Program Evaluation: Methods and Design

Joshua Manning, Jesse Lecy

01 August, 2018

Contents

Welcome

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

Part I

REGRESSION

Chapter 1

R Markdown

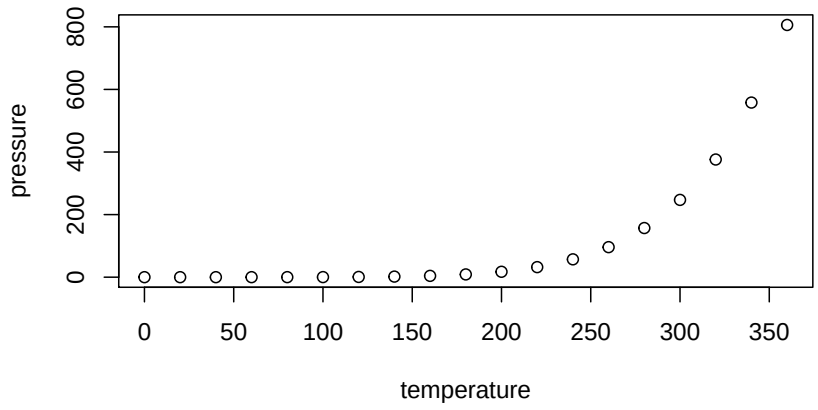
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
#>      speed           dist
#>  Min.   : 4.0   Min.   :  2.00
#>  1st Qu.:12.0   1st Qu.: 26.00
#>  Median :15.0   Median : 36.00
#>  Mean   :15.4   Mean   : 42.98
#>  3rd Qu.:19.0   3rd Qu.: 56.00
#>  Max.   :25.0   Max.   :120.00
```

1.1 Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Chapter 2

Program Impact

2.0.1 Why do we use statistical models for program evaluation?

In unit 1 we introduced the general purpose of statistics and quantitative methods

In unit 2 we will provide a more formal definition of how statistics is used and what it can tell us about a program

2.0.2 Reminder from Unit 1: Advantages of Statistical Evaluation

- Statistical models can be used with to analyze data that will show the quality, program impact, and where the impact occurs.
- Quantitative and statistical analysis can attempt to give an unbiased evaluation. It can provide us with outcomes associated with probabilities, level of confidence, and the size of the effect. It can also provide information about the specific relationship between the program, the effect, and the impact.
- Variables can be used to represent the program and its effect.

2.1 The Regression Equation

In unit 1 we saw that the regression line is represented by a linear equation (Eq. 1.1). We also saw regression coefficients, often represented by the upper case Greek letter Beta, show the slope of the regression line. In this case the slope is Beta with a subscript of 1. Here we only have up to B_1. This means that there is only one independent variable. Over next few weeks we will begin to see that we can have more than on independent variable represented by $B_1, B_2, B_3, ...B_i$.

(Eq. 2.1) Regression Equation: $Y = B_0 + B_1X_1$

The Beta with a subscript 0, (B_0), shows the intercept or where the regression line crosses the horizontal axis, Eq 2.1. How can we interpret this? If the value of x_1 is 0 for a data point then B_1 is multiplied by 0, which would make $B_1X_1 = B_10 = 0$ Because it is equal to zero all that remains is Eq 2.2.

(Eq. 2.1) Regression Equation: $Y = B_0 + B_1X_1$

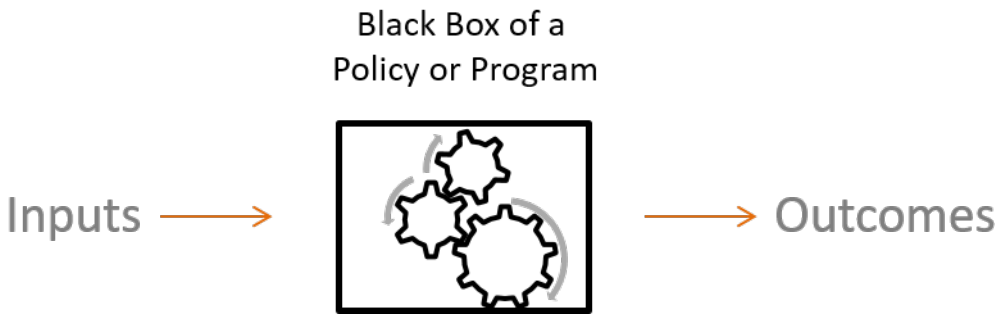
(Eq. 2.2) Intercept of Regression Equation: $Y = B_0$

2.1.1 Hypothetical Example of Interpreting the Regression Equation

A manager may want to increase productivity of the employees by implementing a new or altered program in the organization. There may be many ways that this could happen, such as better technologies, implementing work teams, or creating a better work environment. The manager proposes a simple solution: increase pay. Therefore, the manager believes that employees will be more productive with more pay.

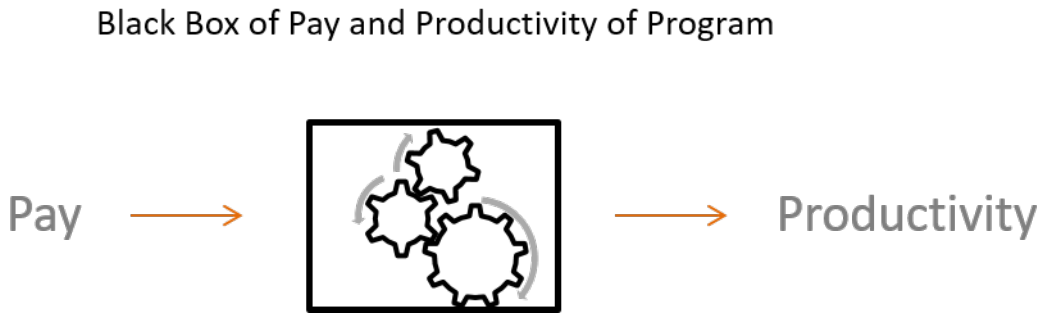
The manager implements the pay raise and the employees are very satisfied with the increased income. Everything seems to be going well, but how does the manager know if the increased pay was associated with increased productivity. In other words: did the program work? The manager decided to simply compare raw productivity data would not lead to enough confidence that it was successful. The solution is using statistical methods for evaluation.

Recall that program evaluation can be interpreted as a blackbox model (Fig. 2.1) with inputs and outcomes.



(Fig. 2.1)

In this case there are inputs of pay raise and outcomes of productivity levels.



(Fig. 2.2)

Because the manager has inputs (pay raise) and outputs (productivity) the manager can use these to run a regression analysis.

(Eq. 2.3) Regression Eq: $Productivity = B_0 + B_1 PayRaise$

The manager ran the regression analyis and the results show that $B_0 = 50$ units of productivity and that $B_1 = 5$.

(Eq. 2.4) Regression Eq: $Productivity = 50 + 5PayRaise$

How should the manager interpent these results? Because B_0 is 50 it means that this is the expected productivity if there was no pay raise. Because B_1 is 5 it means that for each unit or dollar of a pay raise you would expect productivity to increase by 5 units.

(Eq. 2.4) Regression Eq: $Productivity = 50 + 5PayRaise$

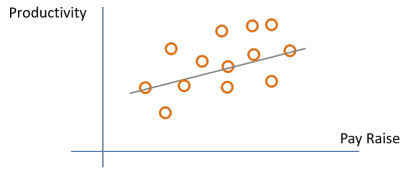


Figure 2.1: The variance of pay productivity

2.2 Confidence Intervals

How can we be confident of our estimate of program impact?

From the example we saw that for each dollar we expect to see an increase of 5 units of productivity. Does this mean that for every employee each dollar of pay raise translates into 5 unit of productivity? That would mean we would be perfectly confident that the regression equation describes each employee perfectly. Therefore the regression equation would resemble Fig. 2.3.

However, in the real world people are all different and there is variance in how the pay raise will impact productivity for each individule employee. Fig. 2.4 shows that some employees have smaller increase in prodcutivity for a dollar of pay rais and some a larger increase.

In statistics we can use confidence intervals. Confidence intervals tell us if we repeat an experiment many times we would expect a certain percentage of the experiments to contain the true mean in the confidence interval. Often people use 95% confidence intervals. Therefore, if the experiment is repeated 100 times we would expect 95 of those experiments to contain the true mean within the confidence interval.

This also applies to regression. At the 95% level if the experiment is repeated 100 times we would expect 95 of those experiments to contain the true B or regresion slop within the confidence interval. Later we will see how confidence intervals are calculated, but in order to calculate the confidence interval you must first calculate the standard error or s.

2.3 Confidence Intervals For Hypothesis Testing

Fig. 2.5 shows a visualization of using confidence to test the hypothesis that B_1 for pay raise is significantly greater than the true population B, (A), or not significantly greater, (B), or Significantly less, (C). For A and C the confidence interval does not contain the null or true population B and therefore A shows B_1 significantly greater than the true population B and C significanly less than the true population B. However, because in B the confidence interval includes the null of the true population B there is no significant difference.

2.4 The Standard Error

The standard deviation and variance are closely related. The standard deviation essentially shows the average distance between each data point and the mean. The standard error is simply a standard deviation used for the sample mean or the B estimated from the sample. First we will review the variance and standard deviation.

2.5 The Variance and Standard Deviation Of The Mean

Eq.2.6

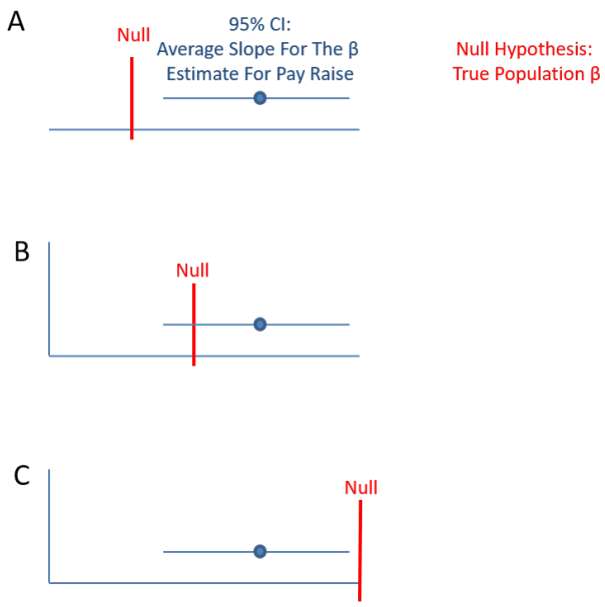


Figure 2.2: Confidence interval of pay raises

Eq.2.7

Eq. 2.6 shows the equation for calculating the variance. σ_x^2 is the symbol for the variance. $\bar{x}$ is the sample mean and x_i is each individual data point. Therefore, x_i - $\bar{x}$ is the size of the deviation of each data point from the mean. The subscript i stands for index, therefore x_1 is the first data point. Σ represents the summation. Therefore, $\Sigma(x_i - \bar{x})$ is the summation of the deviation of each data point from the mean. n is the sample size or number of data points.

When thinking about the variance formula, why not just use the average of the summation of the deviations of each data point? Because the mean will be the center of the data, the sum of deviations will always equal 0. Therefore, if the deviations are squared the deviations all become positive.

Because the variance uses squared deviation it is not in the same units a the data points or the mean. The solution is to take the square root of the variance, Eq. 2.7, which is called the standard deviation or σ. This puts the variance into the same units as the data points and the mean. In addition, it is easier to interpret.

2.5.1 Hand Calculation Of The Variance

Below shows how to calculate the variance by hand using Eq. 2.6. The data set shows the grades of 5 students of a quiz with a maximum of 10 points. The entire data set is represented by X.

$X = [10, 8, 9, 6, 10]$

First calculate the deviation from the mean for each data point.

$10 - 8.6 = 1.4$

$8 - 8.6 = -0.6$

$9 - 8.6 = .4$

$6 - 8.6 = -2.6$

$10 - 8.6 = 1.4$

The second step is to square each deviation from the mean.

$$SE_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

Figure 2.3: Standard Error of the Mean

$1.4^2 = 1.96$

$-.6^2 = .36$

$.4^2 = .16$

$-2.6^2 = 6.76$

$1.4^2 = 1.96$

The third step is tha sum or add the squared deviations from the mean. This is called the sum of the squared deviations.

$\Sigma(x_i - \bar{x})^2) = 1.96 + .36 + .16 + 6.76 + 1.96 = 11.2$

The fourth step is to calculate the sample size minus 1.

$5 - 1 = 4$

The final step is to divide the sum of squared deviations by the sample size minus 1.

$\sigma^2 = \Sigma(x_i - \bar{x})^2)/(n - 1)$

$\sigma^2 = 11.2/4 = 2.8$

To calculate the standard deviation we simply take the square root of the variance or 2.8.

$\sigma = \sqrt{\sigma^2} = \sqrt{2.8} = 1.673$

2.6 The Standard Error Of The Sample Mean

Recall that the standard error is used when we do not know the true population standard deviation and only obtain a standard deviation of a sample. One very interesting fact is that the distribution of the means of each sample converges to a normal distribution as the number of samples become very large. This is refered to as a sampling distribution. The concept that the sampling distributions of the means approaches a normal distribution is called the Central Limit Theorem. Because of the Central Limit Theorem the mean and standard error are our best estimates of the population statistics even though we only have one sample. Therefore, we will use the standard error instead of the standard deviation when we conduct inferential statistics from a sample.

2.6.1 Calculating The Standard Error Of The Mean

Eq. 2.8 shows the standard error denoted by $SE_{\bar{x}}$. This is the formual that we will used for statistical analyses, such as confidence intervals.

Eq. 2.8

2.6.2 Hand Calculating The Mean And Standard Error Of Samples.

Recall that we calculated the variance and standard deviation previously with 5 data points.If those 5 data points are the population, we can take substes of those 5 data points as samples. Below are all the possible samples of 3 data points.

Sample Data Sets:

$X_1 = [10, 8, 9]$ $X_2 = [10, 9, 6]$

$X_3 = [10, 6, 10]$ $X_4 = [8, 9, 6]$

$X_5 = [8, 6, 10]$ $X_6 = [8, 10, 10]$

$X_7 = [9, 6, 10]$ $X_8 = [9, 10, 10]$

$X_9 = [6, 10, 10]$ $X_{10} = [6, 10, 8]$

2.6.3 Calculating The Standard Error

The first step is to calculate the standard deviations for each data set. We have alreatdy calculated the standard deviation by had previously. Therefore, the standard deviations are provided below.

$X_1 = 1$ $X_2 = 2.08$

$X_3 = 2.31$ $X_4 = 1.53$

$X_5 = 2$ $X_6 = 1.15$

$X_7 = 2.08$ $X_8 = .578$

$X_9 = 2.31$ $X_{10} = 2$

The final step is to divide the standard deviations by the sqare root of n or the sample size. For each data set the sample size is 3.

$1/\sqrt{(3)} = .58$ $2.08/\sqrt{(3)}$

$2.31/\sqrt{(3)} = 1.33$ $1.53/\sqrt{(3)} = .88$

$2/\sqrt{(3)} = 1.55$ $1.15/\sqrt{(3)} = .67$

$2.08/\sqrt{(3)} = 1.20$ $.578/\sqrt{(3)} = .33$

$2.31/\sqrt{(3)} = 1.33$ $2/\sqrt{(3)} = 1.55$

2.6.4 Recap Of The Standard Error

Conceptually both the standard deviation and the standard error are identical (both measure the 'average' error we expect when we make a guess about the population statistic using a sample statistic, and the formula for the confidence interval will be the same no matter which we use).

The formula for the standard error for each sample statistic has been derived by mathematicians, and provides the theoretical foundations for inferential statistic. We don't have to take multiple samples to calculate the average error directly. Statistical theory has already established the relationship between descriptive statistics of a sample and the standard error.

2.7 Summary: What we covered in this unit

- We introduced the regression model and how to interpret the slope of the program and the impact on the outcome.
- We introduced how to know if there is confidence in a program and the concept of confidence intervals.
- We introduced the standard error and when we use it rather than the standard diviation.
- We calculated step-by-step the variance, the standard deviation, and the standard error.

2.7.1 Looking Ahead

- We will cover the interpretation of the standard error.
- We will cover the details of calculating confidence intervals.
- We will cover the regression model in more detail.

Chapter 3

Confidence Intervals

What is the standard error and how is it used?

In unit 2 we showed that the standard error is a type of standard deviation. This describes how variable the data is and how far the data is from the mean.

The important difference is that the standard error is used when there is an inferential statistical analysis using a sample and the true population statistics are not known.

Confidence interval tell us how confident we can be that the there is a difference between the null hypothesis.

3.1 How The Size Of The Standard Error Represents Confidence

Recall that the standard error describes the variability of the data. The more variable the data the less confidence there is. For example, assume that every week you get paid very different amounts. Some weeks you get paid a large amount, some weeks a reasonable and moderate amount, but other weeks you get paid very little. This type of pay is very variable. Would you have much confidence in the amount of money you have to spend every week? Most likely you would have little confidence. This is because the variability is so large. The standard error is similar. Large standard error lead to less confidence. However, smaller standard errors allow for high confidence. Next we will introduce how to calculate standard errors followed by the details of confidence intervals.

3.1.1 Factors That Influence The Size Of The Standard Error

The main factor that affects the size of the standard error is the sample size. Because the denominator of the standard error is the square root of the sample size (Eq.3.1) as the sample size gets larger the standard error becomes smaller.

3.1.2 The Standard Error Of The Mean Equation

Eq. 3.1

3.1.3 Factors That Influence The Size Of The Standard Error In Regression

We will discuss more details about the regression model in the next few units. Recall that because every data point does not usually fall in a perfect line there is usually space between each data point and the regression

line. The distance between the data point and the regression line is called the error term or residual (Fig.3.1). One way to decrease the size of the standard error is to add more variables.

3.1.4 The Residual In Regression

Fig. 3.1

3.1.5 Factors That Influence The Size Of The Standard Error In Regression

Recall that the regression slope, B, describes the impact on the dependent variable or output. It describes how many units the output will increase or decrease for every one unit of change of the independent variable or input. There is also a standard error of B. The calculation is similar. Eq. 3.1-3.4 show the formulas to calculate the variance and standard deviation of the slope. In Eq. 3.3 SSE is the sum of squared error term. It is analogous to the sum of the deviation of the sum of the deviations from the each data point and the regression line. Eq. 3.5 Shows the formula for the standard error of the regression slope. Just as with the standard error of the mean you must first calculate the variance.

3.1.6 The Variance And Standard Deviation Of The Regression Slope Equation

Eq. 3.2 $SSE = (\hat{y}_i - \bar{y})^2$

Eq. 3.3

Eq. 3.4

3.1.7 The Standard Error Of The Regression Slope Equation

Eq. 3.5

Two other ways to decrease the size of the standard error are (1) increase the sample size and (2) to increase the variance of x. Increasing the sample size will decrease the variance of the regression slope by increasing the denominator of the variance (Eq.3.5). Similarly, increasing the variance of x will increase the denominator of the standard error of the slope leading to a smaller SE_b.

3.2 Standard Errors And Confidence Intervals

Recall that Confidence intervals tell us if we repeat an experiment many times we would expect a certain percentage of the experiments to contain the true mean in the confidence interval. Often people use 95% confidence intervals.

If the experiment is repeated 100 times we would expect 95 of those experiments to contain the true mean within the confidence interval.

In order to calculate confidence intervals we must know how variable the data are. This is estimated by the standard error. We can now use the standard error to calculate the confidence interval

3.2.1 Calculating The Confidence Interval Of The Mean

Recall that the standard error is a type of standard deviation. However, it is used when we are calculating statistics from a sample and do not know the true population statistics. The formulas and equations are very related and similar. You likely have seen in earlier statistics courses how to calculate the t-statistic.

The t-statistic is used to determine the probability that you would get the results that you obtain from your sample if the results were due to random chance. In other words, it is very similar to confidence intervals. If you are interested in a probability or p-value of less than 5% or .05 to be significant then a t-statistic that shows the p-value is less than .05 you believe that the results were not due to random chance, rather they were due because there is a true difference between the mean of your sample and the population mean. For example, a t-statistic that shows a p-value of less than .05 in a study where you wanted to know if a drug worked better than a placebo you would infer that the results of the drug were not due to random chance.

The t-statistic is integrally involved in the calculation of confidence intervals. One reason is that, as with t-tests, we have a probability value or in this case a confidence level. In calculating the t-statistic for the confidence interval we must know the degrees of freedom. A t-statistic for the mean has $n-1$ degrees of freedom or the sample size minus 1. The t-statistic for the regression slope, B, the degrees of freedom are $n-k-1$ or the sample size minus the number of regression paramaters, B's, minus 1.

Chapter 4

The Regression Residual

Viewing The Regression Line As A Conditional Average

In statistics, conditional basically means if you know the value of one variable you either know or can estimate the value of another variable. In other words, if you tell me the level of x, I can tell you the average y for someone with that level of x. In terms of the regression x is the value of the input or independent variable of a program and y is the outcome or dependent variable of the program. Therefore, y is dependent or conditional on x.

4.1 Regression Residuals And Errors

Recall the example of pay raise and productivity. In the real wold people vary. The regression line gives us the best linear fit or prediction of what productivity or y will be if we know the pay raise or x. Therefore y is conditional on x. However, the true values of productivity conditional on each person's productivity varies by person and therefore varies around the regression line instead of being perfectly on it. (Fig. 4.1)

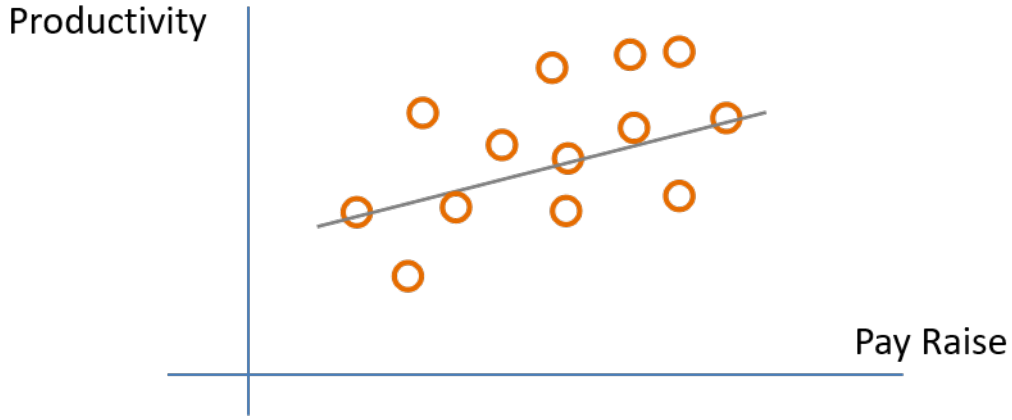


Fig. 4.1

As in Fig. 4.1, the data will cluster around the regression line, but not on it. The residual is calculated from the actual value of y minus the predicted value of y or $\hat{y}$, which is a point on the regression line. (Fig. 4.2)

4.2 The Residual In Regression

Fig. 4.2

4.2.1 Interpreting And Using The Residual In Regression

The regression line tells us our best linear estimate of y conditional on x. This can also be said as y given x, which is represented mathematically as $y|x$. However, because the actual data usually cluster around the regression line each data point likely has a regression residual.

The regression residual can be either positive or negative. In other words, some data points will be above the regression line leading to a positive residual and some data points will be below the regression line leading to a negative residual.

A good example of residuals in regression is "The Math of Cities". This example is based on data from the Census. It basically shows that the size of the city is related to productivity. For this study they used the number of patents as a measure of productivity. This is reasonable because it is a measure of ingenuity and the creation of new ideas and products. To listen to the researchers go to https://www.wnycstudios.org/story/96043-its-alive/ and listen to the time 12:30-17:30.

4.2.2 How The Residual Describes Over-performing and Under-performing

Let us consider the example of "The Math of Cities". What does it mean to be over-performing or under-performing relative to city size? If a city over-performs relative to the regression line the number of patents or productivity will lead to a positive residual. If you under-perform relative to the regression line the number of patents or productivity will lead to a negative residual. Fig. 4.3 shows the regression line and how some cities are above the line and over performing and some below and under-performing.

Fig. 4.3

As we have noted residuals can be positive or negative, and reflect over-performing and under-performing in "The Math and Cities" example. Fig. 4.4 shows the residuals represented by the red line connecting each data point and the regression line. These red vertical lines represent the distance between the data points and the regression line. The positive residuals are the the red lines above the regression line and the negative residuals are the red lines below the regression line. e in the regression equation represents the error or the regression residual.

Fig.4.4

4.3 How Do We Know The Best Fitting Regression Line For Our Data

Finding the best fitting regression line uses mathematics that can be calculated by hand. Luckily computers identify the regression line based upon the criteria of "line of best fit" for the data. In most cases, this means that we are finding the line that minimizes the distance between the line and all data points, i.e. minimizing the error in the model. These errors are the red lines representing the regression residuals (Fig. 4.4).

4.3.1 The Calculation Of The Regression Line

Calculating the regression line and slope coefficients must find the best fitting line. Therefore, it would be a better fit if there is minimum error or residuals for the entire set of data points. Just as we did with the variance, we could not use the absolute error, but rather we squared the deviations from the mean. To calculate the regression line we will also need to square the error or residuals.

Eq. 4.1 shows the regression line that we have studied previously. This equation has a subscript i for x and y. This subscript represents each data point. Therefore, we have a specific value for each y_i or the dependent variable and for each x_i, which is taken directly from data that was collected.

Eq. 4.1

$y_i = B_0 + B_1 x_i + e_i$

To calculate the best fitting regression line we must minimize the error for the entire data set. Therefore, we will minimize the sum of squared errors or residuals. With a little algebra Eq. 4.1 can be rearranged to give you the equation in terms of the error term/residual with it on the left side of Eq. 4.2. This is essentially solving for the error term e. The next step is to sum the squared errors/residuals of each data point. This is shown in Eq. 4.2.

Eq. 4.2

$e_i = y_i - B_0 - B_1 x_i$

Eq. 4.3

$\Sigma(e_i = y_i - B_0 - B_1 x_i)^2$

The next step is to minimize Eq. 4.3 or the sum of squares of the residuals. This requires calculus and is beyond the scope of this course. Once Eq. 4.3 is minimized you can calculate the intercept and slope coefficients of the regression line B_0 and B_1. This is called the ordinary least square regression or OLS. Statistical packages, such as R and others, can give you the OLS regression line.

4.3.2 Implications Of Squaring Residuals

Using sum of squares in both variance and calculating regression lines has important consequences in regard to misinterpreting the meaning of the sum of squares. If we think back to the variance and standard deviation of the mean, because the variance is in terms of squared deviation each unit of increase in the data leads to a squared difference. This leads to an increasingly larger increase in variance for each unit a data point is from the mean.

Her are examples of squared deviations leading to misinterpreting the absolute deviations. These are the results of squaring each deviation from the mean. If a data point deviates 1 point from the mean then 1^2 is 1. If a data point is 2 units from the mean then 2^2 is 4. If a data point is 3 units from the mean then 3^2 is 9. This leads to the squared deviations from the mean increasing much more quickly than the absolute deviations. In this example we can see this by noticing 2^2 being 4 times a deviation of 1 and 3^2 being 9 times a deviation of 1. The solution was to take the square root to put the squared units into the same units as the original data. This is the same with the squared residuals. This is important when considering outliers. Outliers using squared residuals can have an extreme effect on the analysis.

The Regression Line Passes Through Both The Mean Of Y And Mean Of X

Once the sum of the squared residuals is minimized to calculate the regression line the results give the equation for the intercept B_0 (Eq. 4.4) and the equation for the regression slope B_1 (Eq. 4.5). Remember that the "hat" over the $B's$ represent the fitted or estimates and the "bar" over x and y represent the mean. This will imply that the regression line passes through the mean of x and the mean of y.

4.4 The Regression Coefficient Equations Estimates

Eq. 4.4

$\hat{B_0} = \bar{y} - \hat{B_1}\bar{x}$

Eq. 4.5

$\hat{B_1} = (\Sigma(x_i - \bar{x})(y_i - \bar{y}))/(\Sigma(x_i - \bar{x})^2$

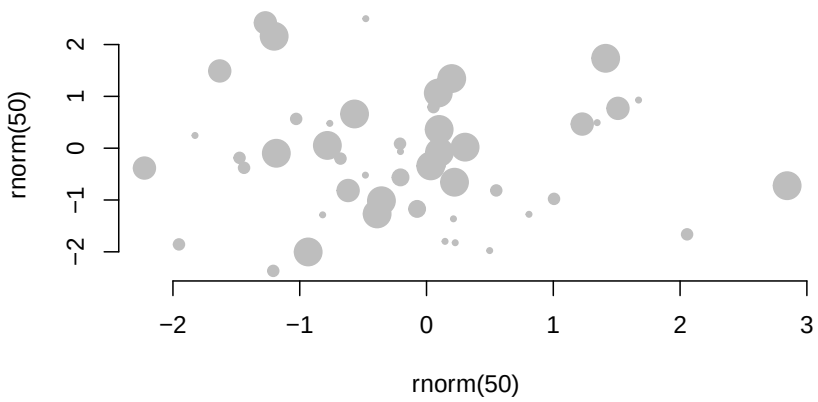


Figure 4.1: Here is a Figure Caption

4.5 Looking Forward

1. We will explore the regression slope further.
2. We will discuss how there is variability and a standard error of the regression slope as there is with the mean.
3. We will also discuss how variables vary together, which can be measured by the covariance.

4.6 Figure Example

```
plot( rnorm(50), rnorm(50), cex=0.5*sample(1:5,25,T), bty="n", pch=19, col="gray" )
```

In **??** we see examples of plotting in R, and another example in **??**.

```
plot( rnorm(50), rnorm(50), cex=0.5*sample(1:5,25,T), bty="n", pch=19, col="firebrick" )
```

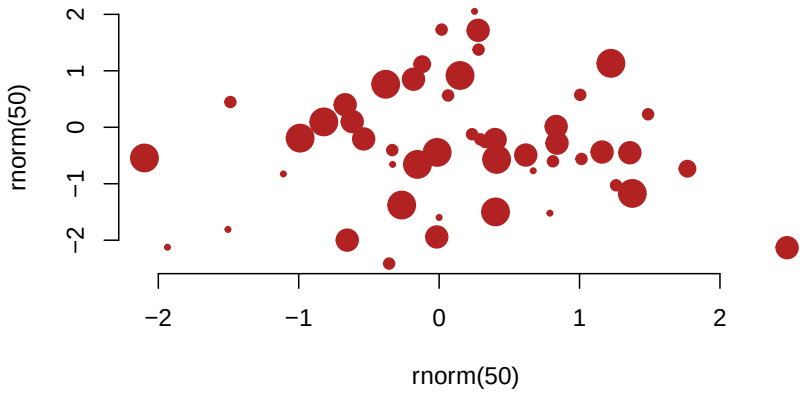


Figure 4.2: plotting example

Chapter 5

The Regression Slope

5.1 Variance Revisited

Recall from Unit 2 we discussed the variability and variance of the data. We therefore are intersted in how the data is spread out from the mean of the data. To assess how the data is spread out from the mean we want a measure for how spread out or how the data deviates from the mean $\bar{x}$ This is accomplished by using the following calculation: $\Sigma(x_i - \bar{x})$, which is the summation of the deviation of each data point from the mean. The complete formula for calculating the variance is Eq. 5.1

Eq. 5.1 $\sigma^2 = \Sigma(x_i - \bar{x})^2)/(n-1)$

We also saw that if you add all of the deviations from the mean, the distance from the mean of each data point, it will always be 0. To solve this problem of the sum of the deviations being 0 we squared each deviation, which you can see in Eq. 5.1.

5.2 The Standard Deviation Revisited

Recall that we calculated the standard deviation by taking the square root of the variance. This is necessary because the squared deviations from the mean are in squared units. However, the original data and the mean are not in squared units. Taking the square root of the variance it puts the standard deviation in the original units.

5.2.1 Squared Residuals

Related to the variace, we minimized the sum of the squared residuals, or the distance of each data point from the regression line. This will give us the best estimation for the regression line.

5.3 Covariance

So far we have discussed how the data varies around the mean and varies around the regression line. Often we are interested in how two variables vary together. In otherwords, as one variable increase does the other variable increase or decrease. How two variables vary with each other is called the covariance. This is related to regression slopes. If the regression slope is positive then as the independent variable increases the dependent variable increases. If the regression slope is negative then as the independent variable increases the dependent variable decreases.

5.3.1 Calculating The Covariance

The equation for the covariance will look similar to that of the variance. With the variance we calculated the sum of the squared deviations from the mean. However, with covariance we have two variables. For the covariance we now calculate the sum of the deviations of the two variables multiplied together, variable 1 or x and variable 2 or y, Eq. 5.2. While similar to the next step in the variance, for the covariance we now add the multiplied deviations of x and y, Eq. 5.3.

Eq. 5.2

$(x_i - \bar{x})(y_i - \bar{y})$

Eq. 5.3

$\Sigma(x_i - \bar{x})(y_i - \bar{y})$

Finally we divide by $n - 1$, or the sample size minus 1. Eq. 5.4. $cov(x, y)$ stands for the covariance. Let us look at Eq. 5.4 more closely. A positive number multiplied by a positive number will result in a positive number. Therefore if $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are either both positive or both negative then the numerator and thus the covariance will be positive. Alternatively, if $(x_i - \bar{x})$ is positive and $(y_i - \bar{y})$ is negative than the numerator and thus the covariance will be negative. The third possibility is if $(x_i - \bar{x})$ is negative and $(y_i - \bar{y})$ is positive than the numerator and thus the covariance will be also be negative.

Eq. 5.4

$cov(x, y) = \Sigma(x_i - \bar{x})(y_i - \bar{y})/n - 1$

We can use the covariance to calculate the correlation between two variables. Corrlation measures the strength of the linear relationship and the direction between two variables. Suppose we were interested in how class size, the independent variable, is related to test scores, the dependent variable. If there is a positive covariance and correlation between the the two variables will increase from bottom left to upper right in a plot, Fig. 5.1.

5.3.2 Positive Correlation

Fig. 5.1

5.3.3 Negative Corrleation

If there is a negative covariance and correlation between the the two variables will decrease from top left to bottom right in a plot,

Fig. 5.2

5.3.4 Low Correlation

If there is a low to no covariance and correlation (the variables are unrelated) between the the two variables will appear randomly scatterd in the plot with no upward or downward to the right, Fig. 5.3. There are two other possibilitties. If the plot is either arranged in a vertical or horizontal pattern then there is also no correlation. This is because the vertical plot means as y increases then x does not change and the horizontal plot means as $x$increases y does not change.

Correlations are always between -1 and $+1$. If the the correlation is closer to 1, then the correlation is positive and has a strong linear relationship. If the correlation is closer the correlation is to -1, then the correlation is negative and has a strong linear relationship. Finally, if the correlation is closer to 0, then the

correlation is low and has a weak linear relationship. A correlation of 0 would mean that the variables are completely unrelated.

5.4 The Slope

Recall the equations for the covariance and variance: Eq. 5.4 and Eq. 5.1. We can divide the covariance by the variance as seen in Eq. 5.5. With a little algebra $(n-1)$ in both the covariance and variance will cancel out each other. This will leave us with $[(\Sigma(x_i - \bar{x})(y_i) - \bar{y})]/[\Sigma(x_i - \bar{x})^2)]$. By cancelling each $\Sigma(x_i - \bar{x})$ in the numerator and denominator we are left with the components in Eq. 5.6. This is an intuitive formula for the slope because, if you recall from algebra the slope of a line is the change in y divided by the change in x.

Eq. 5.5 $cov(x, y)/var(x) = [\Sigma(x_i - \bar{x})(y_i - \bar{y})/n - 1]/[\Sigma(x_i - \bar{x})^2)/(n-1)]$

Eq. 5.6 $(y_i - \bar{y})/(x_i - \bar{x})$

5.5 Looking Ahead

- We will discuss how the variance can be explained with partitioning different variaces
- We will discuss R^2 or the coefficient of determination

Chapter 6

Explaining Variance

The variance can be split up into components of several parts. Intuitively, if you consider a regression model that has two independent variables, x_1 and x_2. This would imply that the dependent variable, y, would have variances from two components that can help explain how the independent variables are related to it.

Two components of the variance are the regression sum of squares, RSS, and the error sum of squares, ESS. Eq's. 6.1 and 6.2 show the equations that are used to calculate RSS and ESS. Finally there are the total sum of squares, TSS, which is simply adding RSS and ESS together.

Eq. 6.1

$RSS = \Sigma(\hat{y_i} - \bar{y})^2$

Eq. 6.2

$ESS = \Sigma(y_i - \hat{y_1})^2$

Eq. 6.3

$TSS = \Sigma(y_i - \bar{y})^2 = RSS + ESS$

6.1 R-Squared

One important measurement in statistics and explaining the variance is R-squared or R^2. Basically R^2 is the percentage of the variance explained by independent variables or the regression model. This can be seen in Eq. 6.4 by seeing it is simply the ratio of the explained variance or simplified to the explained sum of squares divided by the total sum of squares. This will always give us a value between 0 and 1, which also reflects the ratio or percentage. Because we are using how much the regression explains the explained sum of squares is the same as the regression sum of squares or RSS.

Eq. 6.4

$R^2 = RSS/TSS$

Because it is a percentage, sometimes it is easier to use Eq. 6.5. This comes directly from the factor that the ratio and percentage measured by Eq. 6.4 must add to 1.

Eq. 6.5

$R^2 = ESS/TSS$

6.2 Hand Calculation Of The Regression

Below shows the data you previously used to calculate the variance in Unit 2. The data set shows the grades of 5 students of a quiz with a maximum of 10 points. The entire data set is represented by Y. This time we are representing that data with Y because we will treat this as the dependent variable. The data set X will now be used for the hours spent studying for the quiz.

$Y = [10, 8, 9, 6, 10]$

$X = [5, 2, 3, 3, 6]$

As an exercise please use the data above to calculate the slope, intercept, predicted Y, residual, and sum of squared errors for the regression. You can use the equations we just showed and have seen earlier. In addition show the following:

1. Show that the sum of squared errors becomes the numerator in the standard error of the slope
2. Calculate the regression sum of squares (difference between observed and predicted value)
3. Calculate the total sum of squared errors $\Sigma(y_2 - \bar{y})^2$ and show that this is the raw variance instead of the average variance
4. Calculate the R-square as RSS/TSS and 1 - ESS/TSS.
5. Emphasize that RSS + ESS = TSS, or explained variance plus unexplained variance is equal to total variance.

6.3 Partitioning The Variance Of Y

Now that we have introduced how to calculate the different sum of squares we will discuss the partitioning of the variance. When taking any equation you can add and subtract the same value or same additional variable and the original equation stays the same. Eq. 6.6 shows a generall example of this. We simply add and subtract C and therefore does not change the result of the original equation of $A + B$.

Eq. 6.6

$A + B = A + C - C + B$

Recall that part of calculating the variance involves the deviations of the actual data from the mean. Therefore we can use the deviation as an approximation to represent the variance. Eqq. 6.7 shows the deviation. Just as in Eq. 6.6 we can add and subtract the value of $\hat{y_i}$. Eq. 6.8. This will apprimately represent the variance of y.

Eq. 6.7

$y_i - \bar{y}$

Eq. 6.8

$y_i - \hat{y} + \hat{y} - \bar{y}$

If you look closely at Eq. 6.8 you can see that this is now actually a part of the residual sum of squares, Eq. 6.9, and and the regression or explained sum of squares, 6.10.

Eq. 6.9: Residual

$y_i - \hat{y}$

Eq. 6.10: Regression/Explained

$\hat{y} - \bar{y}$

6.3.1 The Relationship Of R-Squared And Partitioned Variance

Because R^2 is the regression or explained sum of squares devided by the total sum of squares we can use the partitioned variance to calculate R^2. Recall that the total sum of squares is simply the residual sum of squares plus the regression sum of squares. Therefore, from that partition of the variance we have the regression or explaned sum of squares and the residual sum of squares. These are all of the components that are needed to calculate R^2 with Eq. 6.4. Another measurement is r, or the correlation. This shows how two variables correlate positively, correlate negatively or not correlated at all. This is related to the covariance. R^2 is simply the square of r.

6.4 Looking Ahead

- We will discuss Bellentine Diagrams and how they relate to correlation.
- We will use the diagrams to visualize the partitioning of the variance now that we have introduced the mathematics.
- We will discuss how the variance of each variable and the size of the circles in the diagrams represent variance.

Chapter 7

Ballentine Venn Diagrams

Diagrams, graphs, and charts can be very helpful in statistics. These are tools that help to visually represent equations and concepts that are used in statistics. Bellentine Venn Diagrams are a great way to many of the concepts that we have just recently discussed by representing them visually.

7.0.1 Concepts That Bellentine Venn Diagrams Can Represent

These diagrams help show the following concepts:

-Variance (larger variance means larger circle)

-Covariance (larger covariance means more overlap between two circles)

-Residual (portion of Y that is not covered by all independent variables)

-Explained variance (portion of the variance of y that is accounted for by x)

-R-square (ratio of the residual to total variance)

-Slope (ratio of covariance to variance of x)

-Standard error (ratio of residual to variance of x)

7.0.2 Variance

The Venn Diagrams very nicely show these statistical concept and the variance is a very clear concept that is represented. Recall that variance is how spread the data is from the mean. Therefore, we can represent the variance with a circle and then the wider the circle is or the larger the diameter is a larger variance. Fig. 7.1 shows that x_1 has a smaller variance than x_2.

Fig. 7.1

Eq. 7.1 shows the variance of y, Eq. 7.2 shows the variance of x_1, and Eq. 7.3 Shows the variance of of x_2.

$var(y) = \Sigma(y_i - \bar{y})^2$

7.1 Covariance

The covariance is represented by the overlap of the circles in the diagrams. Because the covariance shows how two variables relate, this means that they share variation together and therefore must overlap. In Fig.