What I'm Currently Working On

Gaurav Khanna

April 1, 2018

Contents

List of Figures

List of Tables

Chapter 1

Miscellaneous

$$\left(\frac{10 \times 9 \times 8 \times 7}{6 \times 5}\right) \times (4 \times 3) + (2 \times 1) = 2018 \tag{1.0.1}$$

Chapter 2

Linear Algebra Review

Chapter 3

Probability and Stats Stuff

3.1 Notes - Brase et.al., *Understandable Statistics*

- Statistics is the study of how to collect, organize, analyze, and interpret numerical information from data - Individuals are people or objects included in the study. - A variable is the characteristic of the individual to be measured or observed.

- A quantitative variable has a numerical value for which operations such as addition or variance makes sense - A qualitative variable is describes an individual by placing them in a group or a category

Good example: climbers of Mt. Everest. We can know their age, height, weight, years of climbing experience – these are quantitative variables. We can also collect information about their gender, nationality etc. These are qualitative or category variables.

- Population data – the variable is collected from every individual of interest - Sample data – the variable is only collected from a subset of individuals of interest. If we have data from all the climbers of Mt. Everest, it is population data. If only a some of the climbers, it's sample data. Note: whether data is population data or sample data depends on the context and what boundary conditions you have placed. So the analysis of interest is all of the climbers of Mt. Everest during 2017, then if you collect data from all of these climbers, that's population data. You are not interested in all of the people who have ever climbed Mt. Everest. If you only got data on some of the climbers ion 2017, then its sample data.

- A simple random sample of n measurements from a population is a subset of the population selected in a manner such that } - every sample of size n from the population has an equal chance of being selected as any other sample of size n } - every member of the population has an equal chance of being included in a particular sample }

Great illustration of random sampling on. p14 – the lottery. Suppose a state's lottery system let's you buy a card and select 6 numbers form 1 to 42. You group of numbers is then compared to the winning group of six numbers selected by a simple random sampling. If your numbers match the randomly selected ones, you win big! Let's see if this meets the criteria for a random sample:

- Is the number 25 as likely to be selected as the number 5 in the winning group of six numbers? Yes, because each number from 1 - 42 has an equal chance of being selected. - Could all the winning numbers be even? Yes, since six even numbers is one of the possible group of six numbers - If I always play the numebrs 1,2,3,4,5,6, can I ever win? Yes, because that particular set of numbers is as likely as any of the other 5,245,786 groups of 6 numbers to be selected.

Sampling techniques:

- Stratified sampling: entire population is divided into subgroups, or strata, based on specific characteristics. For example, dividing a population based on age, income, education level, etc. All members of a stratum share one or more characteristics. Random samples can be drawn from each stratum.

- Systematic sampling: members of the population are sequentially numbered. The, from a starting point, every kth member of the population is included in the sample.

- Cluster sampling: entire population is divided into pre-existing segments or clusters. The clusters are randomly selected and every member of the selected cluster is included in the sample.

Pareto chart: a bar graph in which the bar heights represents the frequency of an event. In addition, the bars are arranged from left to right according to decreasing magnitude.

3.2 Covariance

The **covariance** between two jointly distributed real-valued random variables X and Y is defined as

$$cov(X,Y) = E\left[\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)\right] \tag{3.2.1}$$

That is, the covariance is the expected product of their deviation from their individual expected values (or means).

This quantity measure the strentgth of the *linear* relationship between two variables.

Since variance is sometimes denoted as σ_X^2, co-variance is sometimes written as σ_{XY}^2.

We can write this in a simplified form

$$\begin{aligned}
cov(X,Y) &= E\left[\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)\right] \\
&= E\left[XY - X\overline{Y} - \overline{X}Y + \overline{XY}\right] \\
&= E(XY) - E\left(X\overline{Y}\right) - E\left(-\overline{X}Y\right) + E\left(\overline{X}\ \overline{Y}\right) \\
&= E(XY) - \overline{X}\ \overline{Y} - \overline{X}\ \overline{Y} + \overline{X}\ \overline{Y} \\
&= E(XY) - \overline{X}\ \overline{Y}
\end{aligned} \tag{3.2.2}$$

We can also write Eqn.(3.2.1) in the summation notation as

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right) \tag{3.2.3}$$

Equating the above with Eqn.(3.2.2)

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right) &= E(XY) - \overline{X}\ \overline{Y} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i\ Y_i) - \overline{X}\ \overline{Y}
\end{aligned} \tag{3.2.4}$$

When merging linear regression notes into All Math file, make a reference to above equation as it appears in the derivation in Eqn.(4.3.12).

The best way to understand covariance is to visualize it. We saw in Section XYZ how variance gives us a measure of scatter about a mean. Covariance gives us a measure how much two variables move away from their mean together.

In Figure 3.2.1 we show two hypothetical sets of data. Intuitively, we can see the data in the top of the figure does not show any real pattern or link between the x and y coordinates. It is instructive to divide the data field into four quadrants by using two lines: a horizontal one at $\overline{Y}$ and a vertical one at $\overline{X}$. We can then measure the covariance of the points in each of the quadrants with the help of Eqn.(3.2.3).

We see that where $\left(X_i - \overline{X}\right)$ and $\left(Y_i - \overline{Y}\right)$ have the same sign, we will have positive contributions to the covariance in Eqn.(3.2.3). When $\left(X_i - \overline{X}\right)$ and $\left(Y_i - \overline{Y}\right)$ have opposite signs, the contribution of those data points will be negative in Eqn.(3.2.3). The covariance is simply the average of all of these contributions.

It wouldn't surprise us to learn that the covariance of the data in the bottom of Figure 3.2.1 is much higher for the data set shown at the top. We will look at specific data sets in Section XYZ and connect the variance, covariance, and correlation of data.

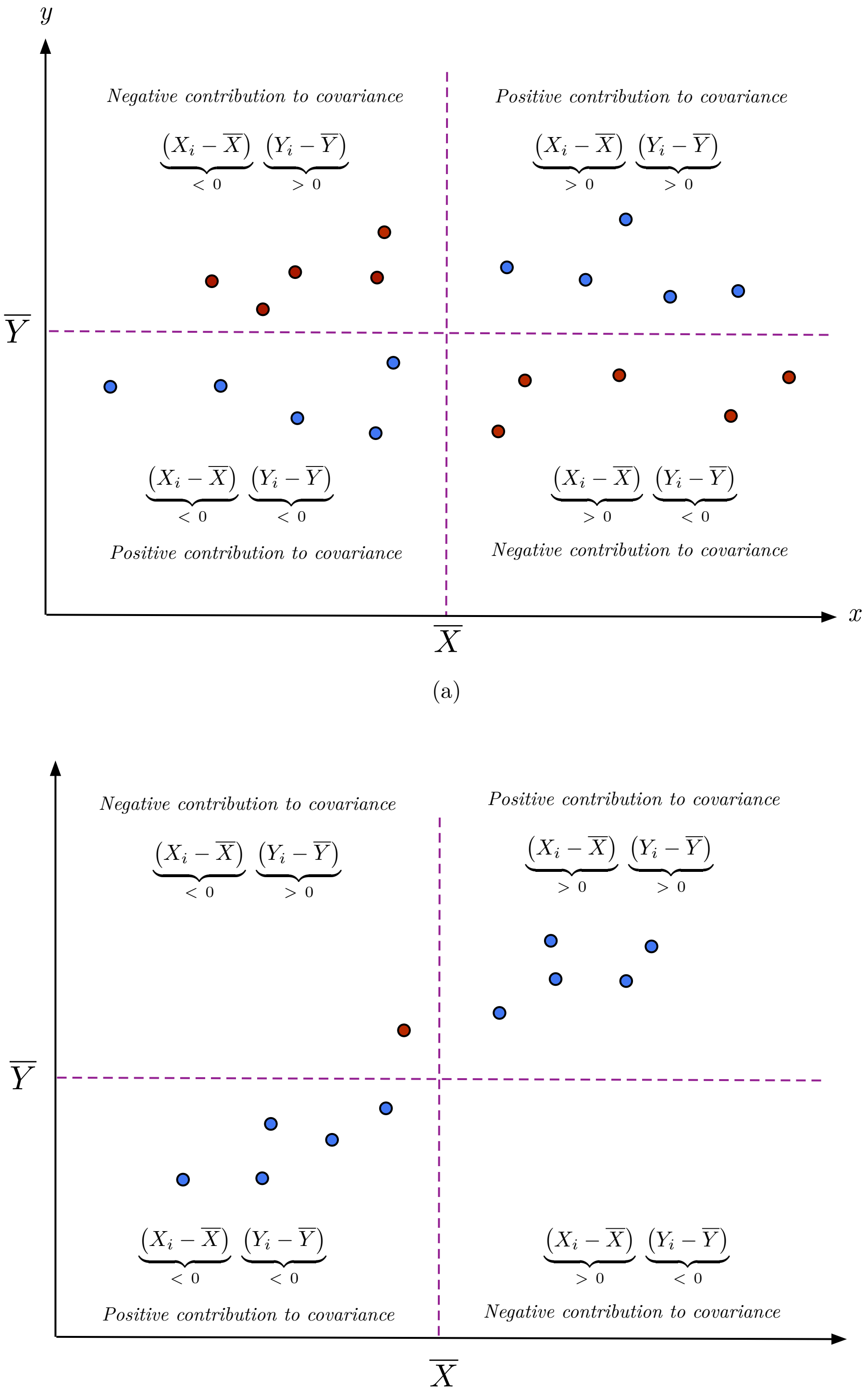
We may also write the covariance without reference to the mean values of X and Y

$$cov(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}\left(X_i - X_j\right)\left(Y_i - Y_j\right) \tag{3.2.5}$$

Furthermore, if X_i and Y_i take on the values (x_i, y_i) with equal probability $1/n$, then we can express their covariance as

$$cov(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{1}{2}\left(X_i - X_j\right)\left(Y_i - Y_j\right) \tag{3.2.6}$$

Need to verify these equations


y
Negative contribution to covariance
$\underbrace{(X_i - \overline{X})}_{<0}\ \underbrace{(Y_i - \overline{Y})}_{>0}$
Positive contribution to covariance
$\underbrace{(X_i - \overline{X})}_{>0}\ \underbrace{(Y_i - \overline{Y})}_{>0}$
$\overline{Y}$
$\underbrace{(X_i - \overline{X})}_{<0}\ \underbrace{(Y_i - \overline{Y})}_{<0}$
Positive contribution to covariance
$\underbrace{(X_i - \overline{X})}_{>0}\ \underbrace{(Y_i - \overline{Y})}_{<0}$
Negative contribution to covariance
x
$\overline{X}$
Negative contribution to covariance
$\underbrace{(X_i - \overline{X})}_{<0}\ \underbrace{(Y_i - \overline{Y})}_{>0}$
Positive contribution to covariance
$\underbrace{(X_i - \overline{X})}_{>0}\ \underbrace{(Y_i - \overline{Y})}_{>0}$
$\overline{Y}$
$\underbrace{(X_i - \overline{X})}_{<0}\ \underbrace{(Y_i - \overline{Y})}_{<0}$
Positive contribution to covariance
$\underbrace{(X_i - \overline{X})}_{>0}\ \underbrace{(Y_i - \overline{Y})}_{<0}$
Negative contribution to covariance
$\overline{X}$

(a)

(b)

Figure 3.2.1:

3.3 Correlation

3.4 Central Limit Theorem

Chapter 4

Linear Regression

4.1 General Notes

With regression, we want to use one set of numbers to make a preiction on the value of another set

Correlation is a part of what we need for regression. Regression and correlation give related, but distinct information

Correlation gives you a measurement that can be interpreted independently of the scale of two variables. It is, by definition, bounded by ±1. If two variables are very highly correlated, then the correlation is close to 1. Similarly, if two variables are highly correlated but with opposing signs, then the correlation is close to -1.

When you do a regression analysis, the slope of the fitted line could not give you information about the correlation. The slope of the regression line is meant to tell you the expected change in the dependent variable with a unit change in the independent variable X. You cannot calculate this from the correlation alone.

The salient point from this discussion is that the correlation is unit-less whereas the slope of the regression curve has units of Y/X.

4.2 General Linear Model with Single Variable

First – good discussion about the difference between a functional relationship and a statistical relationship.

From Kutner et. al. *Applied Linear Statistical Models*: A regression model is a formal means of expressing the two essential ingredients of a statistical relation.

1. A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion.

2. A scattering of points around the curve of statistical relationship.

These two characteristics are embodied in a regression model by postulating that:

1. There is a probability distribution of Y for each level of X.

2. The means of these probability distributions vary in some systematic fashion with X.

Excellent graph in Figure 1.4 of Kutner that helps you think about this analysis: for each X there is a probability distribution of Y. Then at each X, there is a mean of the probability distribution of Y. If you connect these means together, you will get a regression curve.

Regression models may differ in the form of the regression function (linear, curvilinear), in the shape of the probability distributions of Y (symmetrical, skewed), and in other ways. Whatever the variation, the concept of a probability distribution of Y for any given X is the formal counterpart to the empirical scatter in a statistical relation. Similarly, the regression curve, which describes the relation between the means of the probability distributions of Y and the level of X, is the counterpart to the general tendency of Y to vary with X systematically in a statistical relation.

Let's first tackle the case where there is only one predictor variable, X and the regression function is linear. The model can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{4.2.1}$$

There is an alternative way of writing the model which will come in handy when we discuss the matrix formulation of regression

$$Y_i = X_0\beta_0 + \beta_1 X_i + \epsilon_i \tag{4.2.2}$$

where we define $X_0 = 1$.

For this model, we define

- The index i runs from $1, \ldots, n$. Note that unless specified otherwise, the summations in this model are $\sum_i^n$.
- X_i the value of the predictor variable in the *ith* trial. X_i is a known constant.
- Y_i the value of the response variable in the *ith* trial or when $X = X_i$.
- β_0, β_1 parameters of the regression function.
- ϵ_i is a random error term with the following properties:
  - The expectation value of the error terms is zero:

    $$E(\epsilon_i) = 0 \tag{4.2.3}$$

  - The variance of the error terms is

    $$\sigma^2(\epsilon_i) = \sigma^2 \tag{4.2.4}$$

  - The error terms are uncorrelated, so that the co-variance is zero for all i, j such that $i \neq j$.

    $$Cov(\epsilon_i, \epsilon_j) = 0 \tag{4.2.5}$$

Figure 4.2.1 is recreated from *Kutner* and is a great illustration of these terms.

Let's explore a few salient points about this model:

1. Its important to note that the model given in Eqn.(4.2.1) is not the regression function yet. That is, we have just stated what the model is, but have not determined the optimal parameters for β_0, β_1 that will provide the "best fit" to the data.

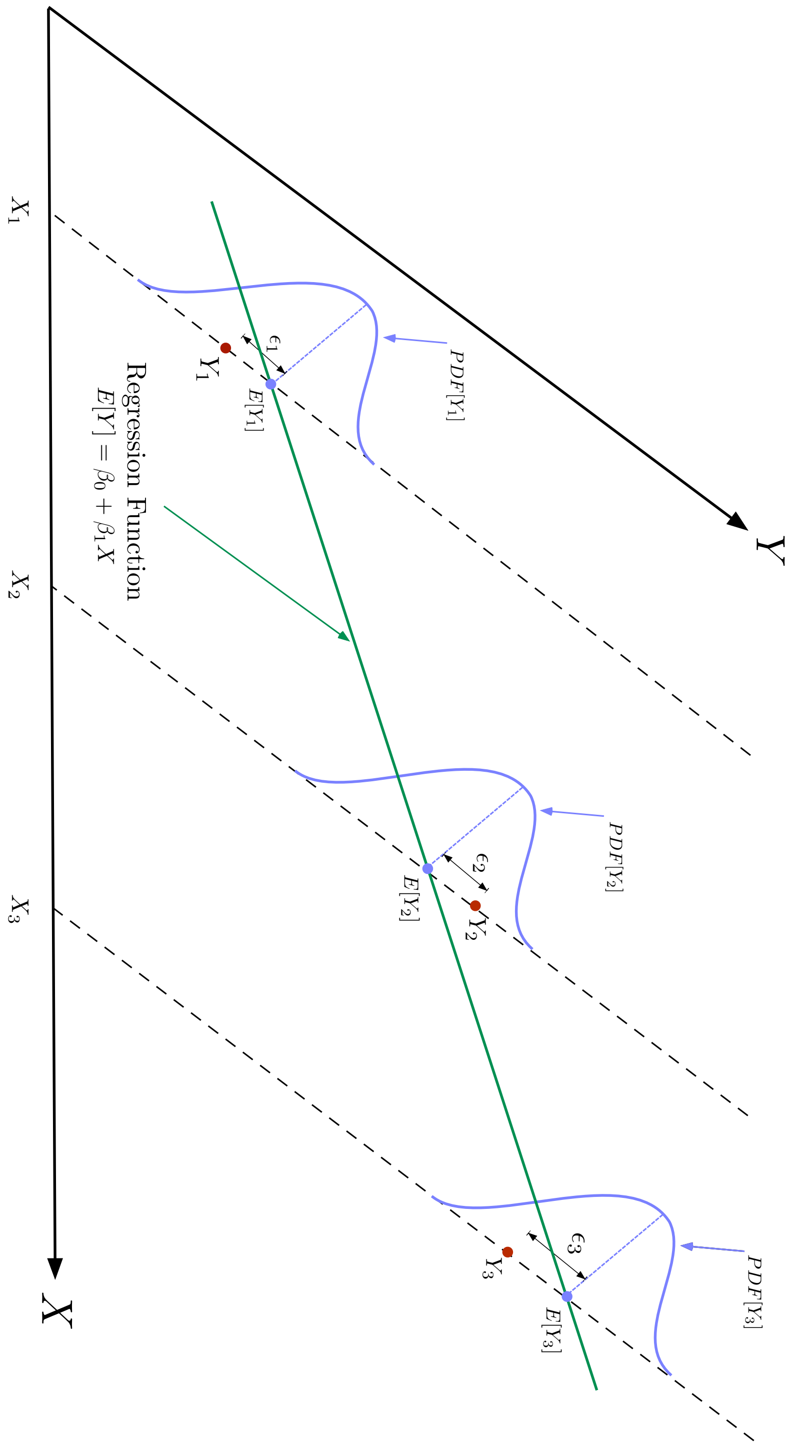

Y
X
X_1
X_2
X_3
Regression Function
$E[Y] = \beta_0 + \beta_1 X$
Y_1
$E[Y_1]$
ϵ_1
$PDF[Y_1]$
Y_2
$E[Y_2]$
ϵ_2
$PDF[Y_2]$
Y_3
$E[Y_3]$
ϵ_3
$PDF[Y_3]$

Figure 4.2.1:

2. This model is linear in the predictor variable since X_i only appears in the first power. This is also referred to as a **first-order model**.

3. The response, Y_i in the i^{th} trial is the sum of two components: (1) the constant term $\beta_0 + \beta_1 X_i$ and (2) the random term ϵ_i. Therefore, Y_i itself is a random variable. It follows that

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= E(\beta_0) + E(\beta_1 X_i) + E(\epsilon_i) \\ &= \beta_0 + \beta_1 X_i + 0 \end{aligned}$$

   or

$$E(Y_i) = \beta_0 + \beta_1 X_i \tag{4.2.6}$$

   In the above steps, we have made use of the expectation value of the sum of random variables is the sum of the expectation value of each variable, as was discussed in Section XYZ in Eqn.(**??**).

   We interpret this result as follows: the response Y_i, when the level of X in the i^{th} trial is X_i, comes from a probability distribution whose mean is $E(Y_i) = \beta_0 + \beta_1 X_i$. This is shown in Figure 4.2.1 as well.

   Note that we are not saying Y_i will be exactly $\beta_0 + \beta_1 X_i$, since each observed Y_i is given by Eqn.(4.2.1). The observed value will higher than, lower than, or equal to $\beta_0 + \beta_1 X_i$. For a given X_i, the *distribution* of Y_i values has a mean given by $\beta_0 + \beta_1 X_i$.

4. The regression model or the **regression function** for the model given in Eqn.(4.2.1) is a collection of these means. In Figure 4.2.1, the regression function is shown as the green line that connects all the points $E(Y) = \beta_0 + \beta_1 X$.

5. The response Y_i in the i^{th} trial differs from the value of the regression term by the error ϵ_i. The error terms are assumed to have a constant variance σ_ϵ^2. Therefore, the responses Y_i have the same constant variance; that is

$$\sigma_{Y_i}^2 = \sigma_\epsilon^2 \tag{4.2.7}$$

This is an important assumption to keep in mind about the model. The regression models assumes that the probability distributions of Y_i have the same variance, regardless of the level of the predictor variable X_i.

6. The error terms are assumed to be uncorrelated. Since the error terms ϵ_i and ϵ_j are uncorrelated, so are the responses Y_i and Y_j.

4.2.1 Why is $E(\epsilon_i) = 0$?

- It is an assumption we made in the model

- Basically, the errors represent everything that the model does not have into account. And why is that? Because it would be extremely unlikely for a model to perfectly predict a variable, as it is impossible to control every possible condition that may interfere with the response variable. The errors may also include reading or measuring inaccuracies. Considering the regression line of best fit, the errors are based on the distance from each point to that line.

The Central Limit Theorem is behind the assumption of the errors following a normal distribution. It states that the distribution of the sum of a large number of independent random variables will tend towards a normal distribution. And actually, in the real world, the majority of the observable errors appear to be distributed that way; which helps us to extrapolate to the unobservable errors.

Another assumption made is that each data point has its own independent associated error, i.e., the errors are independent from one another, which helps us assume they occur randomly.

And because the errors occur randomly, it is expected each data point has equal probability of appearing above or bellow the line of best fit created by the regression (positive error values for the data points with a higher value than the one predicted by the line, and negative error values for the data points with a smaller value predicted by the line), meaning if you summed up every error it would result in a value very close to zero.

4.3 Method of Least Squares

The **Method of Least Squares** is a technique which can identify the "optimal" parameters β_i for the regression function. Note that this technique does not, a priori, tell us what model function (i.e., linear, curvilinear, parabolic) to use. Once we have selected the function, it gives us a way to find the optimal values of the β parameters that fit the data.

Once we employ this method and find the parameters, we have a specific regression function. This technique uses all of the data to come up with a function that will give the $E(Y_i)$ at any given point. It's kinda cool if you think about it; the value of $E(Y_i)$ was influenced by all the other points in the data set.

4.3.1 Derivation of Regression Parameters

To understand how this technique works, we first consider the deviation of Y_i from its expected value, the latter being Eqn.(4.2.6). That is,

$$\begin{aligned} q_i &= Y_i - E(Y_i) \\ &= Y_i - (\beta_0 + \beta_1 X_i) \end{aligned} \tag{4.3.1}$$

As the name of this technique implies, we are going to be interested in the square of this expression

$$(q_i)^2 = [Y_i - (\beta_0 + \beta_1 X_i)]^2 \tag{4.3.2}$$

We now define a quantity that is the sum of these squares

$$\begin{aligned} Q &= \sum_i^n (q_i)^2 \\ &= \sum_i^n [Y_i - (\beta_0 + \beta_1 X_i)]^2 \end{aligned} \tag{4.3.3}$$

In regression analysis, Q is sometimes referred to as the **cost function**. We use the squares because it is possible that the magnitude of all the positive deviations of Y_i from the mean get cancelled out by the magnitude of the negative deviations. That is, if we had used the sum of the q_i, then its possible that this could be zero. Every term in Eqn.(4.3.3) contributes a positive value to the sum. Recall this is similar to the discussion in Section XYZ when we first discuss the variance.

When first learning about regression analysis, sometimes people get the cost function and regression model confused. They are, of course, two very different things. The linear regression model in the simplest case is a linear in its dependence on the regression parameters (as we shall see, the predictor variables, X, can be expressed as polynomials or other forms). However, Q is defined to be a positive, quadratic function.

The method of least squares states that the optimal values for β will minimize the value of Q. That is, we set the partial derivatives of Q with respect to the β parameters equal to zero and solve for β_0 and β_1.

We start with first determining the partial derivatives of Q with respect to the regression parameters.

$$\frac{\partial Q}{\partial \beta_0} = 2\sum_i^n \left[Y_i - (\beta_0 + \beta_1 X_i)\right](-1) \tag{4.3.4}$$

and

$$\frac{\partial Q}{\partial \beta_1} = 2\sum_i^n \left[Y_i - (\beta_0 + \beta_1 X_i)\right](-X_i) \tag{4.3.5}$$

We are interested in minimizing Q, so we set these derivatives equal to zero.

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= 2\sum_i^n \left[Y_i - (b_0 + b_1 X_i)\right](-1) = 0 \\ \frac{\partial Q}{\partial \beta_1} &= 2\sum_i^n \left[Y_i - (b_0 + b_1 X_i)\right](-X_i) = 0 \end{aligned} \tag{4.3.6}$$

Following *Kutner*, we have made a change in notation here. These are two equations with two unknowns; the solution to the values of β_0, β_1 that will satisfy these two equations are designated as b_0, b_1. The distinction here is that β_0, β_1 are arbitrary values of the regression model, whereas b_0, b_1 are specific values that minimize Q through least squares regression.

We start with getting an expression for b_0 that will really come in handy

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= 0 \\
2\sum_i^n \left[Y_i - (b_0 + b_1 X_i)\right](-1) &= 0 \\
\sum_i^n \left[Y_i - b_0 - b_1 X_i)\right] &= 0 \\
\sum_i^n Y_i - nb_0 - b_1 \sum_i^n X_i &= 0
\end{aligned} \tag{4.3.7}$$

which leads to

$$\sum_i^n Y_i = nb_0 + b_1 \sum_i^n X_i \tag{4.3.8}$$

Similarly, a handy expression for b_1 is derived as follows

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_1} &= 0 \\
2\sum_i^n \left[Y_i - (b_0 + b_1 X_i)\right](-X_i) &= 0 \\
\sum_i^n X_i \left[Y_i - (b_0 + b_1 X_i)\right] &= 0 \\
\sum_i^n X_i Y_i - b_0 \sum_i^n X_i - b_1 \sum_i^n X_i^2 &= 0
\end{aligned}$$

which leads to

$$\sum_i^n X_i Y_i = b_0 \sum_i^n X_i + b_1 \sum_i^n X_i^2 \tag{4.3.9}$$

We will eventually derive explicit expressions for b_0 and b_1. According to convention, Eqns.(4.3.8) and (4.3.9) are referred to as the **normal equations** for linear regression. They will be used elsewhere in regression analysis and other derivation.

To solve for b_0, we simply re-arrange terms in Eqn.(4.3.8) to get

$$\begin{aligned} -nb_0 &= -\sum_i^n Y_i + b_1 \sum_i^n X_i \\ b_0 &= \frac{1}{n}\sum_i^n Y_i - b_1 \, \frac{1}{n}\sum_i^n X_i \end{aligned}$$

or

$$\boxed{\boldsymbol{b_0 = \overline{Y} - b_1\overline{X}}} \tag{4.3.10}$$

The derivation of a suitable expression for b_1 is more involved algebraically. Of course, we will work it out in full! We will first derive a few expressions that will aid us in expressing Eqn.(4.3.9) in terms of b_1 in a more intuitive form. Essentially, when we solve the normal equations for b_1, we will want to express it in terms of the variance and covariance of X and Y. It's not obvious from looking at Eqn.(4.3.9), but the variance and covariance are hiding in there; we will tease this out.

The first involves simply rewriting the average of X and Y

$$\begin{aligned} \frac{1}{n}\sum_i^n X_i = \overline{X} \to \sum_i^n X_i = n\overline{X} \\ \frac{1}{n}\sum_i^n Y_i = \overline{Y} \;\; \to \sum_i^n Y_i = n\overline{Y} \end{aligned} \tag{4.3.11}$$

Next, we will see how we can play with the following expression which appears in the covariance of two variables

$$\begin{aligned}\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) &= \sum_i^n \left[X_iY_i - X_i\overline{Y} - \overline{X}Y_i + \overline{X}\ \overline{Y}\right] \\ &= \sum_i^n X_iY_i - \overline{Y}\sum_i^n X_i - \overline{X}\sum_i^n Y_i + \overline{X}\ \overline{Y}\sum_i^n 1 \\ &= \sum_i^n X_iY_i - \overline{Y}\left(n\overline{X}\right) - \overline{X}\left(n\overline{Y}\right) + \overline{X}\ \overline{Y}(n) \\ &= \sum_i^n X_iY_i - n\overline{X}\ \overline{Y} - n\overline{X}\ \overline{Y} + n\overline{X}\ \overline{Y} \\ &= \sum_i^n X_iY_i - n\overline{X}\ \overline{Y}\end{aligned} \tag{4.3.12}$$

where in the third step we have used Eqns.(4.3.11). From here, we can complete the steps in two different ways

$$\begin{aligned}\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) &= \sum_i^n X_iY_i - n\overline{X}\ \overline{Y} \\ &= \sum_i^n X_iY_i - \overline{X}\left(n\overline{Y}\right) \\ &= \sum_i^n X_iY_i - \overline{X}\left(\sum_i^n Y_i\right) \\ &= \sum_i^n \left[X_iY_i - \overline{X}Y_i\right] \\ &= \sum_i^n Y_i\left(X_i - \overline{X}\right)\end{aligned} \tag{4.3.13}$$

Again in the third step we have used Eqns.(4.3.11). We could also equivalently done these steps as follows

$$
\begin{aligned}
\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) &= \sum_i^n X_i Y_i - n\overline{X}\,\overline{Y} \\
&= \sum_i^n X_i Y_i - \overline{Y}\left(n\overline{X}\right) \\
&= \sum_i^n X_i Y_i - \overline{Y}\left(\sum_i^n X_i\right) \qquad (4.3.14) \\
&= \sum_i^n \left[X_i Y_i - \overline{Y} X_i\right] \\
&= \sum_i^n X_i \left(Y_i - \overline{Y}\right)
\end{aligned}
$$

Eqns.(4.3.13) and (4.3.14) may look different, but they are actually equivalent.

Finally, let's see how we can play with the following expression which appears in the variance of two variables

$$
\begin{aligned}
\sum_i^n \left(X_i - \overline{X}\right)^2 &= \sum_i^n \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right) \\
&= \sum_i^n X_i \left(X_i - \overline{X}\right) - \sum_i^n \overline{X}\left(X_i - \overline{X}\right) \\
&= \sum_i^n X_i \left(X_i - \overline{X}\right) - \overline{X}\,\cancelto{0}{\sum_i^n \left(X_i - \overline{X}\right)}
\end{aligned}
$$

We know that $\sum_i^n \left(X_i - \overline{X}\right) = 0$ from Eqn.(**??**); i.e., the sum of the deviations of X_i from the mean is zero. We can now express this in three equivalent ways, all of which we will use at some point. I'm being deliberately pedantic and listing them out as separate derivations only because we will refer to them in the future and it's just easier to see things this way.

1.
$$\sum_{i}^{n}\left(X_i-\overline{X}\right)^2=\sum_{i}^{n}X_i\left(X_i-\overline{X}\right) \tag{4.3.15}$$

2.
$$\begin{aligned}\sum_{i}^{n}\left(X_i-\overline{X}\right)^2&=\sum_{i}^{n}X_i\left(X_i-\overline{X}\right)\\ &=\sum_{i}^{n}\left(X_i^2-\overline{X}X_i\right)\end{aligned} \tag{4.3.16}$$

3.
$$\begin{aligned}\sum_{i}^{n}\left(X_i-\overline{X}\right)^2&=\sum_{i}^{n}\left(X_i^2-\overline{X}X_i\right)\\ &=\sum_{i}^{n}X_i^2-\overline{X}\sum_{i}^{n}X_i\\ &=\sum_{i}^{n}X_i^2-\overline{X}\left(n\overline{X}\right)\\ &=\sum_{i}^{n}X_i^2-n\overline{X}^2\end{aligned} \tag{4.3.17}$$

We now have all the expressions we need to express b_1 is a more friendly format. Starting with Eqn.(4.3.9)

$$\begin{aligned}\sum_{i}^{n}X_iY_i&=b_0\sum_{i}^{n}X_i+b_1\sum_{i}^{n}X_i^2\\ &=b_0\left(n\overline{X}\right)+b_1\sum_{i}^{n}X_i^2\end{aligned} \tag{4.3.18}$$

Once again, in the second step we have used Eqns.(4.3.11). We can now use Eqn.(4.3.10) for b_0 and continue

$$\begin{aligned}\sum_{i}^{n}X_iY_i&=\left(\overline{Y}-b_1\overline{X}\right)\left(n\overline{X}\right)+b_1\sum_{i}^{n}X_i^2\\ &=n\overline{X}\,\overline{Y}-nb_1\overline{X}^2+b_1X_i^2 \qquad (4.3.19)\\ \sum_{i}^{n}X_iY_i-n\overline{X}\,\overline{Y}&=b_1\left(-n\overline{X}^2+\sum_{i}^{n}X_i^2\right)\end{aligned}$$

which leads to

$$b_1 = \frac{\sum_i^n X_i Y_i - n\overline{X}\ \overline{Y}}{\sum_i^n X_i^2 - n\overline{X}^2} \tag{4.3.20}$$

We will now express the numerator and denominator of Eqn.(4.3.20) to expressions that we have derived above. Starting with the numerator

$$\begin{aligned}\sum_i^n X_i Y_i - n\overline{X}\ \overline{Y} &= \sum_i^n X_i Y_i - \overline{X}\ \left(n\overline{Y}\right) \\ &= \sum_i^n X_i Y_i - \overline{X}\sum_i^n Y_i \\ &= \sum_i^n \left(X_i Y_i - \overline{X} Y_i\right) \\ &= \sum_i^n Y_i \left(X_i - \overline{X}\right)\end{aligned} \tag{4.3.21}$$

The above result is actually Eqn.(4.3.13). So the numerator of b_1 in Eqn.(4.3.20) is

$$\sum_i^n X_i Y_i - n\overline{X}\ \overline{Y} = \sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

Now the denominator of Eqn.(4.3.20) can be expressed as

$$\begin{aligned}\sum_i^n X_i^2 - n\overline{X}^2 &= \sum_i^n X_i^2\ -\ \overline{X}\left(n\overline{X}\right) \\ &= \sum_i^n X_i^2 -\ \overline{X}\left(\sum_i^n X_i\right) \\ &= \sum_i^n \left(X_i^2 - \overline{X} X_i\right)\end{aligned}$$

This is the same result we derived in Eqn.(4.3.16). So the denominator of b_1 in Eqn.(4.3.20) is

$$\sum_i^n X_i^2 - n\overline{X}^2 = \sum_i^n \left(X_i - \overline{X}\right)^2$$

Making these substitutions for the numerator and denominator into Eqn.(4.3.20) we get

$$\boxed{\boldsymbol{b_1} = \frac{\sum_i^n \left(\boldsymbol{X_i} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}}\right)}{\sum_i^n \left(\boldsymbol{X_i} - \overline{\boldsymbol{X}}\right)^{\boldsymbol{2}}}} \tag{4.3.22}$$

The numerator is n times the covariance of X and Y. The denominator is n times the variance of X. So we can intuitively express

$$b_1 = \frac{Cov(X,Y)}{Var(X)} \tag{4.3.23}$$

Now we see the purpose of all of this laborious algebra! While there are many ways of expressing Eqn.(4.3.22), writing it in terms of the covariance of X, Y and the variance of X is the most helpful.

4.3.2 Notation and Terminology Interlude

Since *Kutner* is overly pedantic about the notation, let's follow their lead and take a moment to revisit our terminology.

The values of b_0, b_1 are also referred to as **point estimators** of the regression function, β_0, β_1, respectively. Remember that β_0, β_1 can take on any value; b_0 and b_1 are specific outcomes from minimizing the the cost function Q.

Similarly, we will refer to an estimate of the regression function as

$$\hat{Y} = b_0 + b_1 X \tag{4.3.24}$$

For any given value X_i, we will call $\hat{Y}_i$ the **fitted value** of Y for the i^{th} case and calculate as follows

$$\hat{Y}_i = b_0 + b_1 X_i \qquad i = 1, \ldots, n \tag{4.3.25}$$

We have referred to the value of the response variable (i.e., the actual observed value) as *the response* and called it Y_i. We have referred to $E(Y)$ as the mean response and it is the mean of the probability distribution of Y

corresponding to the level X of the predictor variable. Therefore $\hat{Y}$ is the point estimator of the mean response when the level of the predictor variable is X.

The fitted value $\hat{Y}_i$ is distinct from the observed value Y_i. We will explore the difference between these two values when we discuss residuals.

Kutner p.21: Extension of the Gauss-Markov Theorem: $\hat{Y}$ is an unbiased estimator of $E(Y_i)$, with minimum variance in the class of unbiased linear estimators.

4.3.3 Regression Through the Origin

Sometimes we know *a priori* that the regression function has to go through the origin. For example, if the input variable X is amount of inventory in a store and response variable Y is the sales, we can safely say that if there is zero inventory there are no sales!

The regression model for this case is the same as Eqn.(4.2.1) except that the y-intercept is zero so $\beta_0 = 0$. So the regression model reduces to

$$Y_i = \beta_1 X_i + \epsilon_i \tag{4.3.26}$$

The parameters have the same definition as before. For this case, the regression function is given by

$$E(Y_i) = \beta_1 X_i \tag{4.3.27}$$

Finding the point estimator is more straightforward than before. As we did previously, we define q_i as follows

$$\begin{aligned} q_i &= Y_i - E(Y_i) \\ &= Y_i - \beta_1 X_i \end{aligned} \tag{4.3.28}$$

We now define a quantity that is the sum of these squares

$$\begin{aligned} Q &= \sum_i^n (q_i)^2 \\ &= \sum_i^n \left[Y_i - \beta_1 X_i\right]^2 \end{aligned} \tag{4.3.29}$$

To determine the point estimator, we minimize Q with respect to β_1 and set it to zero

$$\frac{\partial Q}{\partial \beta_1} = 0$$

$$2\sum_i^n \left[Y_i - \beta_1 X_i\right](-X_i) = 0$$

$$\sum_i^n X_i \left[Y_i - \beta_1 X_i\right] = 0$$

$$\sum_i^n X_i Y_i - \beta_1 \sum_i^n X_i^2 = 0$$

or

$$\beta_1 = b_1 = \frac{\sum_i^n X_i Y_i}{\sum_i^n X_i^2} \tag{4.3.30}$$

Following the format introducded for the general linear case, we may also write the regression function as

$$\hat{Y} = b_1 X \tag{4.3.31}$$

4.3.4 Residuals

p.22 and Chapter 3 of Kutner. Discuss here and also when complete the LR matrix discussion.

The residuals are a very useful way of determining whether a given regression model is appropriate for the data we have.

We first make a distinction between the error in the model term $\epsilon_i = Y_i - E(Y_i)$ and the residual, $e_i = Y_i - \hat{Y}_i$. The ϵ_i represents the vertical deviation of Y_i from the unknown true regression line, and is therefore unknown. However, the e_i are the vertical deviation of Y_i from the fitted value $\hat{Y}_i$. We know the regression curve after we calculate the point estimators, we can calculate the residuals e_i for each point i; therefore the e_i are known.

4.3.5 Properties of Least Squares Regression

It's worth spending some time to evaluate the intuition and consequences of these regression results.

- Note that if $\beta_1 = 0$, then $\beta_0 = \overline{Y}$. Intuitively, we know that if $\beta_1 = 0$, the Y has no dependence on X. That is another way of saying that all of the Y values are scattered vertically. In this case, the value of Y that will minimize the sum of the squares is $\overline{Y}$. We will discuss this further in Section XYZ.

  Reference discussion from R^2 regression that $y = \mu$ minimizes the sum of the squares. This discussion is another way of saying that.

- In the regression model, β_0 and β_1 are the regression parameters. According to the model, β_1 is the slope of the regression line and β_0 is the y-intercept. When the scope of the model data includes $X = 0$, β_0 gives the mean of the probability distribution at $X = 0$. When the scope of the model does not cover $X = 0$, β_0 does not have any meaning as a separate term in the regression model.

- Further discussion – does it make intuitive sense that b_1 would be the ratio of the covariance and the variance? Yes – at least the units work out; units are y/x.

There are several properties of the fitted regression line that follow directly from the least squares normal equations Eqns.(4.3.8) and (4.3.9). We will cover each of these and provide the proof.

- The sum of the residuals equals zero:

$$\sum_{i=1}^{n} e_i = 0 \tag{4.3.32}$$

  Proof: We know that

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ &= Y_i - (b_0 + b_1 X_i) \end{aligned} \tag{4.3.33}$$

Plugging this back into the sum of the residuals gives us

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)] \\ &= \sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i \\ &= 0\end{aligned} \tag{4.3.34}$$

where in the last step we have used the fist normal equation Eqn.(4.3.8).

- The sum of the squared residuals, $\sum_{i=1}^{n} e_i^2$, is a minimum when the regression parameters are b_0, b_1.

  Proof: This is a direct consequence of the minimization of the cost function. First, we write the expression for the sum of squared residuals, we get

  $$\sum_{i=1}^{n} e_i^2 = [Y_i - (b_0 + b_1 X_i)]^2 \tag{4.3.35}$$

  Comparing this directly to Eqn.(4.3.3) we see that is the same expression for the cost function, except that Eqn.(4.3.3) was expressed in terms of β. We already derived that the cost function is minimized when $\beta_0 \to b_0$ and $\beta_1 \to b_1$. Inserting these into the cost function gives us that

  $$\begin{aligned}Q &= \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \\ Q_{minimum} &= Q\left(\beta_0 \to b_0, \beta_1 \to b_1\right) \\ &= \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2 \\ &= \sum_{i=1}^{n} e_i^2\end{aligned} \tag{4.3.36}$$

- The sum of the observed values Y_i equals the sum of the fitted values $\hat{Y}_i$.

Proof: We work out the expression for the sum of the fitted values explicitly

$$\begin{aligned}\sum_{i=1}^{n} \hat{Y}_i &= \sum_{i=1}^{n} [b_0 + b_1 X_i] \\ &= nb_0 + b_1 \sum_{i=1}^{n} X_i \\ &= nb_0 + b_1 n\overline{X} \\ \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i &= b_0 + b_1\overline{X} \\ &= \overline{Y}\end{aligned} \tag{4.3.37}$$

where in the last step we have used the first normal equation Eqn.(4.3.10). Now, simply re-writing $\overline{Y}$ using Eqn.(4.3.11), we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i &= \overline{Y} \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} \hat{Y}_i &= \sum_{i=1}^{n} Y_i\end{aligned} \tag{4.3.38}$$

- The sum of the residuals weighted by X_i is zero. That is

$$\sum_{i=1}^{n} X_i e_i = 0 \tag{4.3.39}$$

Proof: We work out the weighted sum explicitly

$$\begin{aligned}\sum_{i=1}^{n} X_i e_i &= \sum_{i=1}^{n} X_i \left(Y_i - \hat{Y}_i\right) \\ &= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \hat{Y}_i \\ &= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \left(b_0 + b_1 X_i\right) \\ &= \sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2\end{aligned} \tag{4.3.40}$$

For the first term on the left, we can substitute the result from the second normal equation Eqn.(4.3.9). We then get

$$\begin{aligned}\sum_{i=1}^{n} X_i e_i &= \left[b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2\right] - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 \\ &= \left[b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2\right] - \left[b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2\right] \\ &= 0\end{aligned} \tag{4.3.41}$$

- The sum of the residuals weighted by $\hat{Y}_i$ is zero. That is

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0 \tag{4.3.42}$$

Proof: This will follow directly from Eqns.(4.3.32) and (4.3.41). We work out the weighted sum explicitly

$$\begin{aligned}\sum_{i=1}^{n} \hat{Y}_i e_i &= \sum_{i=1}^{n} \left(b_0 + b_1 X_i\right) e_i \\ &= \sum_{i=1}^{n} b_0 e_i + \sum_{i=1}^{n} b_1 X_i e_i \\ &= b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} X_i e_i \\ &= 0\end{aligned} \tag{4.3.43}$$

where the first term on the right is zero from Eqn.(4.3.32) and the second term on the right is zero because of Eqn.(4.3.41).

- The regression line always goes through the point $(\overline{X}, \overline{Y})$.

  Proof: The fitted line is given by $\hat{Y} = b_0 + b_1 X_i$. When $X_i = \overline{X}$, we get

  $$\begin{aligned} \hat{Y} &= b_0 + b_1 \overline{X} \\ &= \overline{Y} \end{aligned} \tag{4.3.44}$$

  where in the last step we have used Eqn.(4.3.10).

4.4 Least Squares Regression - Examples

- Use my made up example
- Use the Talouca Company example in Kutner
- Any examples worked out in Mathematica
- Examples from Kutner
- Since by now we would have covered variance, covariance, and correlation, we can tie these all together with linear regression. Verify that our "calculation done by hand" matches the output from Excel and Mathematica and the formulas are consistent.

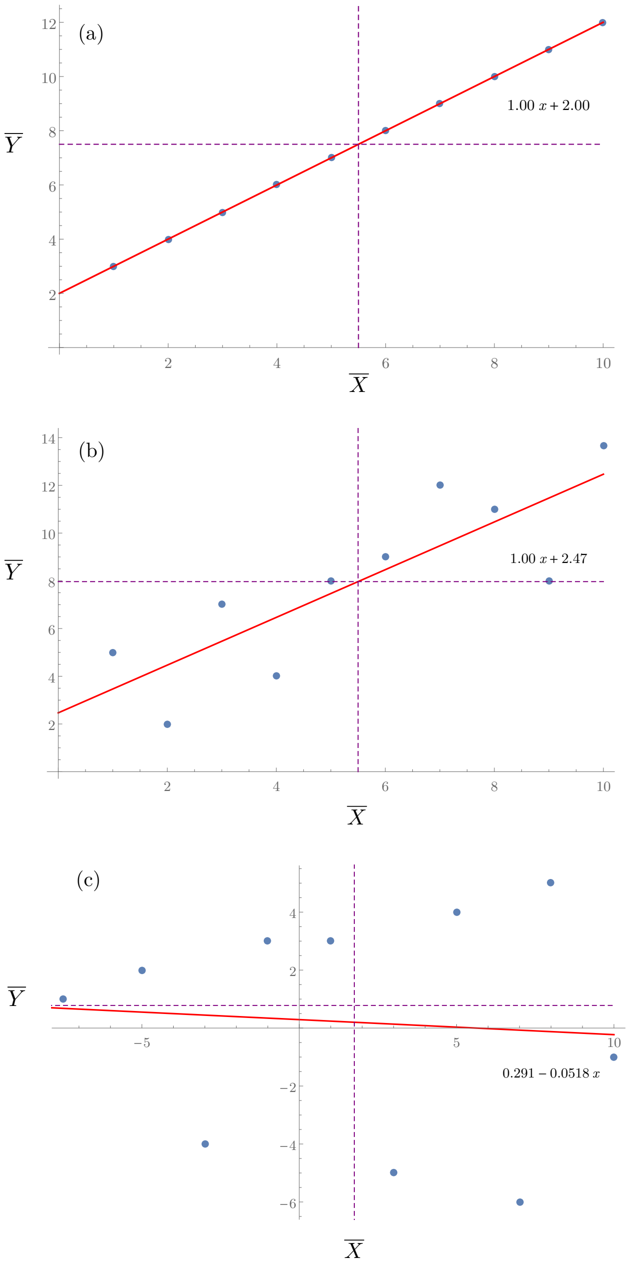

(a)
12
10
8
6
4
2
$\overline{Y}$
1.00 x + 2.00
2
4
6
8
10
$\overline{X}$
(b)
14
12
10
8
6
4
2
$\overline{Y}$
1.00 x + 2.47
2
4
6
8
10
$\overline{X}$
(c)
4
2
−2
−4
−6
$\overline{Y}$
−5
5
10
0.291 − 0.0518 x
$\overline{X}$

Figure 4.4.1:

4.5 R^2 - The “Goodness” of the Fit

4.6 Matrix Formulation of Linear Regression

4.6.1 Regression Model Setup

We'll now move another step closer to generalizing the linear regression problem by introducing linear algebra. Specifically, expressing the linear regression problem in matrix notation accomplishes a few key things. First, it gives us the ability to express enormous amounts of data from trials or observations in a compact form. It also gives us a structure for handling more complex multi-variate problems, where the response function is influenced by many input variables, in the same familiar, compact form.

This is the roadmap we will follow for developing the matrix formulation for linear regression for a single variable.

1. Work out the formulation of the regression model in matrix terms.

2. We will "proactively" calculate some needed matrix products, so we have them handy to insert into derivations.

3. Derive the normal equations in matrix form

4. Calculate the inverse of an important matrix $\left(X^T X\right)^{-1}$

5. Work out the solution for the point estimators.

For this formulation, we'll make use of the alternate form of the regression model from Eqn.(4.2.2). Specifically, for a set of data, we can write

$$\begin{aligned} Y_1 &= \beta_0 X_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 X_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots & \qquad\quad \vdots \\ Y_n &= \beta_0 X_0 + \beta_1 X_n + \epsilon_n \end{aligned} \tag{4.6.1}$$

To get this into a matrix form we can work with, we define the response as a vector, Y, consisting of n observations of the response variable

$$\underset{n\times 1}{\boldsymbol{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \tag{4.6.2}$$

Although we are not working on multi-variate linear regression yet, we have (technically speaking) broached the subject of multiple variables by introduced X_0 into the mix. Nevermind that we have defined $X_0 = 1$; it's a variable nonetheless! So we define the following notation for X

$$\underset{n\times 2}{\boldsymbol{X}} = \begin{bmatrix} X_{10} & X_{11} \\ X_{20} & X_{21} \\ \vdots & \vdots \\ X_{i0} & X_{i1} \\ \vdots & \vdots \\ X_{n0} & X_{n1} \end{bmatrix} \tag{4.6.3}$$

The above expression for the preidctor variable is sometimes referred to as the **design matrix**. Here the notation can be interpreted as

$$X_{ij} \rightarrow \text{The value of the } j^{th} \text{ variable during the } i^{th} \text{ trial} \tag{4.6.4}$$

By introducing this notation, we are foreshadowing the more general case which we will cover in Section XYZ. Think of the first column of this matrix as all of the values of X_{i0} for the various i trials. Since X_0 is defined to be 1, the entire left column can be set to 1. Similarly, the second column can be thought of all of the values of the variable X_1 for the various i trials.

In this section, we are building the matrix notation for the case where Y is dependent on only one predictor variable that actually changes in value. That is, Y will vary with $X_j = X_1$ only. So in this section, we will remove the j subscript for this discussion simply refer to the various trial values of

X as X_i. Therefore, we simplify the design matrix from Eqn.(4.6.3) as

$$\underset{n\times 2}{\boldsymbol{X}} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \tag{4.6.5}$$

Continuing on, we define the regression parameters, β, as a vector

$$\underset{2\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{4.6.6}$$

And finally (for now, at least!), we define the error terms, ϵ, as a vector

$$\underset{n\times 1}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{4.6.7}$$

With these defintions, we can write Eqns.(4.6.1) compactly as follows

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.6.8}$$

As a sanity check, let's work this explicitly; that is, does the above equation lead to Eqns.(4.6.1)? First sanity check is to see if the dimensions of the matrix work out

$$\begin{aligned} \underset{n\times 1}{\boldsymbol{Y}} &= \underset{n\times 2}{\boldsymbol{X}} \quad \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}} \\ n \times 1 &= \quad n \times 1 \quad + \quad n \times 1 \end{aligned} \tag{4.6.9}$$

The main check here was to ensure that the dimensions of matrix $\boldsymbol{X\beta}$ are indeed $n \times 1$. Now we work out the matrix math explictly

$$
\begin{aligned}
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\
&= \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \qquad (4.6.10) \\
&= \begin{bmatrix} \beta_0 + \beta_1 X_1 + \epsilon_1 \\ \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots \\ \beta_0 + \beta_1 X_n + \epsilon_n \end{bmatrix}
\end{aligned}
$$

Matching the the rows from the response variable on the left to the corresponding ones on the right, we see that we do indeed recreate Eqns.(4.6.1).

Note that matrix $\boldsymbol{X\beta}$ is equivalent to the vector of the expected values of Y_i. Since $E(Y_i) = \beta_0 + \beta_1 X_i$, we can write the expectation values vector as

$$
E(\boldsymbol{Y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} \qquad (4.6.11)
$$

Just to close the loop, we may write the response vector as

$$
\boldsymbol{Y} = E(\boldsymbol{Y}) + \boldsymbol{\epsilon} \qquad (4.6.12)
$$

That is, the response vector is the sum of two vectors: the vector containing the expected values of Y and another vector containing the error terms.

Without discussing in detail right now, we'll also write the variance-covariance matrix of error terms for this case as

$$\sigma^2(\epsilon) = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & 0 \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \tag{4.6.13}$$

This is a scalar matrix that can also be written as

$$\sigma^2(\epsilon) = \sigma^2 \boldsymbol{I} \tag{4.6.14}$$

4.6.2 Matrix Interlude

With the response vector $\boldsymbol{Y}$ and the design matrix $\boldsymbol{X}$ defined, it will be useful to work out some key products of matrices that will involve these two. Since they will come up in several of the derivations we will do, it'll be handy have these worked out ahead of time so that we can keep things flowing (pedagogically speaking!).

First we calculate $\boldsymbol{Y}^T\boldsymbol{Y}$. If $\boldsymbol{Y}$ is a $n \times 1$ vector, then $\boldsymbol{Y}^T$ is a $1 \times n$ vector, and therefore $\boldsymbol{Y}^T\boldsymbol{Y}$ is a 1×1 vector, which is a scalar. We can calculate this product as

$$\begin{aligned} \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} Y_1^2 + Y_1^2 + \ldots + Y_n^2 \end{bmatrix} \\ &= \sum_i^n Y_i^2 \end{aligned} \tag{4.6.15}$$

So $\boldsymbol{Y}^T\boldsymbol{Y}$ is a compact way of writing the sum of squared terms.

Next we calculate $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{X}$. To verify the dimensions, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $n \times 2$ matrices and $\boldsymbol{X}^T$ is a $2 \times n$ matrix. So both $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{X}$ will be 2×2 matrices.

These matrices can be calculated as follows

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \\ &= \begin{bmatrix} n & \sum_i^n X_i \\ \\ \sum_i^n X_i & \sum_i^n X_i^2 \end{bmatrix}\end{aligned} \tag{4.6.16}$$

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{Y} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \\ &= \begin{bmatrix} \sum_i^n Y_i \\ \\ \sum_i^n X_i Y_i \end{bmatrix}\end{aligned} \tag{4.6.17}$$

In regression analysis, one of the key matrices we will have to calculate is $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$; i.e., the inverse of Eqn.(4.6.16). Referring to Eqn.(**??**), we need to first calculate the determinant of the $\left(\boldsymbol{X}^T\boldsymbol{X}\right)$ using Eqn.(**??**)

$$\begin{aligned}
D &= n\sum_i^n X_i^2 - \left(\sum_i^n X_i\right)\left(\sum_i^n X_i\right) \\
&= n\left[\sum_i^n X_i^2 - \frac{\left(\sum_i^n X_i\right)\left(\sum_i^n X_i\right)}{n}\right] \\
&= n\left[\sum_i^n X_i^2 - \frac{\left(\sum_i^n X_i\right)}{n}\left(\sum_i^n X_i\right)\right] \\
&= n\left[\sum_i^n X_i^2 - \overline{X}\sum_i^n X_i\right] \\
&= n\left[\sum_i^n X_i\left(X_i - \overline{X}\right)\right] \\
&= n\sum_i^n \left(X_i - \overline{X}\right)^2
\end{aligned} \tag{4.6.18}$$

Where in the last step we have used Eqn.(4.3.15). We can now use this result and Eqn.(**??**) to calculate $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$

$$\begin{aligned}
\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} &= \begin{bmatrix} \frac{\sum_i^n X_i^2}{D} & -\frac{\sum_i^n X_i}{D} \\ -\frac{\sum_i^n X_i}{D} & \frac{n}{D} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\sum_i^n X_i^2}{n\sum_i^n \left(X_i - \overline{X}\right)^2} & -\frac{\sum_i^n X_i}{n\sum_i^n \left(X_i - \overline{X}\right)^2} \\ -\frac{\sum_i^n X_i}{n\sum_i^n \left(X_i - \overline{X}\right)^2} & \frac{n}{n\sum_i^n \left(X_i - \overline{X}\right)^2} \end{bmatrix}
\end{aligned} \tag{4.6.19}$$

We can simplify this matrix. Let's take look at the various elements. First we look at $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{22}$

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{22} = \frac{n}{n\sum_i \left(X_i - \overline{X}\right)^2} = \frac{1}{\sum_i^n \left(X_i - \overline{X}\right)^2} \tag{4.6.20}$$

Next we look at $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{12} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{21}$. Using our favorite expression from Eqn.(4.3.11) and substituting into the numerator, we get

$$\begin{aligned}
\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{12} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{21} &= -\frac{\sum_i^n X_i}{n\sum_i^n \left(X_i - \overline{X}\right)^2} \\
&= -\frac{n\overline{X}}{n\sum_i^n \left(X_i - \overline{X}\right)^2} \\
&= -\frac{\overline{X}}{\sum_i^n \left(X_i - \overline{X}\right)^2}
\end{aligned} \tag{4.6.21}$$

And finally, for $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{11}$, we make use of Eqn.(4.3.17); that is

$$\begin{aligned}
\sum_i^n \left(X_i - \overline{X}\right)^2 &= \sum_i^n X_i^2 - n\overline{X}^2 \\
\sum_i^n \left(X_i - \overline{X}\right)^2 + n\overline{X}^2 &= \sum_i^n X_i^2
\end{aligned} \tag{4.6.22}$$

Plugging the above result into the numerator of the $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{11}$ term gives us

$$\begin{aligned}
\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{11} &= \frac{\sum_i^n X_i^2}{n\sum_i^n \left(X_i - \overline{X}\right)^2} \\
&= \frac{\sum_i^n \left(X_i - \overline{X}\right)^2 + n\overline{X}^2}{n\sum_i^n \left(X_i - \overline{X}\right)^2} \\
&= \frac{1}{n} + \frac{\overline{X}^2}{\sum_i^n \left(X_i - \overline{X}\right)^2}
\end{aligned} \tag{4.6.23}$$

So we can plug all of these back into Eqn.(XYZ) to get

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\overline{X}^2}{\sum_i^n \left(X_i - \overline{X}\right)^2} & -\frac{\overline{X}}{\sum_i^n \left(X_i - \overline{X}\right)^2} \\ -\frac{\overline{X}}{\sum_i^n \left(X_i - \overline{X}\right)^2} & \frac{1}{\sum_i^n \left(X_i - \overline{X}\right)^2} \end{bmatrix} \tag{4.6.24}$$

4.6.3 Normal Equations in Matrix Form

In the earlier section, we derived the normal equations in Eqns.(4.3.8) and (4.3.9). We will now derive the equivalent of those in matrix notation.

We start with the cost function in Eqn.(4.3.3) write it as follows

$$Q = \sum_i^n \left[Y_i - (\beta_0 + \beta_1 X_i)\right]\ \left[Y_i - (\beta_0 + \beta_1 X_i)\right] \tag{4.6.25}$$

If we define a vector

$$\boldsymbol{Y} - \boldsymbol{X\beta} = \begin{bmatrix} Y_1 - (\beta_0 + \beta_1 X_1) \\ Y_2 - (\beta_0 + \beta_1 X_2) \\ \vdots \\ Y_n - (\beta_0 + \beta_1 X_n) \end{bmatrix} \tag{4.6.26}$$

Then we see that

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

$$\begin{bmatrix} Y_1 - (\beta_0 + \beta_1 X_1) & Y_2 - (\beta_0 + \beta_1 X_2) & \cdots & Y_n - (\beta_0 + \beta_1 X_n) \end{bmatrix} \begin{bmatrix} Y_1 - (\beta_0 + \beta_1 X_1) \\ Y_2 - (\beta_0 + \beta_1 X_2) \\ \vdots \\ Y_n - (\beta_0 + \beta_1 X_n) \end{bmatrix}$$

Carrying out the matrix multiplication as usual gives us

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta}) &= [Y_1-(\beta_0+\beta_1X_1)]\,[Y_1-(\beta_0+\beta_1X_1)]\\ &+[Y_2-(\beta_0+\beta_1X_2)]\,[Y_2-(\beta_0+\beta_1X_2)]+\dots\\ &+[Y_n-(\beta_0+\beta_1X_n)]\,[Y_n-(\beta_0+\beta_1X_n)]\\ &=\sum_i^n[Y_i-(\beta_0+\beta_1X_i)]\;\;[Y_i-(\beta_0+\beta_1X_i)]\\ &=Q\end{aligned}$$

So Q can be compactly expressed as shown above. It is worth pointing out that even though we have expressed it in terms of the product of two vectors, it is still a scalar quantity.

As before, we want to minimize Q. We will get Q into a format which will make it easier to take the partial derivates. Using some basic properties of matrices (see Section XYZ), we note that

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X\beta})^T &= \boldsymbol{Y}^T-(\boldsymbol{X\beta})^T\\ &=\boldsymbol{Y}^T-\boldsymbol{\beta}^T\boldsymbol{X}^T\end{aligned}$$

Substituting this into our expression for Q we get

$$\begin{aligned}Q&=(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta})\\ &=\left(\boldsymbol{Y}^T-\boldsymbol{\beta}^T\boldsymbol{X}^T\right)(\boldsymbol{Y}-\boldsymbol{X\beta})\\ &=\boldsymbol{Y}^T\boldsymbol{Y}-\boldsymbol{Y}^T\boldsymbol{X\beta}-\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}+\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}\end{aligned}$$

Note that the dimensions of $\boldsymbol{Y}^T\boldsymbol{X\beta}$ are

$$\text{Dimensions of }\boldsymbol{Y}^T\boldsymbol{X\beta}=(1\times n)(n\times 2)(2\times 1)=1\times 1$$

Since this quantity is a scalar, it is equal to its transpose. That is

$$\begin{aligned}\boldsymbol{Y}^T\boldsymbol{X\beta}&=\left(\boldsymbol{Y}^T\boldsymbol{X\beta}\right)^T\\ &=\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}\end{aligned}$$

Substituting this into the cost function from above, we get

$$\begin{aligned} Q &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned} \tag{4.6.27}$$

To find the value of β_0 and β_1 that minimize Q, we have to calculate the following quantity first

$$\frac{\partial}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \dfrac{\partial Q}{\partial \beta_0} \\ \dfrac{\partial Q}{\partial \beta_1} \end{bmatrix} = 0 \tag{4.6.28}$$

The quantities that we will have to determine are the same as before:

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= 0 \\ \frac{\partial Q}{\partial \beta_1} &= 0 \end{aligned} \tag{4.6.29}$$

Looking at the cost function we derived in Eqn.(4.6.27), we can right away eliminate the derivative of $\boldsymbol{Y}^T\boldsymbol{Y}$ since there is no dependence on β; that is

$$\frac{\partial}{\partial \beta_0}\boldsymbol{Y}^T\boldsymbol{Y} = \frac{\partial}{\partial \beta_1}\boldsymbol{Y}^T\boldsymbol{Y} = 0 \tag{4.6.30}$$

So the task at hand is to find the derivatives with respect to $\boldsymbol{\beta}$ for the other two terms of Q. Let's take them one at a time and we'll leverage some of the results already calculated

First we calculate the derivative of the $-2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}$ term. Since we calculated $\boldsymbol{X}^T\boldsymbol{Y}$ in Eqn.(XYZ), we can use that result to calculate

$$\begin{aligned} -2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} &= -2\begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix} \begin{bmatrix} \sum_i^n Y_i \\ \sum_i^n X_i Y_i \end{bmatrix} \\ &= -2\beta_0 \sum_i^n Y_i - 2\beta_1 \sum_i^n X_i Y_i \end{aligned} \tag{4.6.31}$$

Therefore

$$\begin{aligned}
\frac{\partial}{\partial \beta_0}\left[-2\beta_0\sum_i^n Y_i - 2\beta_1\sum_i^n X_iY_i\right] &= -2\sum_i^n Y_i \\
\frac{\partial}{\partial \beta_1}\left[-2\beta_0\sum_i^n Y_i - 2\beta_1\sum_i^n X_iY_i\right] &= -2\sum_i^n X_i\sum_i^n Y_i
\end{aligned} \tag{4.6.32}$$

or in matrix notation

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}}\left[-2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}\right] &= \begin{bmatrix} -2\sum_i^n Y_i \\ -2\sum_i^n X_iY_i \end{bmatrix} \\
&= -2\boldsymbol{X}^T\boldsymbol{Y}
\end{aligned} \tag{4.6.33}$$

Next we calculate the derivative of the $\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$ term. We use our previously derived result for $\boldsymbol{X}^T\boldsymbol{X}$ as follows

$$\begin{aligned}
\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} &= \begin{bmatrix}\beta_0 & \beta_1\end{bmatrix}\begin{bmatrix} n & \sum_i^n X_i \\ \sum_i^n X_i & \sum_i^n X_i^2 \end{bmatrix}\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} \\
&= \begin{bmatrix}\beta_0 & \beta_1\end{bmatrix}\begin{bmatrix} n\beta_0 + \beta_1\sum_i^n X_i \\ \beta_0\sum_i^n X_i + \beta_1\sum_i^n X_i^2 \end{bmatrix} \\
&= \beta_0\left(n\beta_0\right) + \beta_1\beta_0\sum_i^n X_i + \beta_0\left(\beta_1\sum_i^n X_i\right) + \beta_1\left(\beta_1\sum_i^n X_i^2\right) \\
&= n\beta_0^2 + 2\beta_0\beta_1\sum_i^n X_i + \beta_1^2\sum_i^n X_i^2
\end{aligned} \tag{4.6.34}$$

The derivatives are therefore

$$\begin{aligned}
\frac{\partial}{\partial \beta_0}\left[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\right] &= 2n\beta_0 + 2\beta_1\sum_i^n X_i \\
\frac{\partial}{\partial \beta_1}\left[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\right] &= 2\beta_0\sum_i^n X_i + 2\beta_1\sum_i^n X_i^2
\end{aligned} \tag{4.6.35}$$

or in matrix notation

$$\frac{\partial}{\partial \boldsymbol{\beta}}\left[\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}\right] = 2 \begin{bmatrix} n\beta_0 + \beta_1 \sum_i^n X_i \\ \\ \beta_0 \sum_i^n X_i + \beta_1 \sum_i^n X_i^2 \end{bmatrix} \tag{4.6.36}$$

We recognize the matrix on the right as $\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}$; we saw it earlier in this derivation in the second line of Eqn.(4.6.34). Therefore

$$\frac{\partial}{\partial \boldsymbol{\beta}}\left[\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}\right] = 2\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} \tag{4.6.37}$$

Combining Eqns.(4.6.33) and (4.6.37), we write the derivative of Q with respect to $\boldsymbol{\beta}$ as

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T \boldsymbol{Y} + 2\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} \tag{4.6.38}$$

Setting this equal to zero yields the normal equations in matrix form for the simple linear regression model

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{\beta}} &= 0 \\ -2\boldsymbol{X}^T \boldsymbol{Y} + 2\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} &= 0 \\ \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} &= \boldsymbol{X}^T \boldsymbol{Y} \end{aligned} \tag{4.6.39}$$

A quick sanity check: does the expression $\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{Y}$ look like the normal equations we derived earlier in Eqns.(4.3.8) and (4.3.9)? Let's find out by calculating this expression explitly and comparing row by row of the resultant matrix:

$$\begin{aligned} \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} &= \boldsymbol{X}^T \boldsymbol{Y} \\ \begin{bmatrix} n & \sum_i^n X_i \\ \\ \sum_i^n X_i & \sum_i^n X_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \\ \beta_1 \end{bmatrix} &= \begin{bmatrix} \sum_i^n Y_i \\ \\ \sum_i^n X_i Y_i \end{bmatrix} \\ \begin{bmatrix} n\beta_0 + \beta_1 \sum_i^n X_i \\ \\ \beta_0 \sum_i^n X_i + \beta_1 \sum_i^n X_i^2 \end{bmatrix} &= \begin{bmatrix} \sum_i^n Y_i \\ \\ \sum_i^n X_i Y_i \end{bmatrix} \end{aligned} \tag{4.6.40}$$

Just comparing the rows, we see that the first row gives us the equivalent of which is Eqn.(4.3.8)

$$\sum_i^n Y_i = n\beta_0 + \beta_1 \sum_i^n X_i$$

and the second row gives us the equivalent of Eqn.(4.3.9)

$$\sum_i^n X_i Y_i = \beta_0 \sum_i^n X_i + \beta_1 \sum_i^n X_i^2$$

And finally, to complete this derivation, we return to Eqn.(4.6.39) and have to isolate the vector $\boldsymbol{\beta}$. For this, we multiply each side by the inverse of $\boldsymbol{X}^T\boldsymbol{X}$; that is

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$$

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Since $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\right) = \boldsymbol{I}$, we get the point estimators for the regression function

$$\boxed{\boldsymbol{\beta} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{\boldsymbol{-1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}} \tag{4.6.41}$$

4.7 Linear Regression - Matrix Examples

- Use this section to repeat previous examples and show how we derive the same results using the matrix formulation.
- Take 2 of the same problems from the previous sections and repeat them in matrix formulation

Let's take a look at the problem we looked at in Section XYZ using the tradition method and see if we can replicate the results using our matrix formulation.

$$Y = \begin{bmatrix} -15.3 \\ -11.07 \\ -10.72 \\ -10.29 \\ -9.66 \\ -6.05 \\ -7.12 \\ -5.19 \\ -2.6 \\ -1.38 \\ 0.7 \\ 1.64 \\ 3.06 \\ 4.56 \\ 6.2 \\ 8.3 \\ 9.84 \\ 12.11 \\ 9.92 \\ 15.75 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & -10 \\ 1 & -9 \\ 1 & -8 \\ 1 & -7 \\ 1 & -6 \\ 1 & -5 \\ 1 & -4 \\ 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \end{bmatrix} \tag{4.7.1}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots & 1 \\ -10 & -9 & -8 & -7 & -6 & -5 & -4 & \cdots & 9 \end{bmatrix} \tag{4.7.2}$$

$$X^T X = \begin{bmatrix} 20 & -10 \\ -10 & 670 \end{bmatrix} \tag{4.7.3}$$

$$\left(X^T X\right)^{-1} = \begin{bmatrix} \dfrac{67}{1330} & \dfrac{1}{1330} \\ \dfrac{1}{1330} & \dfrac{1}{665} \end{bmatrix} \tag{4.7.4}$$

$$b = \left(X^T X\right)^{-1} \; X^T \; Y$$

$$= \begin{bmatrix} \dfrac{67}{1330} & \dfrac{1}{1330} \\ \dfrac{1}{1330} & \dfrac{1}{665} \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ -10 & -9 & -8 & \cdots & 9 \end{bmatrix} \quad \begin{bmatrix} -15.3 \\ -11.07 \\ -10.72 \\ -10.29 \\ -9.66 \\ -6.05 \\ -7.12 \\ -5.19 \\ -2.6 \\ -1.38 \\ 0.7 \\ 1.64 \\ 3.06 \\ 4.56 \\ 6.2 \\ 8.3 \\ 9.84 \\ 12.11 \\ 9.92 \\ 15.75 \end{bmatrix} \qquad (4.7.5)$$

$$= \begin{bmatrix} 0.385579 \\ 1.50116 \end{bmatrix}$$

which means $b_0 = 0.385579$ and $b_1 = 1.50116$, which are identical to the results we saw in XYZ.

4.8 Multi-Variate Linear Regression

4.8.1 Linear Regression with Two Variables

Take notes on discussion on p.214

The main point here is that for a lot of problems, trying to predict the response variable based on a single predictor variable is inadequate. For example, housing price data.

To tackle this problem, we'll first build on our previous work and just move one step up in complexity: having two predictor variables instead of just one.

The regression model first defined in Eqn.() becomes for two predictor variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \tag{4.8.1}$$

As before, there is an alternative formulation of the model

$$Y_i = X_{i0}\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad (X_{i0} \equiv 1) \tag{4.8.2}$$

In this notation,

X_{ij} → The value of the j^{th} predictor variable in the i^{th} trial

Note that as before, we have

$$E(\epsilon_i) = 0$$

The regression function for the model in Eqn.(4.8.1)

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{4.8.3}$$

Recreate figure 6.1 in Kutner

When we had a single predictor variable, X_1, the response function was a line. With two variables, the response is a plane. Sometimes this is referred to as the **regression surface** or **response surface**.

According to figure 6.1, the distance between Y_i and $E(Y_i)$ is ϵ and represents the distance between Y_i and the regression surface.

How do we interpret the regression coefficients in the multiple regression function? β_1 is the change int he mean response $E(Y)$ per unit increase in X_1 when X_2 is held constant. Similarly, β_2 is the mean response $E(Y)$ per unit increase in X_2 when X_1 is held constant. The regression parameters β_1, β_2 are sometimes referred to as partial regression coefficients.

We can also interpret the meaning of β_1 and β_2 from calculus

$$\begin{aligned} \beta_1 &= \frac{\partial E(Y)}{\partial X_1} \\ \beta_2 &= \frac{\partial E(Y)}{\partial X_2} \end{aligned} \tag{4.8.4}$$

When the effect of X_1 on Y does not depend on the level of X_2, and the effect of X_2 on Y does not depend on the level of X_1, the two predictor variables are said to have additive effects or not to interact. The linear regression model in Eqn.(4.8.1) is designed for predictor variables whose effects on the mean response do not interact.

When we have more than two predictor variables, the response surface becomes difficult to visualize. However, in some cases the response surface of a plane can be used as an approximation to more complex surfaces, For example, within a specific range of X_j, you may have a situation where two out of the many other X_j are sufficient to model the response of Y.

Good reference to note the algebraic equivalents for the first-order regression model with two predictor variables. See p.239 in Kutner.

4.8.2 General Linear Model

We now define the general linear model as follows:

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i\ p-1} + \epsilon_i \tag{4.8.5}$$

where we define

- $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are the parameters of the regression function. We will see later why we picked $p-1$ (instead of p) as the number of predictor variables.

- $X_{i1}, X_{i2}, \ldots, X_{i,\ p-1}$ are the predictor variables and are known constants. The X_{i1} notation means the value of the j^{th} predictor variable in the i^{th} trial.

- The ϵ_i are the error terms. They are independent and normally distributed $N(0, \sigma^2)$.

- $X_{i0} \equiv 1$

We may write this model more compactly as

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \epsilon_i \tag{4.8.6}$$

As before, we know that $E(\epsilon_i) = 0$, the response function is given by

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} \tag{4.8.7}$$

Note that the $X_1 + X_2 + \ldots + X_{p-1}$ do not need to represent different predictor variables, as we shall see in Section (XYZ).

In setting up the expression for the solution for the β parameters, we will employ the power of matrix algebra. We'll be able to represent the results for the general linear regression model in a notation that mirrors that for simple regression in Section XYZ.

Y_i and ϵ_i are defined as before and are $n \times 1$ vectors. The parameter vector from Eqn.(**??**) is now expanded to be an $n \times 1$ vector

$$\underset{2\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \tag{4.8.8}$$

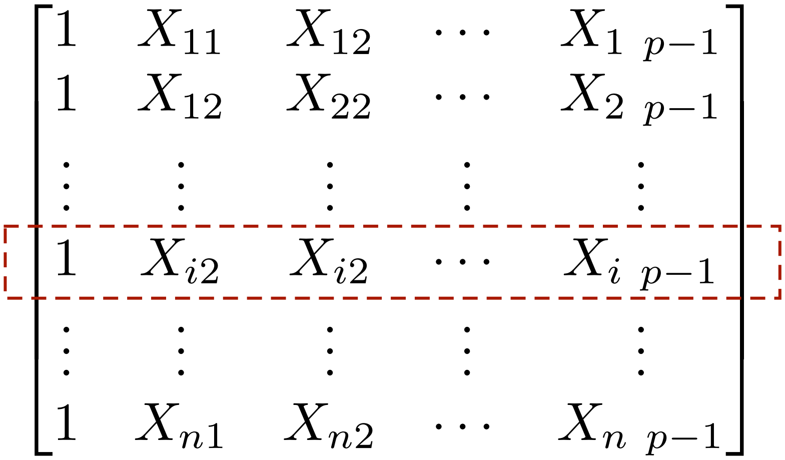


Figure 4.8.1:

The design matrix is also expanded to include all of the $p-1$ variables and n trials. So it is an $n \times p$ matrix given by

$$\underset{n\times p}{\boldsymbol{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1\ p-1} \\ 1 & X_{12} & X_{22} & \cdots & X_{2\ p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{i2} & X_{i2} & \cdots & X_{i\ p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n\ p-1} \end{bmatrix} \tag{4.8.9}$$

Figure 4.8.1 helps us interpret how we can visualize this general design matrix. Think of each row as all the data points from the i^{th} trial of an experiment or the training data. As you go across the row from left to right, you insert the value of each of the predictor variables. Teh first column of ones represents X_i0 which is defined to be 1.

The beautiful thing about matrix algebra is that the generalized model can also be expressed in the same way as the simple model from Eqn.(4.6.11).

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{4.8.10}$$

We can see that for the general case, the matrix dimensions also work out as expected

$$\text{Dimensions: } \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

$$n \times 1 = (n \times p)(p \times 1) + (n \times 1) \tag{4.8.11}$$

$$n \times 1 = n \times 1 + n \times 1$$

You may have wondered why we picked $p-1$ as the number of regression variables. We can see that it was done deliberately because of the column of 1s in the matrix which come from the predictor variable X_i0. By defining the number of predictor variables to be $p-1$, the final column dimension of the matrix is p.

The cost function of the regression model can we written as

$$Q = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} + \ldots + -\beta_{p-1} X_{i\ p-1}\right)^2 \tag{4.8.12}$$

The least squares normal equations are still given by Eqns.(XYZ) and (XYZ) and the least squares estimators can be derived using the same expression as before in Eqn.(XYZ)

$$\boldsymbol{\beta} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{4.8.13}$$

4.8.3 General Linear Model - Use Cases

p.218 in notes; All good examples.

From p.221 - What is clear from these examples is that the term “linear” refers to the model in Eqn.(4.8.5) being linear in the parameters β_i. Since we have seen more complex response surfaces, “linear” does not refer to the shape of the response hyperplane. To put it another way, a regression model is linear in the parameters when it can be written in the form

$$Y_i = C_{i0}\beta_0 + C_{i1}\beta_1 + C_{i2}\beta_2 + \ldots + C_{i\ p-1}\beta_{p-1} + \epsilon_i \tag{4.8.14}$$

where the C_{ik} are coefficients involving the predictor variables.

Just for reference, what are some examples of non-linear models?

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \epsilon_i \tag{4.8.15}$$

4.8.4 General Linear Model - Examples

Dwine Studios

4.8.5 What if the X^TX Matrix is Non-Invertible?

We may run into a situation where the matrix $\left(\boldsymbol{X}^T\boldsymbol{X}\right)$ has no inverse. The two most common causes for this are:

1. You have redundant features; i.e, the features are linearly dependent. For example, in a housing example, you have one parameter that gives the size of the house in square feet and another that gives it in square meters. It's the same information; one value can be derived from the other.

   Verify the theorem / rule in linear algebra that if two vectors in a matrix are linearly dependent, then it may not be invertible

2. We have too many features relative to the training examples. The way to take care of this is to delete some features or use regularization.

Chapter 5

Classification and Logistic Regression

In linear regression, we assume that the values of the response variable Y and the things we want to predict are quantitiative. However, sometimes we want to make a prediction about qualitative variables, also referred to as **categorical variables**. The process of predicting categorical variables is sometimes referred to as **classification**. The term is appropriate, since predicting a qualitative response for an observation involves assigning it to a class (i.e., "yes" or "no", "malignant" or "benign", "legitimate" or "fraudulent", etc.).

Three widely used classifiers: logistic regression, linear discriminant analysis, and K-nearest neighbors.

5.1 Logistic Regression - Examples

- Use data sets from Coursera class
- Good data set and examples from UCLA IDRE site:
  IDREatUCLA

Chapter 6

Intermediate Value Theorem

Theorem 6.0.1: Intermediate Value Theorem
} Suppose $f : [a, b] \to \mathbb{R}$ be continuous function over a closed interval $I = [a, b]$. In addition, suppose $f(a) \neq f(b)$. If $L \in \mathbb{R}$ satisfies $f(a) < L < f(b)$ or $f(a) > L > f(b)$, then there exists a point $c \in (a, b)$ where $f(c) = L$.

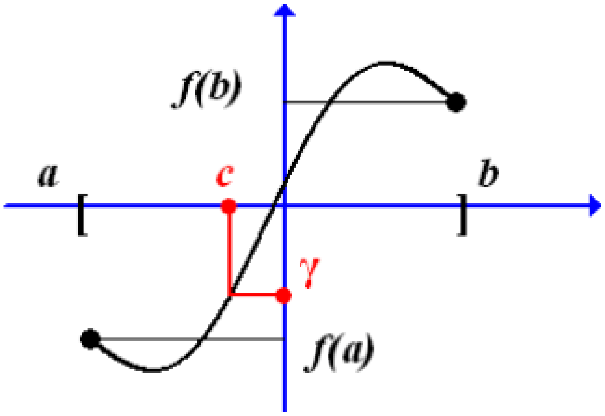


Figure 6.0.1:

Proof taken from Dr. M. Coleman - University of Manchester
Before we discuss the proof, let's see what this means intuitively. In Figure XYZ, we have a continuous function on the closed interval $[a, b]$. We then

pick any value, L, that lies between $f(a)$ and $f(b)$. From the figure, it seems plausible that the line $y = L$ would intersect the function at least once. That point of intersection happens at some point $x = c$ and that $a < c < b$.

Note that the Intermediate Value Theorem only tells you that the function will have a value of M somewhere between $[a, b]$; it does not tell you where. It simply states that it exists.

■ Redraw this figure and replace γ with L.

Chapter 7

Intermediate Value Theorem - Follow-up

Interestingly, we can use the Intermediate Value Theorem to prove the existence of roots of equations. Let's take an example:

Example: Show that $p(x) = 2x^3 - 5x^2 - 10x + 5$ has a root somewhere in the interval [-1,2].

What we're really asking here: is there a number, c, such that $p(c) = 0$ somewhere between $-1 < c < 2$?

To use the Intermediate Value Theorem here, we first note that $p(x)$ is a polynomial and so it is continuous over this interval. Note also that $p(-1) = 8$ and $p(2) = -19$, so we know the function must pass through 0 at some point in this interval between $p(-1)$ and $p(8)$. Therefore, there must exist a $x = c$ such that $p(c) = 0$. So the function p has at least one root in this interval.

`[Make this example more formal - p.189 in Anton]`

It's worth pointing out that the Intermediate Value Theorem can only tell you if a c exists. However, if M does not lie between $f(a)$ and $f(b)$, the Intermediate Value Theorem doesn't tell you that c cannot exist – you simply cannot make the determination that there is a $p(c) = M$ using this theorem.

■ Find a good figure that illustrates this.

■ Step-by-step slides - use for the main proof. Use some of his figures as well. I like how he approaches the proof from a basic case, and then shows how you can generalize it for $\gamma \neq 0$ and $f(a) > L > f(b)$. Make the point early on in the discussion that the proof of the basic case can easily be extended to the more general case.

■ Some examples in the UC Davis course section on Intermediate Value Theorem.

■ Use figure in Wikipedia for now. Recreate it later. Can also include figures from Anton.