1

Bi-Variate Hypothesis Testing and Model Fit

The previous chapters discussed the logic of OLS regression and how to derive OLS estimators. Now that simple regression is no longer a mystery, we will shift the focus to bi-variate hypothesis testing and model fit. Note that the examples in this chapter use the `tbur` data set. We recommend that you try the analyses in the chapter as you read.

1.1 Hypothesis Tests for Regression Coefficients

Hypothesis testing is the key to theory building. This chapter is focused on empirical hypothesis testing using OLS regression, using examples drawn from the accompanying dataset (tbur.data). Here we will use the responses to the political ideology question (ranging from 1=strong liberal, to 7=strong conservative), as well as responses to a question concerning the survey respondents' level of risk that global warming poses for people and the environment.[1]

Using the data from these questions, we posit the following hypothesis:

> H_1: On average, as respondents, become more politically conservative, they will be less likely to express increased risk associated with global warming.

The null hypothesis, H_0, is $\beta = 0$, positing that a respondents ideology has no relationship with their views about the risks of global warming for people and the environment. Our working hypothesis, H_1, is $\beta < 0$. We expect β to be less than zero because we expect a *negative* slope between our measures of ideology and levels of risk associated with global warming, given that a larger numeric value for ideology indicates a more conservative respondent. Note that this is a *directional* hypothesis since we are positing a negative relationship. Typically, a directional hypothesis implies a one-tailed test where the critical value is 0.05 on one side of the distribution. A *non-directional* hypothesis, $\beta \neq 0$ does not imply a particular direction, it only implies that there is a relationship. This requires a two-tailed test where the critical value is 0.025 on both sides of the distribution.

[1] The question wording was as follows: "On a scale from zero to ten, where zero means no risk and ten means extreme risk, how much risk do you think global warming poses for people and the environment?"

To test this hypothesis, we run the following code in R.

Before we begin, for this chapter, we need to make a special data set that just contains the variables glbccrisk and ideol with missing values removed.

```
#Filtering a data set with only variables glbcc_risk and ideol
ds.omit <- filter(ds) %>%
  select(glbcc_risk,ideol) %>%
  na.omit()
#Run the na.omit function to removed the value
```

```
ols1 <- lm(glbcc_risk ~ ideol, data = ds.omit)
summary(ols1)
```

```
##
## Call:
## lm(formula = glbcc_risk ~ ideol, data = ds.omit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.726 -1.633  0.274  1.459  6.506
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.81866    0.14189   76.25   <2e-16 ***
## ideol       -1.04635    0.02856  -36.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.479 on 2511 degrees of freedom
## Multiple R-squared:  0.3483,Adjusted R-squared:  0.348
## F-statistic:  1342 on 1 and 2511 DF,  p-value: < 2.2e-16
```

To know whether to accept of reject the null hypothesis, we need to first understand the standard error associated with the model and our coefficients. We start, therefore, with consideration of the residual standard error of the regression model.

1.1.1 Residual Standard Error

The residual standard error (or standard error of the regression), measures spread of our observations around the regression line. As will be discussed below, the residual standard error is used to calculate the standard errors of the regression coefficients, A and B.

The formula for the residual standard error is as follows:

$$S_E = \sqrt{\frac{\Sigma E_i^2}{n-2}} \tag{1.1}$$

To calculate this in R, based on the model we just ran, we create an object called Se and use the sqrt and sum commands.

```
Se <- sqrt(sum(ols1$residuals^2)/(length(ds.omit$glbcc_risk)-2))
Se

## [1] 2.479022
```

Note that this result matches the result provided by the `summary` function in `R`, as shown above.

For our model, the results indicate that: Y_i = 10.8186624 -1.0463463$X_i + E_i$. Another sample of 2513 observations would almost certainly lead to different estimates for A and B. If we drew many such samples, we'd get the sample distribution of the estimates. Because we typically cannot draw many samples, we need to estimate the sample distribution, based on our sample size and variance. To do that, we calculate the standard error of the slope and intercept coefficients, $SE(B)$ and $SE(A)$. These standard errors are our estimates of how much variation we would expect in the estimates of B and A across different samples. We use them to evaluate whether B and A are larger that would be expected to occur by chance, if the real values of B and/or A are zero (the null hypotheses).

The standard error for B, $SE(B)$ is:

$$SE(B) = \frac{S_E}{\sqrt{TSS_X}} \tag{1.2}$$

where S_E is the residual standard error of the regression, (as whown earlier in equation 9.1). TSS_X is the total sum of squares for X, that is the total sum of the squared deviations (residuals) of X from its mean $\bar{X}$; $\sum(X_i - \bar{X})^2$. Note that the greater the deviation of X around its mean as a proportion of the standard error of the model, the smaller the $SE(B)$. The smaller $SE(B)$ is, the less variation we would expect in repeated estimates of B across multiple samples.

The standard error for A, $SE(A)$, is defined as:

$$SE(A) = S_E * \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{TSS_X}} \tag{1.3}$$

Again, the SE is the residual standard error, as shown in equation 9.1.
For A, the larger the data set, and the larger the deviation of X around its mean, the more precise our estimate of A (i.e., the smaller $SE(A)$ will be).

We can calculate the SE of A and B in `R` in a few steps. First, we create an object `TSSx` that is the total sum of squares for the X variable.

```
TSSx <- sum((ds.omit$ideol-mean(ds.omit$ideol, na.rm = TRUE))^2)
TSSx

## [1] 7532.946
```

Then, we create an object called `SEa`.

```
SEa <- Se*sqrt((1/length(ds.omit$glbcc_risk))+(mean(ds.omit$ideol,na.rm=T)^2/TSSx))
SEa

## [1] 0.1418895
```

Finally, we create `SEb`.

```
SEb <- Se/(sqrt(TSSx))
SEb

## [1] 0.02856262
```

Using the standard errors, we can determine how likely it is that our estimate of β differs from 0; that is how many standard errors our estimate is away from 0. To determine this we use the t value. The t score is derived by dividing the regression coefficient by its standard error. For our model, the t value for β is as follows:

```
t <- ols1$coef[2]/SEb
t

##     ideol
## -36.63342
```

The t value for our B is -36.6334214, meaning that B is -36.6334214 standard errors away from zero. We can then ask: What is the probability, p *value*, of obtaining this result if $\beta = 0$? According to the results shown earlier, $p = 2e - 16$. That is remarkably close to zero. This result indicates that we can reject the null hypothesis that $\beta = 0$.

In addition, we can calculate the confidence interval (CI) for our estimate of B. This means that in 95 out of 100 repeated applications, the confidence interval will contain β.

In the following example, we calculate a 95% CI. The CI is calculated as follows:

$$B \pm 1.96(SE(B)) \tag{1.4}$$

We can easily calculate this in `R`. First, we calculate the upper limit then the lower limit and then we use the `confint` function to check.

```
Bhi <- ols1$coef[2]-1.96*SEb
Bhi

##     ideol
## -1.102329

Blow <- ols1$coef[2]+1.96*SEb
Blow

##      ideol
## -0.9903636

confint(ols1)

##                  2.5 %     97.5 %
## (Intercept) 10.540430 11.0968947
## ideol       -1.102355 -0.9903377
```

As shown, the upper limit of our estimated B is -0.9903636, which is far below 0, providing further support for rejecting H_0.

So, using our example data, we tested the working hypothesis that political ideology is negatively related to expressed risk of global warming to people and the environment. Using simple OLS regression, we find support for that working hypothesis, and can reject the null.

1.2 Measuring Goodness of Fit

Once we have constructed a regression model, it is natural to ask: how `good` is the model at explaining variation in our dependent variable? We answer this question with a number of statistics that indicate "model fit". Basically, these statistics provide measures of the degree to which the estimated relationships account for the variance in the dependent variable, Y.

There are several ways to examine how well the model "explains" the variance in Y. First, we can examine the covariance of X and Y, which is a general measure of the sample variance for X and Y. Then we can use a measure of sample `correlation`, which is the standardized measure of covariation. Both of these measures provide indicators of the degree to which variation in X can account for variation in Y. Finally, we can examine R^2, also know as the coefficient of determination, which is the standard measure of the goodness of fit for OLS models.

1.2.1 Sample Covariance and Correlations

The sample covariance for a simple regression model is defined as:

$$S_{XY} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \tag{1.5}$$

Intuitively, this measure tells you, on average, whether a higher value of X (relative to its mean) is associated with a higher or lower value of Y. Is the association negative or positive? Covariance can be obtained quite simply in `R` by using the the `cov` function.

```
Sxy <- cov(ds.omit$ideol, ds.omit$glbcc_risk)
Sxy

## [1] -3.137767
```

The problem with covariance is that its magnitude will be entirely dependent on the scales used to measure X and Y. That is, it is non-standard, and its meaning will vary depending on what it is that is being measured. In order to compare sample covariation across different samples and different measures, we use the sample `correlation`.

The sample correlation, r, is found by dividing S_{XY} by the product of the standard deviations of X, S_X, and Y, S_Y.

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}} \tag{1.6}$$

To calculate this in `R`, we first make an object for S_X and S_Y using the `sd` function.

```
Sx <- sd(ds.omit$ideol)
Sx

## [1] 1.7317

Sy <- sd(ds.omit$glbcc_risk)
Sy

## [1] 3.070227
```

Then to find r:

```
r <- Sxy/(Sx*Sy)
r

## [1] -0.5901706
```

To check this we can use the `cor` function in `R`.

```
rbyR <- cor(ds.omit$ideol, ds.omit$glbcc_risk)
rbyR

## [1] -0.5901706
```

So what does the correlation coefficient mean? The values range from +1 to -1, and a value of +1 means there is a perfect positive relationship between X and Y. Each increment of increase in X is matched by a constant increase in Y – with all observations lining up neatly on a positive slope. A correlation coefficient of -1, a perfect negative relationship, would indicate that each increment of increase in X corresponds to a constant decrease in Y – or a negatively sloped line. A correlation coefficient of zero would describe `no relationship` between X and Y.

1.2.2 Coefficient of Determination: R^2

The most often used measure of goodness of fit for OLS models is R^2. R^2 is derived from three components; the total sum of squares, the explained sum of squares, and the residual sum of squares. R^2 is the ratio of **ESS** (explained sum of squares) to **TSS** (total sum of squares).

The components of R^2 are illustrated in Figure 1.1. As shown, for each observation Y_i, variation around the mean can be decomposed into that which is "explained" by the regression and that which is not. In Figure 1.1, the deviation between the mean of Y and the predicted value of Y, $\hat{Y}$, is the proportion of the variation of Y_i that can be explained (or predicted) by the regression. That is shown as a blue line. The deviation of the observed value of Y_i from the predicted value $\hat{Y}$ (aka the residual, as discussed in the previous chapter) is the unexplained deviation, shown in red. Together, the explained and unexplained variation make up the total variation of Y_i around the mean $\hat{Y}$.

To calculate R^2 "by hand" in `R`, we must first determine the total sum of squares, which is the sum of the squared differences of the observed values of Y from the mean of Y, $\Sigma(Y_i - \bar{Y})^2$. Using `R`, we can create an object called `TSS`.

Components of R^2

- *Total sum of squares (TSS)*: The sum of the squared variance of Y

$$\sum E_i'^2 = \sum (Y - \bar{Y})^2$$

- *Residual sum of squares(RSS)*: The variance of Y not accounted for by the model

$$\sum E_i^2 = \sum (Y - \hat{Y})^2 = \sum (Y_i - A - BX_i)^2$$

- *Explained sum of squares (ESS)*: The variance of Y accounted for in the model
  It is the difference between the TSS and the RSS.

$$ESS = TSS - RSS$$

- R^2: The proportion of the total variance of Y explained by the model

$$R^2 = \frac{ESS}{TSS}$$

$$= \frac{TSS - RSS}{TSS}$$

$$= 1 - \frac{RSS}{TSS}$$

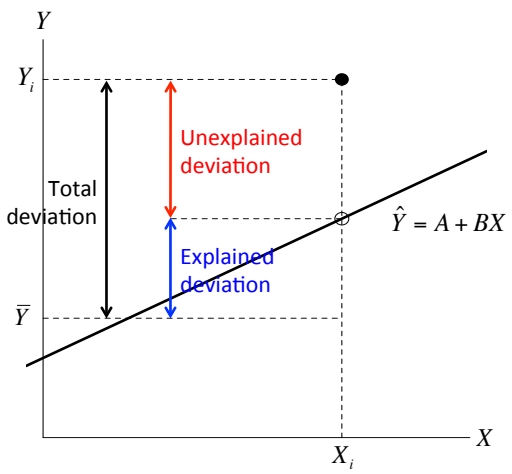

Figure 1.1: The Components of R^2

```
TSS <- sum((ds.omit$glbcc_risk-mean(ds.omit$glbcc_risk))^2)
TSS

## [1] 23678.85
```

Remember that R^2 is the ratio of the explained sum of squares to the total sum of squares (ESS/TSS). Therefore to calculate R^2 we need to create an object called `RSS`, the squared sum of our model residuals.

```
RSS <- sum(ols1$residuals^2)
RSS

## [1] 15431.48
```

Next, we create and object called `ESS`, which is equal to TSS-RSS.

```
ESS <- TSS-RSS
ESS

## [1] 8247.376
```

Finally, we calculate the R^2.

```
R2 <- ESS/TSS
R2

## [1] 0.3483013
```

Note–happily–that the R^2, calculated by "by hand" in `R` matches the results provided by the `summary` command.

The values for R^2 can range from zero to 1. In the simple regression case, a value of 1 indicates that the modeled coefficient (B) "accounts for" all of the variation in Y. Put differently, all of the squared deviations in Y_i around the mean ($\hat{Y}$) are in ESS, with none in the residual (RSS).[2] A value of zero would indicate that all of the deviations in Y_i around the mean are in RSS – all residual or "error". Our example shows that the variation in political ideology (our X) accounts for roughly 34.8 percent of the variation in our measure of perceived risk of global warming (Y).

Visualizing Bivariate Regression

The `ggplot2` package provides a mechanism for viewing the effect of ideology on the dependent variable of perceived risk of global warming. Adding `geom_s`$moothwillcalculateandvisualizearegressionlin$ 7), $theirperceptionoftheriskofglobalwarmingdecreases$.

```
ggplot(ds.omit, aes(ideol, glbcc_risk)) +
  geom_smooth(method = lm)
dev.off()
```

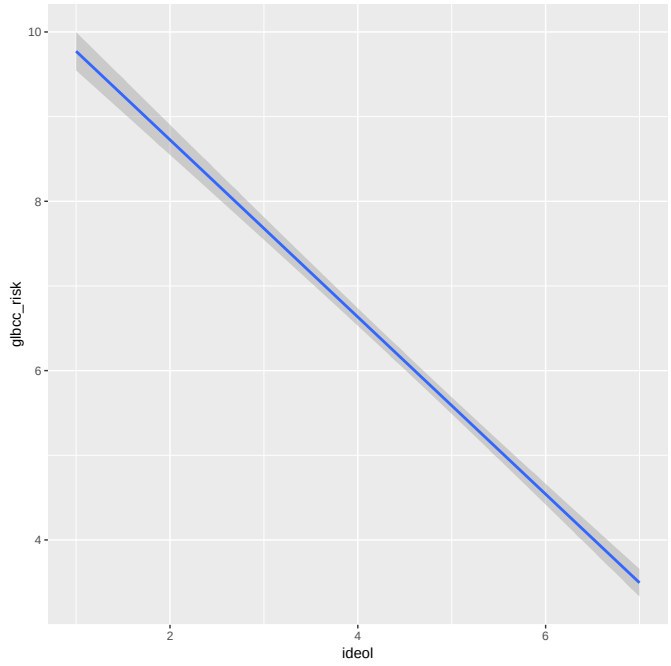

Figure 1.2: Effects Plot of Ideology

`Cleaning up the R Environment`

[2]Note that with a **bivariate model**, R^2 is equal to the square of the correlation coefficient.

If you recall, at the beginning of the chapter, we created several temporary data sets. We should take the time to clear up our workspace for the next chapter.

```
#  we can clean up from the temporary data sets for this chapter.
rm(ds.omit) #remove the omit data set
```

1.3 Summary

This chapter has focused on two key aspects of simple regression models -- hypothesis testing and measures of the goodness of model fit. With respect to the former, we focused on the residual standard error, and its role in determining the probability that our model estimates, B and A, are just random departures from a population in which β and α are zero. We showed, using R, how to calculate the residual standard errors for A and B and, using them, to calculate the t-statistics and associated probabilities for hypothesis testing. For model fit, we focused on model covariation and correlation, and finished up with a discussion of the coefficient of determination -- R^2. So now you are in a position to use simple regression, and to wage unremitting geek-war on those whose models are endowed with lesser $R^2 s$.