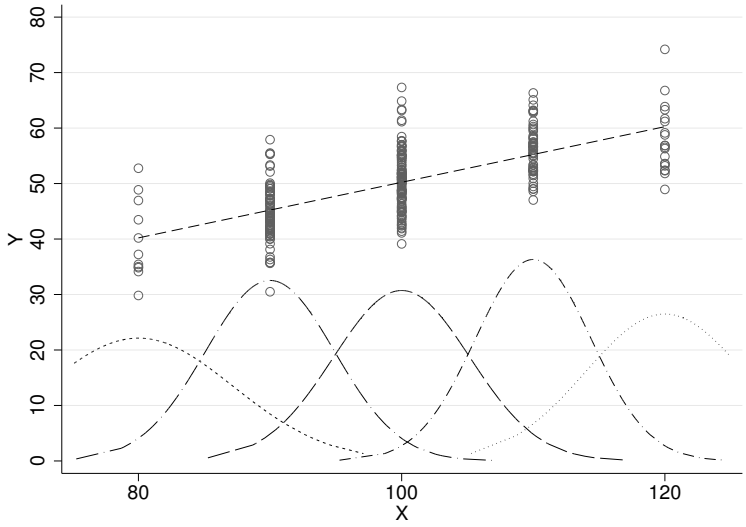

Figure 16.1: Regression with relatively constant variance

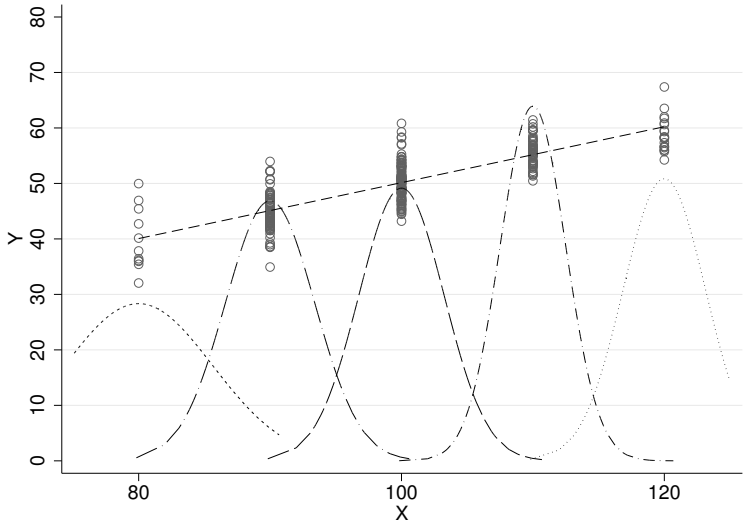

Figure 16.2: Regression with non-constant variance

where σ^2 is the variance of the residuals. Mechanically, it just means that the standard error is a function of the variance of the residuals standardized by the sum of squares in the predictor.

We do not actually know the true variance of the residuals, so we estimate the variance of the residuals with

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{e}_i^2}{N - p} \tag{16.1}$$

and assume that it is constant for all predicted values, $\hat{y}_i$. When the variance of the residuals is not constant, then the formula for the variance of the slope is more complicated

$$\text{var}\left(\beta_1\right) = \frac{\sum_{i=1}^{N} \left(\left(x_i - \bar{x}\right)^2 \sigma_i^2\right)}{\left(\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2\right)^2} \tag{16.2}$$

Now, the variance of the residual is unique for each case (note the subscript i on σ_i^2) and is scaled by the squared deviation from the mean. We a critical problem: since we only have one observation for each prediction, there is no way to estimate the variance of the errors for that prediction. OLS regression software does not know this, so estimates of the standard errors are potentially biased.

The direction of this bias is not consistently reported in statistical texts. Econometricians generally think that the variances are overestimated, leading to difficulty in rejecting hypotheses. This would be annoying, but somewhat acceptable with large samples. However, others believe that the direction of the bias is unknown, so in some cases the null hypothesis rejection is false.

16.2 Testing for heteroskedasticity

When the issue is introduced, often people show a figure like Figure **??**. However, it often doesn't look this straightforward. More often, you need theory to tell you that you should expect heteroskedastic relationships and you should plan ahead for them. We again return to charity and household income; see Tables **??** and **??** in the discussion of logs in **??**.

We expect that the variance of giving is going to be lower for lower-income households (they just do not have as much) and higher for richer households (some give, some do not, some give a lot). We can examine a regression

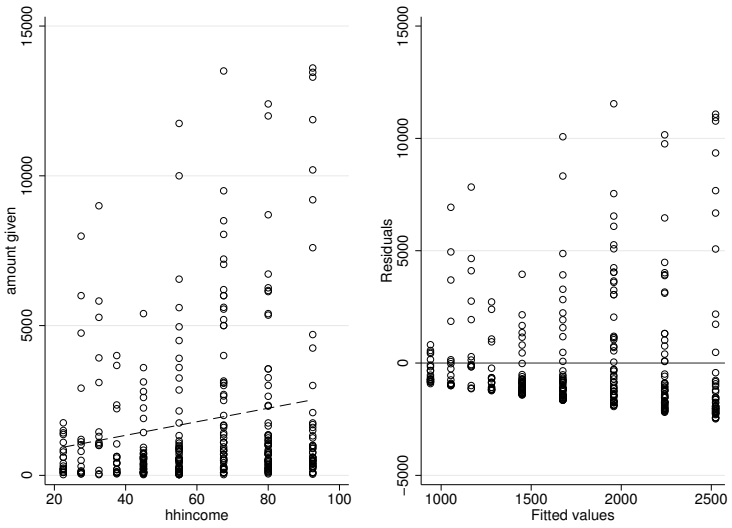

Figure 16.3: Regression of charity on income (left), residuals by fitted values (right)

of charity on income, visually, and inspect a plot of the residuals by the prediction, $\hat{y}_i$. I do this in Figure **??**. Notice how the "spread" of residuals differs based on the prediction.

How can we detect heterskedasticity without having to look at graphics?

One simple test is called the Breusch-Pagan test. It tests for a function to the variance of the residuals based on a set of predictors. The logic of the test is that if there is heteroskedasticity, then the magnitude of the residuals (and thus the variance) would be dependent on some set of predictors. In other words, if we have a regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

we then suppose some variance function

$$\frac{e_i^2}{\sigma^2} = \gamma_0 + \gamma_1 z_{1i} \dots \gamma_k z_{ki} + v_i$$

Where e_i is the residual of each case and σ^2 is the variance of all the residuals. The reason for the denominator is to standardize the residuals to a z-score, which makes the results of the test applicable to a χ^2 distribution. If it is not

immediately obvious how this standardization occurs, recall that a z-score takes the form of

$$z_i = \frac{x_i - \bar{x}}{s}$$

where the numerator, $x_i - \bar{x}$, is the difference between an observed case and the mean and the denominator, s, is the standard deviation. Now consider that a residual is just the difference between an observed case and the predicted value, $e_i = y_i - \hat{y}$, that has a mean of 0.

Since a predicted value is just a conditional mean, we can see the analogue to a z-score numerator. In this notation, we consider σ^2 to be the variance of the residuals. Therefore, σ is the standard deviation of the residuals. With all of these pieces, we can see that the dependent variable of our variance function, $\frac{e_i^2}{\sigma^2}$, is basically a z-score.

By default, we use the predicted values of y as our predictor for the variance function. This is a convenient thing to do when we have several predictors since $\hat{y}$ is just a linear combination of the predictors anyway.

The test procedure is this. Suppose we fit a model to the data

$$y_i = \beta_0 + \sum_p \beta_p x_{pi} + e_i$$

and predicted the fitted values

$$\hat{y}_i = \beta_0 + \sum_p \beta_p x_{pi}$$

and estimated the variance of the residuals, $\hat{\sigma}^2$, using the models sum of squared errors, SSE_{model}:

$$\hat{\sigma}^2 = \frac{SSE_{model}}{N} \tag{16.3}$$

We then calculate a standardized residual

$$u_i = \frac{e_i^2}{\hat{\sigma}^2} \tag{16.4}$$

and fit an auxiliary model using the fitted values as a predictor

$$u_i = \gamma_0 + \gamma_1 \hat{y}_i + v_i$$

We then find the sum of squares regression for this auxiliary regression model ($SSR_{auxiliary}$) and divide it by 2:

$$\theta = \frac{SSR_{auxiliary}}{2} \tag{16.5}$$

This statistic has a χ^2 distribution with 1 degree of freedom. Thus, it can be statistically evaluated. If this statistic is greater than about 3.84, then the model is likely heterskedastic.

The logic is that if the sum of squares for the regression in this auxiliary model is large, then there is a relationship between the *magnitude* of the residual and the predictor(s) of the original model.

16.2.1 Breusch-Pagan example

For example, Model 1 in Table **??** produced a set of residuals that were converted in a standardized outcome. The auxiliary model produced a sum of squares regression of 85.570, see Table **??**. We then calculate our test

$$\theta = \frac{SSR_{auxiliary}}{2} = \frac{85.570}{2} = 42.785$$

Which has a p-value of 0.00000000003; statistically significant.

This is a problem for the analysis as it is. While the point estimates (the slope) are trustworthy, the standard errors are not (they may be too big, they may be too small) making hypothesis testing difficult. The rest of this chapter is devoted to what you can do to generate appropriate standard errors while essentially preserving the model.

16.3 Robust standard errors

The first method to combat heterskedasticity is to directly address the standard errors. Thus, we want to estimate robust standard errors. Recall the issue is that with non-constant variance, the variance estimate of the bivariate slope is

$$\text{var}\left(\beta_1\right) = \frac{\sum_{i=1}^{N}\left(\left(x_i - \bar{x}\right)^2 \sigma_i^2\right)}{\left(\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2\right)^2} \tag{16.6}$$

Recall that the problem was that there was no way to estimate the variance of a single residual σ_i^2. However, Hal White suggested that we substitute with each case's squared residual instead to produce

$$\text{var}(\beta_1) = \frac{\sum_{i=1}^{N}\left((x_i - \bar{x})^2 e_i^2\right)}{\left(\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^2} \tag{16.7}$$

Why would e_i^2 make a good substitute for σ_i^2? Remember that a population variance is the sum of squared deviations from the mean divided by the number in the population

$$\text{var}(x) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

Since the expected mean residual is 0, and we only have 1 observation,

$$\text{var}(e) = \frac{\sum_{i=1}^{1}(e_i - 0)^2}{1} = \frac{\sum_{i=1}^{1}(e_i)^2}{1} = e_i^2$$

Thus, the square of the residual is a good choice. All robust standard errors do is fit the original model, save the residuals, and reuse them to calculate new variances. Model 2 in Table **??** presents robust standard errors. Note that in this case, the standard error of the slope is larger.

16.3.1 Robust standard errors in matrix notation

Recall from **??** that the variance of the slopes is expressed as

$$\mathbf{V}\left(\hat{\mathbf{b}}\right) = \left(\frac{\mathbf{1}}{\mathbf{N}}\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\sigma^{\mathbf{2}}$$

This expression also makes the assumption of constant variance. The expression is more complicated in its general form without this assumption

$$\mathbf{V}\left(\hat{\mathbf{b}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'\left(\mathbf{V}\left(\mathbf{y}\right)\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}} \tag{16.8}$$

When we assume constant variance and uncorrelated errors, then

$$\mathbf{V}\left(\mathbf{y}\right) = \sigma^{\mathbf{2}}I \tag{16.9}$$

Where I is a matrix of 1s in the diagonal and 0s otherwise. For example, if we had 4 observations, then

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

When we multiply this matrix by the scalar such as σ^2, then the diagonals become σ^2 and the off diagonals stay 0:

$$\sigma^{\mathbf{2}} I = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

This all makes $\mathbf{V}(\mathbf{y})$ a very simple entity of the same scalar along the diagonals, allowing us to cancel out most of the $\mathbf{X'X}$s. This leaves us with $\sigma^{\mathbf{2}}(\mathbf{X'X})$ to express the variances of our slopes. However, when we have non-constant variances then we are no longer dealing with a single scalar. Instead, we theoretically have a variance for each case, σ_i^2. This makes

$$\mathbf{V}(\mathbf{y}) = \text{diag}\left\{\sigma_{\mathbf{1}}^{\mathbf{2}} \dots \sigma_{\mathbf{N}}^{\mathbf{2}}\right\} \tag{16.10}$$

The "diag" notation is a short hand for saying the following elements are along the diagonal on a matrix that is otherwise 0s. For example, in our 4 case scenario,

$$\mathbf{V}(\mathbf{y}) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

Of course, as we noted before, we do not know σ_i^2. Instead, as White suggested, we define a matrix where the diagonal is the square of the residuals

$$\mathbf{\Sigma} = \text{diag}\left\{\mathbf{e}_{\mathbf{1}}^{\mathbf{2}} \dots \mathbf{e}_{\mathbf{N}}^{\mathbf{2}}\right\} \tag{16.11}$$

We then use that for our variance estimation

$$\mathbf{V}(\mathbf{b_{robust}}) = (\mathbf{X'X})^{-\mathbf{1}} \mathbf{X'\Sigma X} (\mathbf{X'X})^{-\mathbf{1}} \tag{16.12}$$

This is often called a "sandwich" estimator because we sandwich Σ in between two instances of $(\mathbf{X'X})^{-\mathbf{1}}$.

16.4 Generalized least squares

The second approach to addressing heterskedasticity is to address the variance function directly. In this case, we theorize that the magnitude of the residuals, and thus the variance of them, is correlated with x. For our example, we believe that the variance of giving to charity increases with household income. Thus, we can weight the data in such a way to compensate for this relationship.

We can estimate our models by way of weighted least squares (WLS) and take two approaches to doing so: (a) we can assume we "know" a simple relationship between our errors and a predictor or (b) we can estimate the variance function to make a feasible model.

16.4.1 Weighted least squares

One possible way to think about this relationship is that the observed variance for a residual is a function of the "true" variance proportional to the magnitude of our predictor

$$\text{var}\,(e_i) = \sigma^2 x_i$$

Following White's suggestion that the best guess of the variance of a single residual is the square of the residual, we can easily see that our observed residuals, , are a product of the true residual and the square root of x

$$e_i^2 = u_i^2 x_i \tag{16.13}$$

$$e_i = u_i \sqrt{x_i}$$

$$\frac{e_i}{\sqrt{x_i}} = u_i$$

Thus, the only logical thing to do to estimate the true error, u_i, is to divide the entire model (including the intercept) by $\sqrt{x_i}$

$$\frac{y_i}{\sqrt{x_i}} = \beta_0 \frac{1}{\sqrt{x_i}} + \beta_1 \frac{x_i}{\sqrt{x_i}} + \frac{e_i}{\sqrt{x_i}} \tag{16.14}$$

In effect, this procedure weights the data by

$$w_i = \frac{1}{\sqrt{x_i}} \tag{16.15}$$

The results of this model are presented as Model 3 in Table **??**.

16.4.2 Weighted least squares in matrix form

We again return to matrix algebra to see how this generalizes to the least squares linear model. If we make the strong assumption that variances of the errors are proportional to another variable like $\sigma_i^2 = \sigma^2/w_i^2$, then we know that the likelihood of the model is

$$L\left(\mathbf{b}, \sigma^2\right) = \frac{1}{(2\pi)^{\frac{N}{2}}} e^{\left(-\frac{1}{2}(\mathbf{y}-\mathbf{Xb})'\mathbf{\Sigma^{-1}}(\mathbf{y}-\mathbf{Xb})\right)} \tag{16.16}$$

where

$$\Sigma = \sigma^2 \times \text{diag}\left\{\frac{1}{a_i^2} \ldots \frac{1}{w_N^2}\right\} = \sigma^{\mathbf{2}}\mathbf{W^{-1}} \tag{16.17}$$

With this likelihood function we can get to the formula for the coefficients that maximizes the likelihood and minimizes the squares of the weighted errors, $\sum_{i=1}^{N}\left(w_i e_i\right)^2$

$$\mathbf{b_{WLS}} = \left(\mathbf{X'WX}\right)^{\mathbf{-1}}\mathbf{X'Wy} \tag{16.18}$$

where

$$\mathbf{W} = \text{diag}\left\{w_1 \ldots w_N\right\} \tag{16.19}$$

Compare this the OLS estimator, (**??**)

$$\mathbf{b} = \left(\mathbf{X'X}\right)^{\mathbf{-1}}\mathbf{X'y}$$

It is almost the same thing except that the weight matrix is sandwiched in there. The variance-covariance matrix of these slopes is

$$\mathbf{V}\left(\mathbf{b_{WLS}}\right) = \sigma^{\mathbf{2}}\left(\mathbf{X'WX}\right)^{\mathbf{-1}} \tag{16.20}$$

Compare this to the OLS variances

$$\mathbf{V}\left(\mathbf{b_{OLS}}\right) = \sigma^{\mathbf{2}}\left(\mathbf{X'X}\right)^{\mathbf{-1}} \tag{16.21}$$

Again, it is almost the same thing except for the addition of the weight matrix. However, you may wish to use the robust-weighted variances, which like White's robust variances is a somewhat complicated expression

$$\mathbf{V}\left(\mathbf{b_{WLS,robust}}\right) = \left(\mathbf{X'WX}\right)^{\mathbf{-1}}\mathbf{X'W\Sigma WX}\left(\mathbf{X'WX}\right)^{\mathbf{-1}} \tag{16.22}$$

However, you should be able to see the resemblance to the typical robust errors we just covered in **??**.

$$\mathbf{V}\left(\mathbf{b_{robust}}\right) = \left(\mathbf{X'X}\right)^{\mathbf{-1}}\mathbf{X'\Sigma X}\left(\mathbf{X'X}\right)^{\mathbf{-1}} \tag{16.23}$$

16.4.3 Feasible generalized least squares

Weighted least squares assumes that the observed variance function is simple and known

$$\text{var}\left(e_i\right) = \sigma^2 x_i$$

It really could be

$$\text{var}\left(e_i\right) = \sigma^2 x_i^2$$

or

$$\text{var}\left(e_i\right) = \sigma^2 x_i^{\frac{1}{2}}$$

I doubt that it ever really is known. Thus, the power of x can be another parameter to estimate, λ

$$\text{var}\left(e_i\right) = \sigma^2 x_i^{\lambda} \tag{16.24}$$

When we took the square root above, we were weighting the data by $\frac{1}{x^{\frac{1}{2}}}$; $\lambda = \frac{1}{2}$. Now, we would like to create weights allowing the power of x to be flexible. How can we estimate λ? We can take the log of the expression to get

$$\ln\left(\sigma^2 x_i^{\lambda}\right) = \ln\left(\sigma^2\right) + \lambda\left(\ln\left(x_i\right)\right) \tag{16.25}$$

by taking the exponent of both sides we get

$$e^{\ln\left(\sigma^2 x_i^{\lambda}\right)} = e^{\ln\left(\sigma^2\right) + \lambda\left(\ln\left(x_i\right)\right)}$$

$$\sigma^2 x_i^{\lambda} = e^{\gamma_0 + \gamma_1 z_i} \tag{16.26}$$

so basically

$$\gamma_0 = \ln\left(\sigma^2\right),$$

$$\gamma_1 = \lambda,$$

and

$$z_i = \ln(x_i)$$

Did you catch what we did? One of the greatest tricks in the book of social science is to take a multiplicative relationship with unknown powers like

$$\sigma^2 x_i^{\lambda}$$

and transforming it into something resembling a linear function that we can estimate

$$w_i = \frac{1}{e^{\gamma_0 + \gamma_1 \ln(x_i)}} \tag{16.27}$$

Thus, we can estimate weights were we allow λ to be anything. Since γ_0 or $\ln(\sigma^2)$ is just a constant, we are still weighting proportional to x, only now we are scaling x based on empirical information. Generalized least squares requires us to know the exact weights, but this method allows us to make a *feasible* and empirical estimation of the weights, and that is why it is called feasible generalized least squares (FGLS). Doing this in a software package is relatively straightforward.

1. Estimate the naive variance of our residuals, which in practice is the square of the OLS residual
2. Run a non-linear least squares model, where these squared residuals are function of the the exponent of a regression line that uses $\ln(x)$ as a predictor
3. Estimate the weights with the inverse of the linear prediction
4. Re-estimate the model using these weights

In Model 4 from Table **??**, the estimate of γ_1 is 1.76 which implies $\text{var}(e_i) = \sigma^2 x_i^{1.76}$, quite different than $\text{var}(e_i) = \sigma^2 x_i^1$ that we thought we knew in weighted least squares model. It also produced the smallest standard error.

What makes this method so attractive is not only can we empirically estimate the functional form of the variance, but we can put in any variable we want into that non-linear model. The matrixes for estimation are the

Table 16.1: Models predicting charitable giving that handle heterskedasticity

Coefficients	Model 1	Model 2	Model 3	Model 4
hhincome	22.614***	22.614**	22.186***	21.989***
	(6.777)	(7.135)	(6.023)	(5.868)
Intercept	432.067	432.067	457.873	468.022
	(434.469)	(381.719)	(334.358)	(283.750)
Model Statistics				
N	328.000	328.000	328.000	328.000
F	11.134	10.046	13.567	14.040
R^2	0.033	0.033	0.040	0.041
df Regression	1.000	1.000	1.000	1.000
df Error	326.000	326.000	326.000	326.000
Model 1 is OLS regression				
Model 2 is OLS regression with robust SEs				
Model 3 is WLS regression				
Model 4 is FGLS regression, $\gamma_1 = 1.76$				
SEs in parentheses, $**p < 0.01, ***p < 0.001$				

same as in the weighted least squares, or generalized least squares, model. What makes this different is the method we used to estimate the weights. How do these methods compare?

We see that compared to the OLS model, the robust standard errors increased the standard error and reduced the level of significance. However, the generalized least squares model had a slightly smaller standard error, and the feasible generalized least squares had the smallest standard error of all. In this case, it did not make that much of a difference, but in a situation in which a paper is on the line, these methods can prove invaluable.

Table 16.2: Model of residuals from Model 1 in Table **??** on fitted values

Coefficients	Model
$\hat{y}$	0.001***
	(0.000)
Intercept	-0.879
	(0.558)
Model Statistics	
N	328.000
F	12.173
R^2	0.036
df Regression	1.000
Sum of Squares Regression	85.570
df Error	326.000
Sum of Squares Error	2291.589
Model predicting standardized residuals	
SEs in parentheses, $***p < 0.001$	

Chapter 17

Generalized least squares, in general

The next few chapters will utilize a method of estimation known as generalized least squares. We encountered this method first when talking about solving non-constant variance in Chapter **??**. In this brief chapter, we outline the basics of using generalized least squares.

Generalized least squares is essentially a weighting technique to solve the issue of correlated error terms.

Recall the linear model in matrix form, (**??**)

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

$$\begin{bmatrix} y_i \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p11} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & \cdots & x_{pN} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}$$

Life is easier when we assume each error term, e_i is independent of all other error terms. That allows us to say that they are distributed normally with mean 0 and variance σ^2. If this is not the case, we need to express the residuals as being normal, with a mean 0, but a variance-covariance matrix of Σ_{ee}

$$\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{\Sigma_{ee}}\right) \tag{17.1}$$

The use of Σ is a little confusing, but we use it since it is the capital version of σ. Σ_{ee} is a theoretical matrix in which the variance of the residuals fall

along the diagonal, and the relationship of each residual is expressed in the off-diagonals.

If Σ_{ee} were known, we could write out the log-likelihood function for the slopes of a normally distributed outcome

$$\ln\left(L\left(\mathbf{b},|\mathbf{y}\right)\right)=-\frac{N}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left(\det\boldsymbol{\Sigma}_{\mathbf{ee}}\right)-\frac{1}{2}\left(\mathbf{y}-\mathbf{Xb}\right)'\boldsymbol{\Sigma}_{\mathbf{ee}}^{-\mathbf{1}}\left(\mathbf{y}-\mathbf{Xb}\right) \quad (17.2)$$

and like least squares in general, we are working to minimize the residuals, $\left(\mathbf{y}-\mathbf{Xb}\right)'\boldsymbol{\Sigma}_{\mathbf{ee}}^{-\mathbf{1}}\left(\mathbf{y}-\mathbf{Xb}\right)$. Doing the calculus to find the estimator of $\mathbf{b}$ yields

$$\mathbf{b}=\left(\mathbf{X}'\boldsymbol{\Sigma}_{\mathbf{ee}}^{-\mathbf{1}}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\mathbf{ee}}^{-\mathbf{1}}\mathbf{y} \quad (17.3)$$

with variances of our slopes provided by

$$V\left(\mathbf{b}_{\mathbf{GLS}}\right)=\left(\mathbf{X}'\boldsymbol{\Sigma}_{\mathbf{ee}}^{-\mathbf{1}}\mathbf{X}\right)^{-1} \quad (17.4)$$

As I stated earlier, generalized least squares is essentially a transformation. If we consider the matrix Γ to be the square-root of Σ_{ee}^{-1}, then $\Gamma'\Gamma=\Sigma_{ee}^{-1}$. In that case, $X^{*}=\Gamma X$ and $y^{*}=\Gamma y$, and the estimator of the slopes is the familiar

$$\mathbf{b}=\left(\mathbf{X}^{*'}\mathbf{X}^{*}\right)^{-\mathbf{1}}\mathbf{X}^{*'}\mathbf{y}^{*} \quad (17.5)$$

which is why GLS is generally accomplished with weights. Of course, we need to guess at these weights, as we do with non-constant variance problems, see Chapter **??**.

Chapter 18

Logistic regression

Generalized linear models, of which the logit is one, form a large branch of statistical models that seek to use many of the tools of linear regression on dependent variables that do not meet the requirements of OLS regression; namely linearity and normality.

We begin by trying to derive the estimator for a dichotomous outcome, finding we have a problem, then offering a solution. We then move from this specific circumstance to a more general framework of generalized linear models that can encompass many types of outcomes that can be expressed in a specific form.

Consider the data in Table **??**. Figure **??** presents a scatter plot of these data. How would we best model y? One possibility would be to use OLS regression and fit a linear probability model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

The big problem with this approach is that some values of x produce nonsense values of y, as we can see in Figure **??**. In this figure, two of the values of x predict values of y that are less than 1, which is out of the range for this type of outcome.

The other more technical problems with this approach is that we don't meet all the assumptions of OLS regression:

- y is a linear function of predictors–NO!
- The expected value of any residual is zero–NO!
- The variance of the residuals in constant–NO!

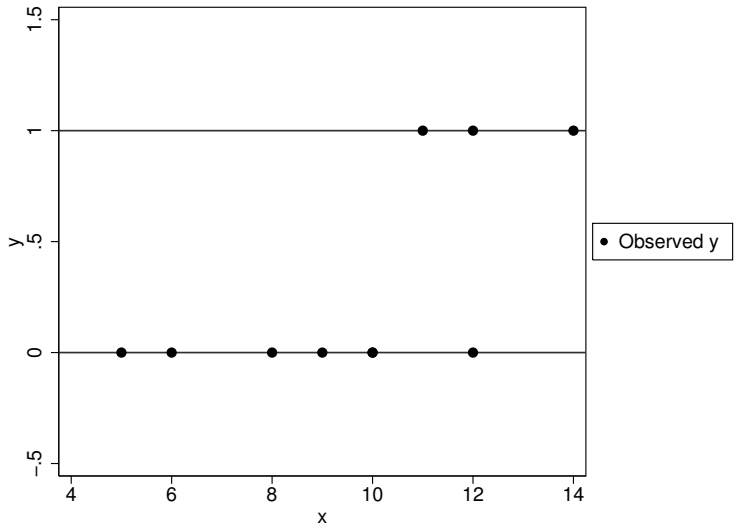

Figure 18.1: Scatterplot of data in Table **??**

- The covariance of all residuals is zero, i.e. $\text{cov}\,(e_i, e_j) = 0$
- The values of the predictors are not random (especially the intercept)
- The values of the predictors are not linear combinations of each other
- The errors are distributed normally–NO!

18.0.1 Dealing with probabilities and odds

To fit a model with these data, we need to deal with the mean of y. Remember that regression is about means, and regression is about predicting the mean of y as a function of covariates. Recall that the mean of a dichotomous variable is a proportion. The issue with regression is that we are constrained by the fact that proportions are bound by 1 and 0, and regressions are infinite lines. The trick is to transform expected (i.e., the mean) outcome into a logit (log odds).

Thus, the mean of y is

$$\bar{y} = \text{proportion of 1s} = \text{chance of 1s} \tag{18.1}$$

Table 18.1: Dichotomous variable y and predictor x

y	x
0	5
0	8
0	12
1	12
0	10
0	10
0	9
1	14
0	6
1	11

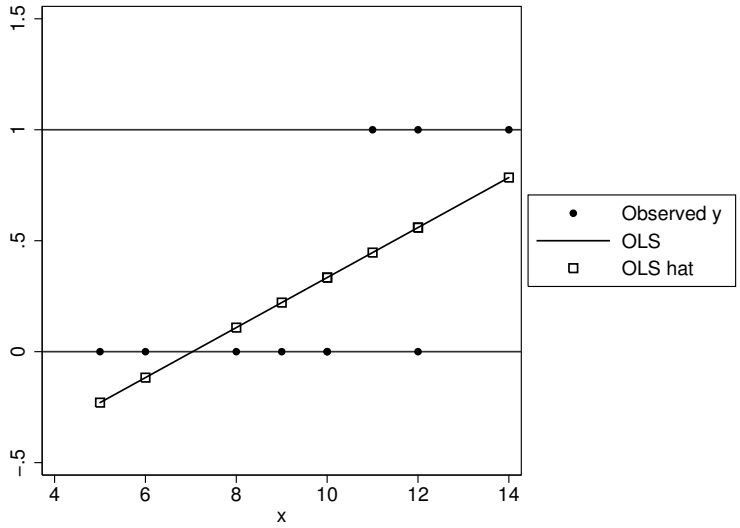

Figure 18.2: Scatterplot of data in Table **??** with linear probability model (OLS)

and then the odds are

$$\text{odds} = \frac{\text{Chance of something happening}}{\text{Chance of something not happening}}$$

or

$$\text{odds} = \frac{\bar{y}}{1-\bar{y}} \tag{18.2}$$

which can be turned into an infinite line with a natural log

$$\log \text{odds} = \ln\left(\frac{\bar{y}}{1-\bar{y}}\right) \tag{18.3}$$

So, let's call the function for the mean of y as a function of x π. This makes the bivariate regression

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \tag{18.4}$$

This gives us a regression model to fit the log-odds as a function of covariates

$$\ln\left(\frac{\Pr(y_i = 1)}{\Pr(y_i = 0)}\right) = \beta_0 + \beta_1 x \tag{18.5}$$

which we can rearrange to get a function for the mean (i.e. probability) of y

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{18.6}$$

To summarize, given a mean of y, which we will call the probability p, we can consider the logit function that produces log-odds (the left panel of Figure **??**)

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \tag{18.7}$$

as the quantity to model, and then once we have the estimates of β_0 and $beta_1$ that predict the log-odds, we can return to probabilities with the inverse logit function (the right panel of Figure **??**)

$$inv.logit(\beta_0, \beta_1, x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{18.8}$$

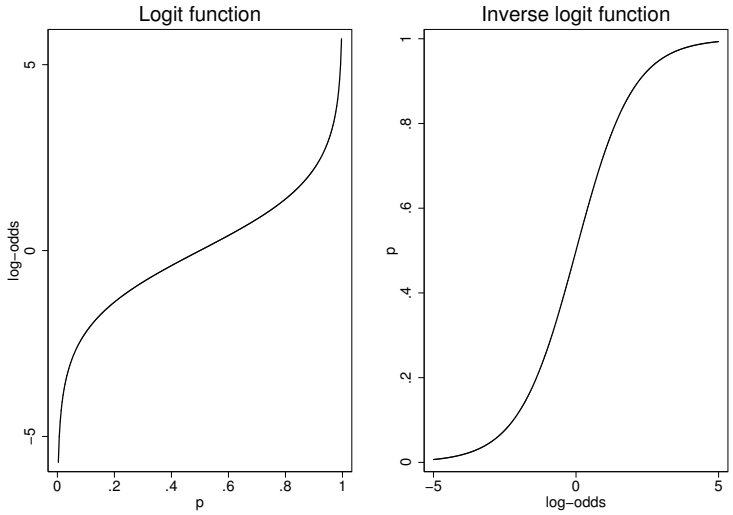

Figure 18.3: Logit and inverse-logit functions

18.0.2 Estimating a regression that predicts probabilities and odds

Recall that the likelihood for a binomial variable is

$$L\left(p|y_1 \ldots y_N\right) = p^{\sum_{i=1}^{N} y_i} \left(1-p\right)^{N-\sum_{i=1}^{N} y_i}$$

where $\Pr\left(y_i = 1\right) = p$, see (**??**). We can rewrite this as

$$L\left(p_i|y_1 \ldots y_N\right) = \prod_{i=1}^{N} \pi\left(x_i\right)^{y_i} \left(1-\pi\left(x_i\right)\right)^{1-y_i} \tag{18.9}$$

or

$$L\left(p_i|y_1 \ldots y_N\right) = \prod_{i=1}^{N} \left(\frac{\pi\left(x_i\right)}{1-\pi\left(x_i\right)}\right)^{y_i} \left(1-\pi\left(x_i\right)\right) \tag{18.10}$$

Since

$$\left(\frac{\pi\left(x_i\right)}{1-\pi\left(x_i\right)}\right) = e^{\beta_0+\beta_1 x_i}$$

and

$$(1 - \pi(x_i)) = \left(1 + e^{\beta_0 + \beta_1 x_i}\right)^{-1}$$

then the likelihood is

$$L(p_i|y_1 \dots y_N) = \prod_{i=1}^{N} e^{y_i \beta_0 + \beta_1 x_i} \left(1 + e^{\beta_0 + \beta_1 x_i}\right)^{-1} \tag{18.11}$$

and the log likelihood is

$$\ln(L(p_i|y_1 \dots y_N)) = \sum_{i=1}^{N} y_i (\beta_0 + \beta_1 x_i) - \ln\left(1 + e^{\beta_0 + \beta_1 x_i}\right) \tag{18.12}$$

Now that we have a likelihood function, we *should* be able to derive the estimators for the intercept and slope. Let's start with the intercept:

$$\frac{\partial \ln(L(p_i|y_1 \dots y_N))}{\partial \beta_0} = \sum_{i=1}^{N} y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \tag{18.13}$$

set to zero and

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \tag{18.14}$$

the solution is the sum of y's needs to be the sum of predicted probabilities. Now, for the slope

$$\frac{\partial \ln(L(p_i|y_1 \dots y_N))}{\partial \beta_1} = \sum_{i=1}^{N} x_i y_i - \frac{x_i}{1 + e^{-\beta_0 + \beta_1 x_i}} \tag{18.15}$$

set to zero and

$$\sum_{i=1}^{N} x_i y_i = \sum_{i=1}^{N} \frac{x_i}{1 + e^{-\beta_0 + \beta_1 x_i}} \tag{18.16}$$

No solution exists.

18.1 Introduction to maximum likelihood estimation

How can we find these parameters? We can take advantage of another problem with non-linear outcomes: the variance is dependent on the mean. For example, the variance of a binomial variable is $\text{var}(y) = \bar{y}(1-\bar{y})$ or $\text{var}(p) = p(1-p)$. We then use the tricks we used to solve non-constant variance (see Chapter **??**, weighted least squares) iteratively. In matrix notation, one procedure of *iterative weighted least squares* is as follows (computers do something slightly different):

1. start with $\beta_0 =$ the mean, and $\beta_1 = 0$
2. calculate weights based on the variance of the mean. In the case of a binomial outcome, the weight for each case is $= \hat{y}_i(1-\hat{y}_i)$, where $\hat{y}_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$
3. redo the least squares estimates of β_0 and β_1 with the new weights and using the residual $(y_i - \hat{y}_i)$ as the outcome. In matrix notation, this is

$$\mathbf{b_{l+1}} = \mathbf{b_l} + \left(\mathbf{X'W_lX}\right)^{-\mathbf{1}} \mathbf{X'}\left(\mathbf{y} - \mathbf{\hat{y}_l}\right) \tag{18.17}$$

4. Each time, calculate the log-likelihood for the new predictions, using the log-likelihood function specific to the type of outcome. In the case of a binomial outcome, this is $\ln\left(L\left(p_i|y_1 \ldots y_N\right)\right) = \sum_{i=1}^{N} y_i\left(\beta_0 + \beta_1 x_i\right) - \ln\left(1 + e^{\beta_0+\beta_1 x_i}\right)$
5. Keep going with steps 2-4 until the changes in the log-likelihood are small

Once we have the parameters, getting the standard errors of the parameters is pretty easy (with matrix algebra). The Asymptotic (i.e. when the samples are big) distribution of the maximum likelihood estimator of the slopes is

$$\text{var}(\mathbf{b}) = \left(\mathbf{X'WX}\right)^{-\mathbf{1}} \tag{18.18}$$

and the square-root of the diagonal elements are the standard errors.

18.1.1 Example logistic regression estimation

While this is boring, it does offer some insight. We begin with the data in Table **??**. The first step is to estimate the mean of our outcome variable, which in this case is 0.3. Using the likelihood function, we see that the log-likelihood for this data is -6.109. We set the intercept to the log-odds (-0.847) and the slope to 0.

Table 18.2: Results from iteration 0

iteration 0			
y	p_0	$\ln(odds_0)$	ll_0
0	0.300	-0.847	-0.357
0	0.300	-0.847	-0.357
0	0.300	-0.847	-0.357
1	0.300	-0.847	-1.204
0	0.300	-0.847	-0.357
0	0.300	-0.847	-0.357
0	0.300	-0.847	-0.357
1	0.300	-0.847	-1.204
0	0.300	-0.847	-0.357
1	0.300	-0.847	-1.204
		log likelihood =	-6.10864

β_0 = -.84729786, β_1 = 0

We then use these proportions to get the variances ($p\,(1-p)$) for the weight matrix and crank out

$$\mathbf{b_{l+1}} = \mathbf{b_l} + \left(\mathbf{X'W_lX}\right)^{-1}\mathbf{X'}\left(\mathbf{y} - \hat{\mathbf{y}}_\mathbf{l}\right)$$

the results are in Table **??**. We see that the predicted probabilities are now different, we have a slope, and the log-likelihood is now up -3.726. We keep going like this, each time getting new variances and slopes, until the log-likelihoods stop changing, see Tables **??** to **??**. I plot the iteration's log-likelihoods in Figure **??**.

With our logistic regression in hand, (β_0 = -16.291254, β_1 = 1.4271929), we can now predict the probability of $y = 1$ with

$$\hat{p}_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \tag{18.19}$$

Table 18.3: Results from iteration 1

iteration 1						
y	p_0	$y - p_0$	$var(p_0)$	p_1	$\ln(odds_1)$	ll_1
0	0.300	-0.300	0.210	0.033	-3.370	-0.034
0	0.300	-0.300	0.210	0.147	-1.760	-0.159
0	0.300	-0.300	0.210	0.596	0.387	-0.905
1	0.300	0.700	0.210	0.596	0.387	-0.518
0	0.300	-0.300	0.210	0.335	-0.686	-0.408
0	0.300	-0.300	0.210	0.335	-0.686	-0.408
0	0.300	-0.300	0.210	0.227	-1.223	-0.258
1	0.300	0.700	0.210	0.812	1.460	-0.209
0	0.300	-0.300	0.210	0.056	-2.833	-0.057
1	0.300	0.700	0.210	0.463	-0.150	-0.771
				log likelihood	=	-3.72638

β_0 = -6.0527866, β_1 = .53664833

Table 18.4: Results from iteration 2

iteration 2						
y	p_1	$y - p_1$	$var(p_1)$	p_2	$\ln(odds_2)$	ll_2
0	0.033	-0.033	0.032	0.004	-5.645	-0.004
0	0.147	-0.147	0.125	0.047	-3.003	-0.048
0	0.596	-0.596	0.241	0.627	0.520	-0.987
1	0.596	0.404	0.241	0.627	0.520	-0.467
0	0.335	-0.335	0.223	0.224	-1.241	-0.254
0	0.335	-0.335	0.223	0.224	-1.241	-0.254
0	0.227	-0.227	0.176	0.107	-2.122	-0.113
1	0.812	0.188	0.153	0.907	2.281	-0.097
0	0.056	-0.056	0.052	0.008	-4.764	-0.008
1	0.463	0.537	0.249	0.411	-0.361	-0.890
				log likelihood	=	-3.12144

β_0 = -10.047784, β_1 = .88065271

Table 18.5: Results from iteration 3

iteration 3						
y	p_2	$y - p_2$	$var(p_2)$	p_3	$\ln(odds_3)$	ll_3
0	0.004	-0.004	0.004	0.000	-7.763	-0.000
0	0.047	-0.047	0.045	0.016	-4.132	-0.016
0	0.627	-0.627	0.234	0.670	0.710	-1.110
1	0.627	0.373	0.234	0.670	0.710	-0.400
0	0.224	-0.224	0.174	0.153	-1.711	-0.166
0	0.224	-0.224	0.174	0.153	-1.711	-0.166
0	0.107	-0.107	0.096	0.051	-2.922	-0.052
1	0.907	0.093	0.084	0.958	3.130	-0.043
0	0.008	-0.008	0.008	0.001	-6.553	-0.001
1	0.411	0.589	0.242	0.377	-0.501	-0.975
				log likelihood	=	-2.92932

β_0 = -13.815145, β_1 = 1.2103927

Table 18.6: Results from iteration 4

iteration 4						
y	p_3	$y - p_3$	$var(p_3)$	p_4	$\ln(odds_4)$	ll_4
0	0.000	-0.000	0.000	0.000	-8.922	-0.000
0	0.016	-0.016	0.016	0.009	-4.749	-0.009
0	0.670	-0.670	0.221	0.693	0.816	-1.182
1	0.670	0.330	0.221	0.693	0.816	-0.366
0	0.153	-0.153	0.130	0.123	-1.967	-0.131
0	0.153	-0.153	0.130	0.123	-1.967	-0.131
0	0.051	-0.051	0.048	0.034	-3.358	-0.034
1	0.958	0.042	0.040	0.973	3.598	-0.027
0	0.001	-0.001	0.001	0.001	-7.531	-0.001
1	0.377	0.623	0.235	0.360	-0.576	-1.022
				log likelihood	=	-2.90239

β_0 = -15.877707, β_1 = 1.3911007

Table 18.7: Results from iteration 5

iteration 5						
y	p_4	$y - p_4$	$var(p_4)$	p_5	$\ln(odds_5)$	ll_5
0	0.000	-0.000	0.000	0.000	-9.149	-0.000
0	0.009	-0.009	0.009	0.008	-4.870	-0.008
0	0.693	-0.693	0.213	0.697	0.835	-1.195
1	0.693	0.307	0.213	0.697	0.835	-0.361
0	0.123	-0.123	0.108	0.117	-2.018	-0.125
0	0.123	-0.123	0.108	0.117	-2.018	-0.125
0	0.034	-0.034	0.033	0.031	-3.444	-0.031
1	0.973	0.027	0.026	0.976	3.687	-0.025
0	0.001	-0.001	0.001	0.000	-7.722	-0.000
1	0.360	0.640	0.230	0.356	-0.592	-1.032
				log likelihood	=	-2.90169

β_0 = -16.279248, β_1 = 1.4261493

Table 18.8: Results from iteration 6

iteration 6						
y	p_5	$y - p_5$	$var(p_5)$	p_6	$\ln(odds_6)$	ll_6
0	0.000	-0.000	0.000	0.000	-9.155	-0.000
0	0.008	-0.008	0.008	0.008	-4.874	-0.008
0	0.697	-0.697	0.211	0.697	0.835	-1.195
1	0.697	0.303	0.211	0.697	0.835	-0.360
0	0.117	-0.117	0.104	0.117	-2.019	-0.125
0	0.117	-0.117	0.104	0.117	-2.019	-0.125
0	0.031	-0.031	0.030	0.031	-3.447	-0.031
1	0.976	0.024	0.024	0.976	3.689	-0.025
0	0.000	-0.000	0.000	0.000	-7.728	-0.000
1	0.356	0.644	0.229	0.356	-0.592	-1.032
				log likelihood	=	-2.90169

β_0 = -16.291254, β_1 = 1.4271929

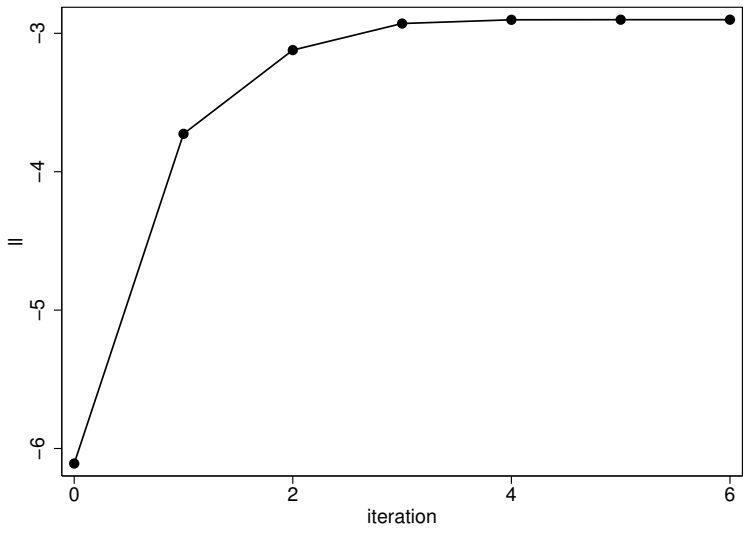

Figure 18.4: Log-likelihoods from maximum likelihood iterations to fit logistic regression to data in Table **??**

These values are plotted in Figure **??**. Notice how the regression line does not cross 0 or 1, unlike Figure **??**.

18.1.2 Model statistics

Models can then be summarized as having good or poor "fit" by looking at the "deviance". The model deviance is

$$G^2 = -2\ln(L) \tag{18.20}$$

We can use this quantity to get a measure of model fit. This is measure is a pseudo-R^2. The idea behind this measure is that we measure how much "deviance" we remove by fitting our model. If we use the log-likelihood from iteration 0 ($\ln(L_0)$) and the log-likelihood from the final iteration ($\ln(L_f)$), we can think of the pseudo-R^2 as

$$R^2_{pseudo} = 1 - \frac{G^2_f}{G^2_0} = 1 - \frac{\ln(L_f)}{\ln(L_0)} \tag{18.21}$$

Another way to think about model fit is with a χ^2 statistic

$$\chi^2 = -2\left(\ln(L_0) - \ln(L_f)\right) \tag{18.22}$$

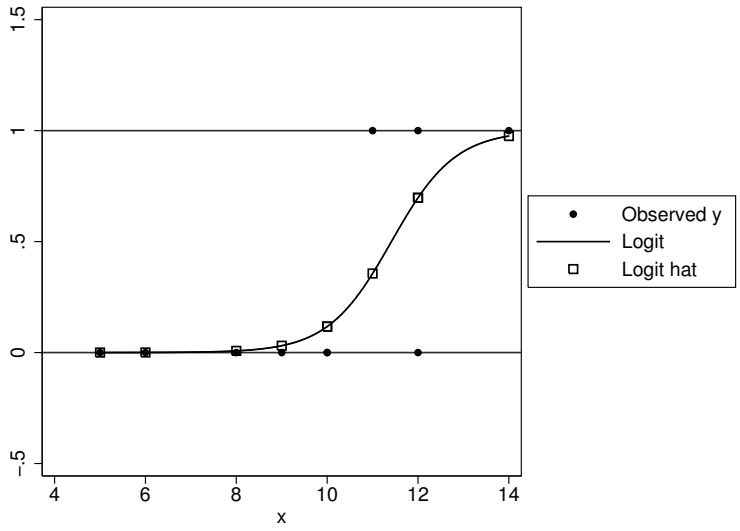

Figure 18.5: Scatterplot of data in Table **??** with logistic regression model

with degrees of freedom equal to predictors, excluding the constant. Why excluding the constant, because we use the constant in the null (0) model, and this χ^2 statistic reflects how much the likelihood function "improved" with our extra parameters. In other words, its a likelihood ratio test (see Chapter **??**)!

This is another method to test the model. Recall that the likelihood for the null model above (Table **??**) was -6.10864, and that by iteration 6, (Table **??**) was -2.90169. This leads to a χ^2 of

$$\chi^2 = -2((-6.10864) - (-2.90169)) = 6.4139$$

which is significant with 1 degree of freedom.

In the model, the slope is $\beta_1 = 1.4271929$ with a standard error (see next section) of 0.9986. This leads to a test statistic of

$$z = 1.4271929/0.9986 = 1.4292$$

Not significant. Unlike OLS regression, the test of the significance for model fit does not directly relate to the test of coefficients.

18.1.3 Variances of slopes

Without getting *too* technical, the variance of the slopes is related to the curvature of the likelihood function at the end of the iterations. In other words, it is related to how "flat" the function is at that point. When it is very flat, the estimates are uncertain and the standard errors are large. If it less flat, and more of a tip, the estimate is precise, and the standard errors are small. The mechanics of how this works is a little beyond the scope of these notes, but the foregoing discussion should shed some intuitive light.

If you compare Figure **??** with Figure **??**, you will notice that the plot of iterations looks very similar to the left portion of the likelihood function. How "flat" this region is when the iteration stops is what drives the standard errors. To better understand this, recall that the variance of a proportion is (**??**)

$$\text{var}(p) = \frac{p(1-p)}{N}$$

As with all standard errors, this depends on the variance, which in this case is $p(1-p)$ and the sample size N. Note that the variance is a function of the mean, p. The shape of the likelihood function is also a function of p, as you recall from Chapter **??**. This leads to the intuition that the variance is linked to the shape of the likelihood function.

If you examine Figure **??** you will notice a few things. First, for any value of p, the shape of the likelihood function is the same. However, as you increase sample size (going down the rows), the variance of p decreases, leading to a more precise estimate. We expect this. But notice what happens when you stay with the same sample size, but change the value of p: as p moves to 0.50, the variance *increases*!

Why does it increase? It depends on how "flat" the likelihood curve is around the estimate. If we zoom in on a row in Figure **??**, instead assuming 100 cases, we get Figure **??**. Notice how the tip of the curve for $p = 0.10$ is "sharp", then how the tip of the curve for $p = 0.50$ is "not as sharp." This "sharpness" is reflected in the variances of the point estimate, which is related directly to what that point estimate actually is; this is different than normal data, where the means and variances are independent.

To connect everything, recall the matrix formula for the variances of the slopes (**??**)

$$\text{var}(\mathbf{b}) = (\mathbf{X'WX})^{-1}$$

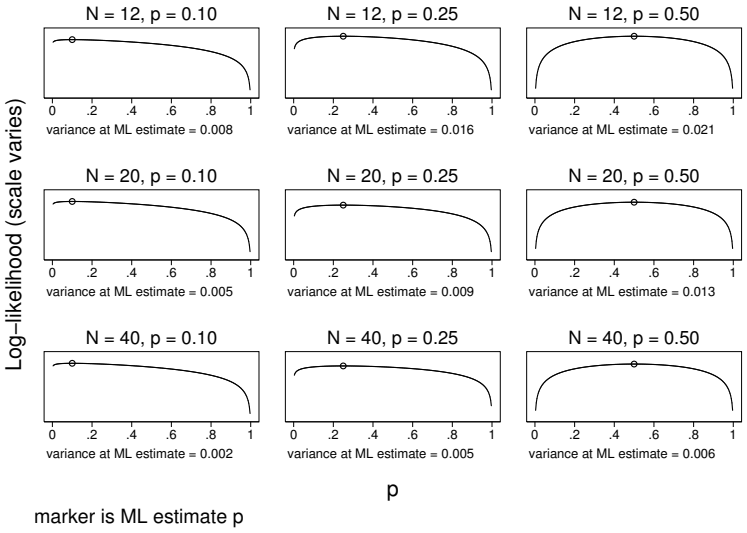

Figure 18.6: Variance of proportion estimator and likelihood functions for various proportions and sample sizes

what is $\mathbf{W}$ again? The variances of the predictions for each case ($\hat{p}_i (1 - \hat{p}_i)$)! Thus, the standard errors for a logistic regression are based on the variances of the predictions.

Since the concept of degrees of freedom for logistic models is difficult, what many people do is a Wald test for the coefficients. That is, we simply use the z distribution to evaluate the null hypothesis:

$$z = \frac{\beta}{SE(\beta)}$$

where

$$SE(\beta) = \sqrt{\text{var}(\beta)}$$

More technical

With Figure **??** we can see that the variance is related to the curvature of the likelihood function, or the second derivative. We can see this by taking

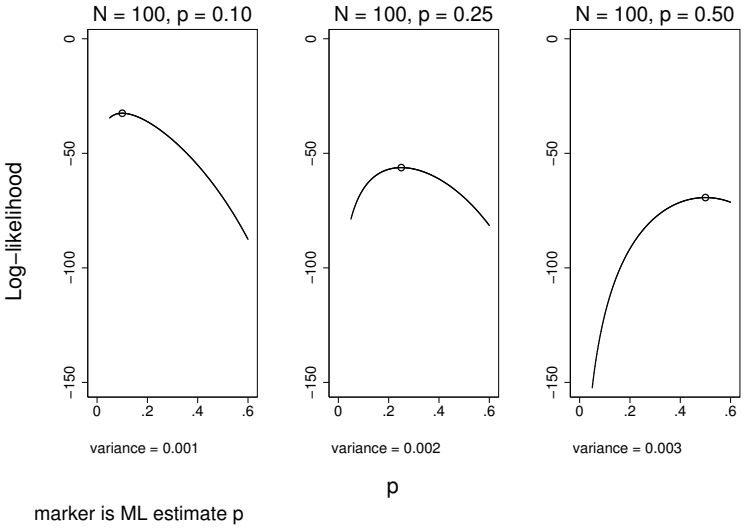

Figure 18.7: Variance of proportion estimator and likelihood functions for various proportions and sample sample size = 100

the second derivative of the likelihood function for the intercept

$$\frac{\partial^2 \ln(L)}{\partial \beta_0^2} = -\sum_{i=1}^{N} \frac{e^{-(\beta_0+\beta_1 x_i)}}{\left(1+e^{-(\beta_0+\beta_1 x_i)}\right)^2}$$

which reduces to

$$\frac{\partial^2 \ln(L)}{\partial \beta_0^2} = -\sum_{i=1}^{N} \hat{p}_i\left(1-\hat{p}_i\right) \tag{18.23}$$

and for the slope

$$\frac{\partial^2 \ln(L)}{\partial \beta_1^2} = -\sum_{i-1}^{N} \left(x_i^2 \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) \left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) \right)$$

which reduces to

$$\frac{\partial^2 \ln(L)}{\partial \beta_1^2} = -\sum_{i=1}^{N} x_i^2 \hat{p}_i\left(1-\hat{p}_i\right) \tag{18.24}$$

Also, covariance is related to

$$\frac{\partial^2 \ln(L)}{\partial \beta_0 \beta_1} = -\sum_{i=1}^{N} x_i \hat{p}_i (1 - \hat{p}_i) \tag{18.25}$$

As you can see, the curvature of the function is directly related to the variance of the prediction in both cases. These second derivatives of the squared slopes (β_0^2) and (β_1^2), as well as the product, ($\beta_0 \beta_1$), are formed into what is called a Hessian matrix which is computed with $\mathbf{X'WX}$. The inverse of $\mathbf{X'WX}$ provides the variances and covariances. This is why the curvature of the likelihood estimation function drives the standard errors of the slopes.

18.2 Interpretation

The core of interpretation in *any* regression is the unit of the outcome. In the case of logistic regression the outcome is not a probability, but the log-odds. Thus, the coefficients are in the units of log-odds.

The way to interpret each coefficient is as change in the log-odds. However, log-odds doesn't have any intrinsic meaning. Instead, it makes more sense to talk about an odds-ratio. This is the ratio the represents what happens when x increases by a single unit. In other words, there are the odds for a value of x, and there are the odds for a value $x + 1$, what is the ratio between the two? The answer is the exponent of the slope for x, e^{β}.

This is easier to understand with an example. Suppose a simple model where x is coded as either 0 or 1 and the outcome y, of course, is coded as 0 or 1. We can think of this as a two by two table as in Table **??**. We can think of the odds that $y = 1$ given $x = 0$ as

$$\text{odds}(y = 1 | x = 0) = \frac{n_{2,1}/n_{+,1}}{n_{1,1}/n_{+,1}} \tag{18.26}$$

and the odds that $y = 1$ given that $x = 1$ as

$$\text{odds}(y = 1 | x = 1) = \frac{n_{2,2}/n_{+,2}}{n_{1,2}/n_{+,2}} \tag{18.27}$$

the ratio of these odds works out to

$$\text{odds-ratio} = \frac{n_{2,2}/n_{1,2}}{n_{2,1}/n_{1,1}} = \frac{38/7}{14/41} = 15.898 \tag{18.28}$$

Table 18.9: Simple 2×2 table

y/x	0	1	total
0	41 ($n_{1,1}$)	7 ($n_{1,2}$)	48 ($n_{1,+}$)
1	14 ($n_{2,1}$)	38 ($n_{2,2}$)	52 ($n_{2,+}$)
total	55 ($n_{+,1}$)	45 ($n_{+,2}$)	100 ($n_{+,+}$)

If we had the actual dataset of 100 cases, we could run a logistic regression

$$\ln\left(\frac{\Pr(y_i=1)}{\Pr(y_i=0)}\right) = \beta_0 + \beta_1 x_i$$

$$\ln\left(\frac{\Pr(y_i=1)}{\Pr(y_i=0)}\right) = -1.074515 + 2.766191 x_i$$

which would produce an estimate of $\beta_1 = 2.766191$, and $e^{2.766191} = 15.898$. How is this the case. Again, it comes back to logs. We can write it out like this

$$\ln(\text{odds}_{x=1}) - \ln(\text{odds}_{x=0}) = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

take the exponent of both sides

$$\exp[\ln(\text{odds}_{x=1}) - \ln(\text{odds}_{x=0})] = \exp[\beta_1]$$

$$\frac{\text{odds}_{x=1}}{\text{odds}_{x=0}} = \exp[\beta_1] \tag{18.29}$$

Thus, the exponent of the slope literally is the ratio of the odds. The log odds for $x = 0$ is the intercept, -1.074515, which equals odds of 0.34146. The odds for $x = 1$ is -1.074515 + 2.766191 = 1.691676, which equals odds of 5.42857. The ratio is 5.42857/0.34146 = 15.898. The nice thing is that this works with continuous (predictor) variables as well.

Language

For example, say the odds-ratio is 1.2, then the odds of the outcome being coded "1" is 120 percent as coded "0." Since "120 percent of the odds" is

still clumsy, we can take out the "of the odds" by subtracting 1.00 from the odds ratio and replace it with "more odds". Thus, with an odds-ratio of 1.2 we can say: It is 20 percent more odds to get a code of "1" than a code of "0."

If the odds-ratio is 1.00, then the odds of "1" is 100 percent the odds of "0"—the same odds, so equal chances, no effect.

If the odds-ratio is less than 1.0, we can use similar language. Say the odds-ratio is 0.8. Then we can say that the outcome being coded "1" is 80 percent the odds as coded "0." "80 percent the odds" means less odds of "1." However, "80 percent the odds" is clumsy as well. One thing you can do is subtract the odds ratio from 1.00 and say that the odds of a "1" code is so much less likely. With an odds-ratio is 0.80, 1.00-0.80 = 0.20, so a code of "1" has 20 percent fewer odds.

Warning!

I was careful not to use the word *chance*. "Chance" often connotes probabilities, and while it is technically not wrong to talk about chance (odds are chances as well), an odds-ratio does not deal with probability. For example, the chance in Table **??** of $y = 1|x = 0$ is about 25 percent (0.25). The chance of $y = 1|x = 1$ is about 84 percent (0.84). Taking the ratio, 0.84/0.25 = 3.360. This is a lot different than the *odds*-ratio of 15.898. The sloppy report or paper will say the chances increased by 16 fold, putting the idea in people's minds that the probabilities increased 1600 percent–not true!

18.2.1 Example logit

In this section we consider whether some basic covariates impact the opinions of a controversial Arizona law. SB1070, as the law is commonly known, was passed in Arizona in 2010 and was intended to curb the influx of illegal migrants into the state. One of the provisions was the ability of law enforcement to legally stop and question anyone they suspected of being in the state illegally. Obviously, this stirred a bit of controversy at the time and led to some data collection. In the analysis below, we consider what demographic and political characteristics of individuals lend themselves to supporting the "stop and question" (SandQ) portion of the law. The descriptive statistics are presented in Table **??**.

Table 18.10: Summary statistics of SB1070 poll

N = 562	Approve SandQ	Years of Educ.	Age < 60	Female	Rep.	Dem.
Mean(not approve)		15.497	0.497	0.644	0.181	0.520
SD(not approve)		2.389				
Mean(approve)		14.460	0.488	0.566	0.504	0.184
SD(approve)		2.167				
Mean(all)	0.685	14.786	0.491	0.591	0.402	0.290
SD(all)		2.289				

Source: Morrison Institute, Arizona State University

We see that in this sample, nearly 70 percent of the sample approve of this provision. We also see that the years of education of those that do not approve is a year higher than those who approve. we also see that of those that do not approve, about half are Democrats, and of those that approve, about half are Republican (the reference group is independent).

Next, we fit the model to this data, and the results are presented in Table **??**. Looking at years of education (centered on high school), we see that the log odds decrease by 0.210 for each year of education; the result is significant. Taking the exponent of this, we see that this leads to a decrease in the odds of about 20 percent, again for each year of education. Age and female are not significant.

However, Republicans are far more likely than independents to approve of the provision. The log odds increase by almost 1 when comparing Republicans to independent voters. This leads to an odds ratio of about 2.7. There is an equal and opposite effect for being a Democrat, with the log odds decreasing by 1, leading to a 66 percent reduction in the odds.

The pseudo-R^2 is about 0.15, meaning that the likelihood function improved by about 15 percent. The likelihood ratio test is about 104, and with 5 degrees of freedom the model is highly significant.

Table 18.11: Logistic regression model predicting approval of Arizona SB1070 law provision of questioning

Variable	Model	Odds-Ratio
Year of education - 12	-0.210***	0.811***
	(0.044)	(0.036)
Age LT 60	-0.063	0.939
	(0.204)	(0.192)
Female	-0.274	0.760
	(0.210)	(0.159)
Republican	0.988***	2.685***
	(0.258)	(0.692)
Democrat	-1.069***	0.343***
	(0.238)	(0.082)
Intercept	4.143***	
	(0.721)	
Model Statistics		
N	562.000	
Log-likelihood (null model)	-350.127	
Log-likelihood (final model)	-298.060	
Pseudo-R^2	0.149	
$\chi^2 LR$	104.133	
df Regression	5.000	
*SE*s in parentheses, $***p < 0.001$		

18.2.2 Marginal predictions

Usually what policy makers want to know is the difference in terms of probability, not odds. This can be a tricky business because while differences in log odds are linear, and thus constant no matter the reference point, this is not the case with probabilities. Going back to our SB1070 model (Table **??**), we can calculate the predicted probability by simply using the regression model to calculate the log odds, then using the inverse logistic function on the log odds.

For example, the intercept of the model in Table **??** is 4.143, and given the coding of the variables, this represents a male independent voter, with 12 years of education, who is aged 60 or older, which works out to a probability of $e^{4.143}/\left(1+e^{4.143}\right) = 0.9844$. This is represented in Figure **??**, which shows right side of the logit curve with points marked for independents, republicans, and democrats, all assuming older, high school aged, male populations. These numbers are literally just plugging in the coefficients. For example, the log odds for democrats is 4.143-1.069 = 3.074, which translates into a probability of $e^{3.074}/\left(1+e^{3.074}\right) = 0.9558$. Thus, for older males with high school educations, the marginal difference in probabilities is 0.9558 - 0.9844 = 0.0286, or about 3 percentage points.

Next, we alter another parameter and reevaluate these marginal changes. For each year of education the log odds decrease by -0.210, so moving from a high school to a college degree decreases the log odds by $-0.210 \times 4 = -0.84$ We then shift all the predicted log odds by this amount and recalculate the differences between the political identities. These are presented in Figure **??**. Notice that the spacing in the vertical lines is the same as in Figure **??**, but the difference in the horizontal lines (the differences in the probabilities) are different. Now, the effect of being a democrat is about 3 percentage points. In other words, when thinking about probabilities, the effect of political affection is twice as large for the college education as for the high school educated. Assumptions make a difference!

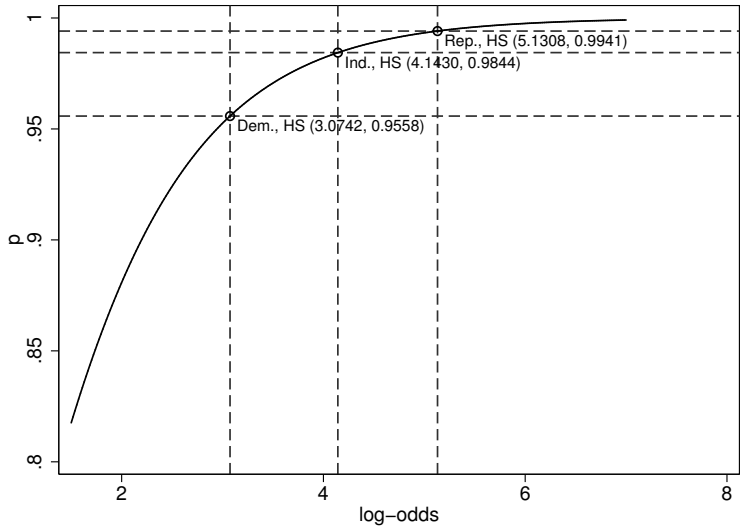

Figure 18.8: Predictions of the probability of approval of Arizona law based on model in Table **??** (12 years of education, male, older population)

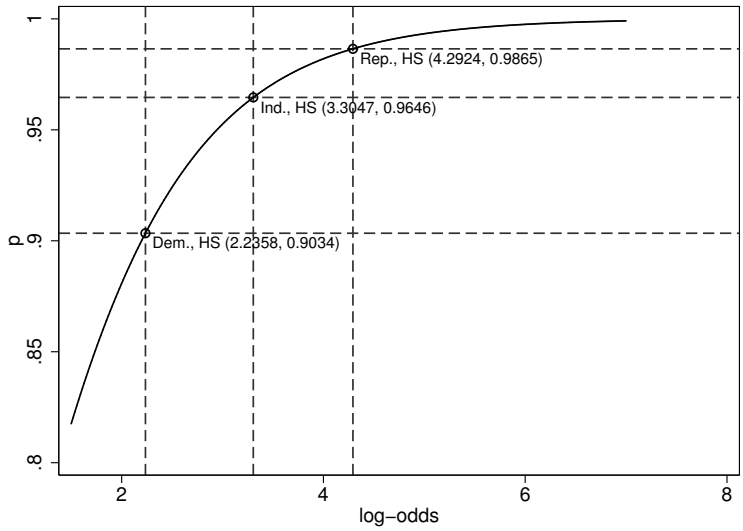

Figure 18.9: Predictions of the probability of approval of Arizona law based on model in Table **??** (16 years of education, male, older population)

Bibliography

Christopher L. Aberson. *Applied power analysis for the behavioral sciences*. Routledge, 2011.

Alan Agresti and Barbara Finlay. *Statistical Methods for the Social Sciences*. Pearson.

Peter C. Austin. A tutorial and case study in propensity score analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate behavioral research*, 46(1):119–151, 2011.

Howard S. Bloom. Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*.

Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

A. Colin Cameron. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

Jack Cohen. *Statistical power analysis for the behavioral sciencies*. Routledge, 1988.

Robert J Connor. Sample size for testing differences in proportions for the paired-sample design. *Biometrics*, pages 207–211, 1987.

Scott R Eliason. *Maximum likelihood estimation: Logic and practice*. Number 96. Sage, 1993.

Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

Ronald Aylmer Fisher and J Henry Bennett. *Statistical methods, experimental design, and scient ific inference*. Oxford University Press, 1973.

John Fox. *Applied Regression Analysis and Generalized Linear Models*. Sage.

David Freedman. *Statistical models: theory and practice*. Cambridge University Press, 2009.

David A. Freedman and Richard A. Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409, 2008.

Jeff Gill. *Essential mathematics for political and social research*. Cambridge University Press, 2006.

Harvey Goldstein, William Browne, and Jon Rasbash. Partitioning variation in multilevel models. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 1(4):223–231, 2002.

William E. Griffiths, R. Carter Hill, and George G. Judge. Learning and practicing econometrics, 1993.

R. Carter Hill, William E. Griffiths, and Guay C. Lim. *Principles of Econometrics*. Wiley.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1): 4–29, 2004.

Guido W. Imbens and Donald B. Rubin. Rubin causal model. In Steven N. Durlauf and Lawrence E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 2008.

Sam Kash Kachigan. *Statistical Analysis: An Interdisciplinary Introduction to Univariate and Multivariate Methods*. Radius, 1986.

David A. Kenny. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin*, 82 (3):345, 1975.

Leslie Kish. Survey sampling. new york: J. *Wiley & Sons*, 643:16, 1965.

Roderick J Little and Donald B Rubin. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health*, 21(1):121–145, 2000.

F. M. Lord. A paradox in the interpretations of group comparisons. *Psychological Bulletin*, a.

F. M. Lord. Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, b.

Robert H Lyles, Hung-Mo Lin, and John M Williamson. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Statistics in medicine*, 26(7):1632–1648, 2007.

Scott E Maxwell, Ken Kelley, and Joseph R Rausch. Sample size planning for statistical power and accuracy in parameter estimation. *Annu. Rev. Psychol.*, 59:537–563, 2008.

Charles E. McCulloch, Shayle R. Searle, , and John M. Neuhaus. *Generalized, Linear, and Mixed Models*. Wiley.

Breed D. Meyer. Natural and quasi-experiments in economics. *Journal of business & economic statistics*, 13(2):151–161, 1995.

Stephen W. Raudenbush and Anthony Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage.

John A Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2007.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B. Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188, 2001.

Donald B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Cengage Learning.

Jessaca Spybrook, Stephen W Raudenbush, Xiao-feng Liu, Richard Congdon, and Andrés Martínez. Optimal design for longitudinal and multilevel research: Documentation for the "optimal design" software. *Survey Research Center of the Institute of Social Research at University of Michigan*, 2006.

Michael Væth and Eva Skovlund. A simple approach to power and sample size calculations in logistic regression and cox regression models. *Statistics in medicine*, 23(11):1781–1792, 2004.