1 The Logic of Multiple Regression

The logic of multiple regression can be readily extended from our earlier discussion of simple regression. As with simple regression, multiple regression finds the regression line (or regression "plane" with multiple independent variables) that minimizes the sum of the squared errors. This chapter discusses the theoretical specification of the multiple regression model, the key assumptions necessary for the model to provide the best linear unbiased estimates (BLUE) of the effects of the Xs on Y, the meaning of the partial regression coefficients, and hypothesis testing. Note that the examples in this chapter use the `tbur` data set.

1.1 Theoretical Specification

As with simple regression, the theoretical multiple regression model contains a **systematic** component—$Y = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$ and a **stochastic** component—ϵ_i. The overall theoretical model is expressed as:

$$Y = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

where

- α is the constant term
- β_1 through β_k are the parameters of IVs 1 through k
- k is the number of IVs
- ϵ is the error term

In matrix form the theoretical model can be much more simply expressed as: $y = X\beta + \epsilon$. The empirical model that will be estimated can be expressed as:

$$\begin{aligned} Y_i &= A + B_1 X_{i1} + B_2 X_{i2} + \ldots + B_k X_{ik} + E_i \\ &= \hat{Y}_i + E_i \end{aligned}$$

Therefore, the residual sum of squares (RSS) for the model is expressed as:

$$\begin{aligned} RSS &= \sum E_i^2 \\ &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - (A + B_1 X_{i1} + B_2 X_{i2} + \ldots + B_k X_{ik}))^2 \end{aligned}$$

1.1.1 Assumptions of OLS Regression

There are several important assumptions necessary for multiple regression. These assumptions include linearity, fixed X's, and errors that are normally distributed.

OLS Assumptions

Systematic Component

- Linearity
- Fixed X

Stochastic Component

- Errors have identical distributions
- Errors are independent of X and other ϵ_i
- Errors are normally distributed

Linearity

When OLS is used, it is assumed that a linear functional form is the correct specification for the model being estimated. Note that linearity is assumed in the *parameters* (that is, for the Bs), therefore the expected value of the dependent variable is a linear function of the parameters, not necessarily of the variables themselves. So, as we will discuss in the next chapter, it is possible to transform the variables (the Xs) to introduce non-linearity into the model while retaining linear estimated coefficients. For example, a model with a squared X term can be estimated with OLS:

$$Y = A + BX_i^2 + E$$

However, a model with a squared B term cannot.

Fixed X

The assumption of fixed values of X means that the value of X in our observations is not systematically related to the value of the other X's. We can see this most clearly in an

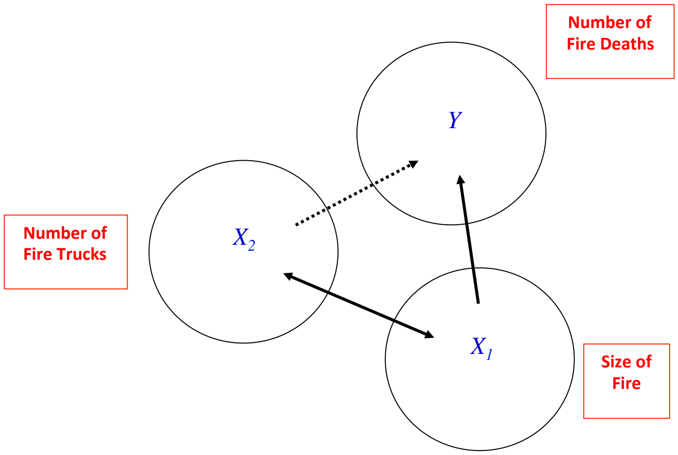

Figure 1.1: Spurious Relationships

experimental setting where the researcher can manipulate the experimental variable while controlling for all other possible Xs through random assignment to a treatment and control group. In that case, the value of the experimental treatment is completely unrelated to the value of the other Xs – or, put differently, the treatment variable is orthogonal to the other Xs. This assumption is carried through to observational studies as well. Note that if X is assumed to be fixed, then changes in Y are assumed to be a result of the independent variations in the X's and error (and nothing else).

1.2 Partial Effects

As noted in Chapter 1, multiple regression "controls" for the effects of other variables on the dependent variables. This is in order to manage possible spurious relationships, where the variable Z influences the value of both X and Y. Figure 1.1 illustrates the nature of spurious relationships between variables.

To control for spurious relationships, multiple regression accounts for the **partial effects** of one X on another X. Partial effects deal with the shared variance between Y and the X's. This is illustrated in Figure 1.2. In this example, the number of deaths resulting from house fires is positively associated with the number of fire trucks that are sent to the scene of the fire. A simple-minded analysis would conclude that if fewer trucks are sent, fewer fire-related deaths would occur. Of course, the number of trucks sent to the fire, and the number of fire-related deaths, are both driven by the magnitude of the fire. An appropriate control for the size of the fire would therefore presumably eliminate the positive association between the number of fire trucks at the scene and the number of deaths (and may even reverse the direction of the relationship, as the larger number of trucks may more quickly

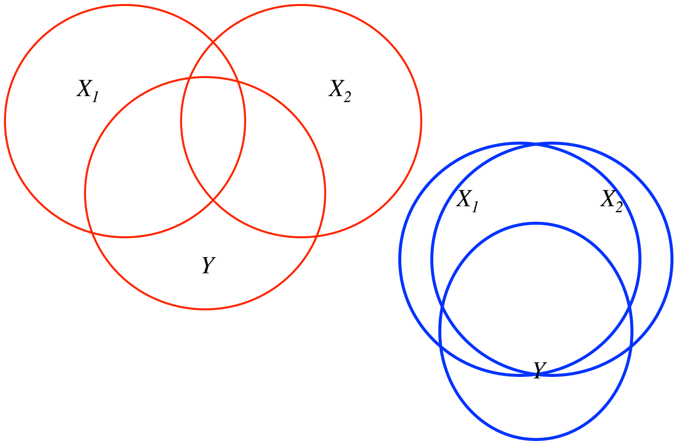

Figure 1.2: Partial Effects

suppress the fire).

In Figure 1.2, the Venn diagram on the left shows a pair of Xs that would jointly predict Y better than either X alone. However, the overlapped area between X_1 and X_2 causes some confusion. That would need to be removed to estimate the "pure" effect of X_1 on Y. The diagram on the right represents a dangerous case. Overall, X_1+X_2 explain Y well, but we don‘t know how the individual X_1 or X_2 influence Y. This clouds our ability to see the effects of either of the Xs on Y. In the extreme case of wholly overlapping explanations by the IVs, we face the condition of **multicolinearity** that makes estimation of the partial regression coefficients (the Bs) impossible.

In calculating the effect of X_1 on Y, we need to remove the effect of the other X's on both X_1 and Y. While multiple regression does this for us, we will walk through an example to illustrate the concepts.

As an example, we will use age and ideology to predict perceived climate change risk.

```
ds.temp <- filter(ds) %>% select(glbcc_risk, ideol, age) %>%
  na.omit()

ols1 <- lm(glbcc_risk ~ ideol+age, data = ds.temp)
summary(ols1)

##
## Call:
## lm(formula = glbcc_risk ~ ideol + age, data = ds.temp)
##
## Residuals:
```

Partial Effects

In a case with two IVs, X_1 and X_2

$$Y = A + B_1 X_{i1} + B_2 X_{i2} + E_i$$

- Remove the effect of X_2 and Y

$$\hat{Y}_i = A_1 + B_1 X_{i2} + E_{iY|X_2}$$

- Remove the effect of X_2 on X_1:

$$\hat{X}_i = A_2 + B_2 X_{i2} + E_{iX_1|X_2}$$

So,

$$E_{iY|X_2} = 0 + B_3 E_{iX_1|X_2}$$

and

$$B_3 E_{iX_1|X_2} = B_1 X_{i1}$$

```
##     Min      1Q  Median      3Q     Max
## -8.7913 -1.6252  0.2785  1.4674  6.6075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.096064   0.244640  45.357   <2e-16 ***
## ideol       -1.042748   0.028674 -36.366   <2e-16 ***
## age         -0.004872   0.003500  -1.392    0.164
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.479 on 2510 degrees of freedom
## Multiple R-squared:  0.3488,Adjusted R-squared:  0.3483
## F-statistic: 672.2 on 2 and 2510 DF,  p-value: < 2.2e-16
```

Note that the estimated coefficient for ideology is -1.0427478. To see how multiple regression removes the shared variance we first regress climate change risk on age and create an object `ols2.resids` of the residuals.

```
ols2 <- lm(glbcc_risk ~ age, data = ds.temp)
summary(ols2)

##
```

```
## Call:
## lm(formula = glbcc_risk ~ age, data = ds.temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4924 -2.1000  0.0799  2.5376  4.5867
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.933835   0.267116  25.958  < 2e-16 ***
## age         -0.016350   0.004307  -3.796  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.062 on 2511 degrees of freedom
## Multiple R-squared:  0.005706,Adjusted R-squared:  0.00531
## F-statistic: 14.41 on 1 and 2511 DF,  p-value: 0.0001504

ols2.resids <- ols2$residuals
```

Note that, when modeled alone, the estimated effect of age on glbccrsk is larger (-0.0164) than it was in the multiple regression with ideology (-0.00487). This is because age is correlated with ideology, and – because ideology is also related to glbccrsk – when we don't "control for" ideology the age variable carries some of the influence of ideology.
Next, we regress ideology on age and create an object of the residuals.

```
ols3 <- lm(ideol ~ age, data = ds.temp)
summary(ols3)

##
## Call:
## lm(formula = ideol ~ age, data = ds.temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9492 -0.8502  0.2709  1.3480  2.7332
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.991597   0.150478  26.526  < 2e-16 ***
## age         0.011007   0.002426   4.537 5.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.725 on 2511 degrees of freedom
## Multiple R-squared:  0.00813,Adjusted R-squared:  0.007735
## F-statistic: 20.58 on 1 and 2511 DF,  p-value: 5.981e-06

ols3.resids <- ols3$residuals
```

Finally, we regress the residuals from ols2 on the residuals from ols3. Note that this regression does not include an intercept term.

```
ols4 <- lm(ols2.resids ~ 0 + ols3.resids)
summary(ols4)

##
## Call:
## lm(formula = ols2.resids ~ 0 + ols3.resids)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7913 -1.6252  0.2785  1.4674  6.6075
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## ols3.resids -1.04275    0.02866  -36.38   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.478 on 2512 degrees of freedom
## Multiple R-squared:  0.3451,Adjusted R-squared:  0.3448
## F-statistic:  1324 on 1 and 2512 DF,  p-value: < 2.2e-16
```

As shown, the estimated B for $E_{iX_1|X_2}$, matches the estimated B for ideology in the first regression. What we have done, and what multiple regression does, is “clean” both Y and X_1 (ideology) of their correlations with X_2 (age) by using the residuals from the bivariate regressions.

1.3 Multiple Regression Example

In this section, we walk through another example of multiple regression. First, we start with our two IV model.

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

##
## Call:
## lm(formula = glbcc_risk ~ age + ideol, data = ds.temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7913 -1.6252  0.2785  1.4674  6.6075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.096064   0.244640  45.357   <2e-16 ***
## age         -0.004872   0.003500  -1.392    0.164
## ideol       -1.042748   0.028674 -36.366   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.479 on 2510 degrees of freedom
## Multiple R-squared:  0.3488,Adjusted R-squared:  0.3483
## F-statistic: 672.2 on 2 and 2510 DF,  p-value: < 2.2e-16
```

The results show that the relationship between age and perceived risk (glbccrsk) is negative and insignificant. The relationship between ideology and perceived risk is also negative and significant. The coefficients of the X's are interpreted in the same way as with simple regression, except that we are now controlling for the effect of the other X's by removing their influence on the estimated coefficient. Therefore, we say that as ideology increases one unit, perceptions of the risk of climate change (glbccrsk) decrease by -1.0427478, controlling for the effect of age.

As was the case with simple regression, multiple regression finds the intercept and slopes that minimize the sum of the squared residuals. With only one IV the relationship can be represented in a two-dimensional plane (a graph) as a line, but each IV adds another dimension. Two IVs create a regression plane within a cube, as shown in Figure 1.3. The Figure shows a scatterplot of GCC risk, age, and ideology coupled with the regression plane. Note that this is a sample of 200 observations from the larger data set. Were we to add more IVs, we would generate a hypercube... and we haven't found a clever way to draw that yet.

In the next example education is added to the model.

```
ds.temp <- filter(ds) %>%
  select(glbcc_risk, age, education, income, ideol) %>%
  na.omit()

ols2 <- lm(glbcc_risk ~ age + education + ideol, data = ds.temp)
summary(ols2)

##
## Call:
## lm(formula = glbcc_risk ~ age + education + ideol, data = ds.temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8092 -1.6355  0.2388  1.4279  6.6334
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.841669   0.308416  35.153   <2e-16 ***
## age         -0.003246   0.003652  -0.889    0.374
## education    0.036775   0.028547   1.288    0.198
## ideol       -1.044827   0.029829 -35.027   <2e-16 ***
## ---
```

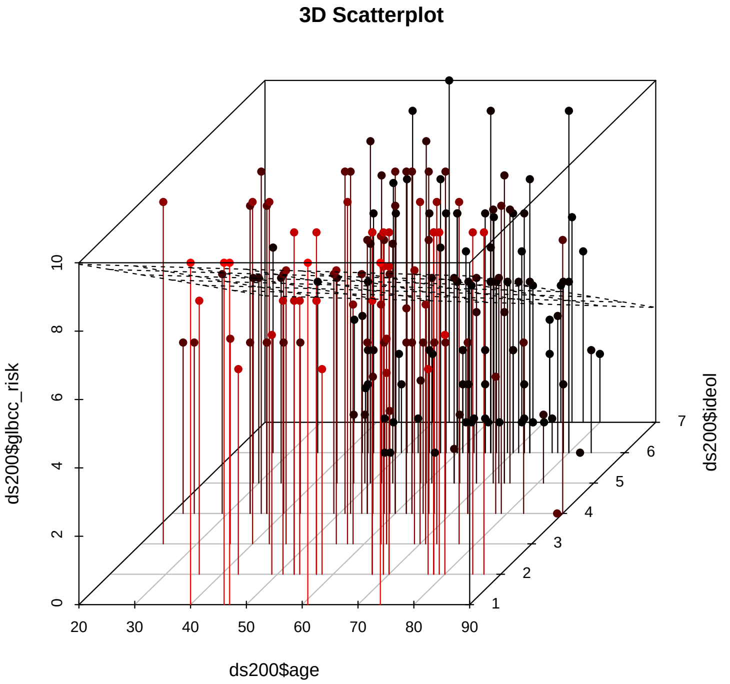

Figure 1.3: Scatterplot and Regression Plane of gcc risk, age, and ideology

```
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.437 on 2268 degrees of freedom
## Multiple R-squared:  0.3607,Adjusted R-squared:  0.3598
## F-statistic: 426.5 on 3 and 2268 DF,  p-value: < 2.2e-16
```

We see that as a respondent’s education increases one unit on the education scale, perceived risk appears to increase by 0.0367752, keeping age and ideology constant. However, this result is not significant. In the final example income is added to the model. Note that the size and significance of education actually increases once income is included, indicating that education only has bearing on the perceived risks of climate change once the independent effect of income is considered.

```
options(scipen = 999)#to turn off scientific notation
ols3 <- lm(glbcc_risk ~ age + education + income + ideol, data = ds.temp)
summary(ols3)

##
## Call:
## lm(formula = glbcc_risk ~ age + education + income + ideol, data = ds.temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7991 -1.6654  0.2246  1.4437  6.5968
##
## Coefficients:
##                   Estimate    Std. Error t value             Pr(>|t|)
## (Intercept) 10.9232861851  0.3092149750  35.326 < 0.0000000000000002 ***
## age         -0.0044231931  0.0036688855  -1.206              0.22810
## education    0.0632823391  0.0299443094   2.113              0.03468 *
## income      -0.0000026033  0.0000009021  -2.886              0.00394 **
## ideol       -1.0366154295  0.0299166747 -34.650 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.433 on 2267 degrees of freedom
## Multiple R-squared:  0.363,Adjusted R-squared:  0.3619
## F-statistic:   323 on 4 and 2267 DF,  p-value: < 0.00000000000000022
```

1.3.1 Hypothesis Testing and t-tests

The logic of hypothesis testing with multiple regression is a straightforward extension from simple regression as described in Chapter 7. Below we will demonstrate how to use the standard error of the ideology variable to test whether ideology influences perceptions of the perceived risk of global climate change. Specifically, we posit:

> H_1: As respondents become more conservative, they will perceive climate change to be less risky, all else equal.

Therefore, $\beta_{ideology} < 0$. The null hypothesis is that $\beta_{ideology} = 0$.

To test H_1 we first need to find the standard error of the B for ideology, (B_j).

$$SE(B_j) = \frac{S_E}{\sqrt{RSS_j}} \tag{1.1}$$

where $RSS_j =$ the residual sum of squares from the regression of X_j (ideology) on the other Xs (age, education, income) in the model. RSS_j captures all of the **independent** variation in X_j. Note that the bigger RSS_j, the smaller $SE(B_j)$, and the smaller $SE(B_j)$, the more precise the estimate of B_j.

S_E (the standard error of the model) is:

$$S_E = \sqrt{\frac{RSS}{n-k-1}}$$

We can use `R` to find the RSS for ideology in our model. First we find the S_E of the model:

```
Se <- sqrt((sum(ols3$residuals^2))/(length(ds.temp$ideol)-5-1))
Se

## [1] 2.43312
```

Then we find the RSS, for ideology:

```
ols4 <- lm(ideol ~ age + education + income, data = ds.temp)
summary(ols4)

##
## Call:
## lm(formula = ideol ~ age + education + income, data = ds.temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2764 -1.1441  0.2154  1.4077  3.1288
##
## Coefficients:
##                   Estimate    Std. Error t value             Pr(>|t|)
## (Intercept)  4.5945481422  0.1944108986  23.633 < 0.0000000000000002 ***
## age          0.0107541759  0.0025652107   4.192   0.0000286716948757 ***
## education   -0.1562812154  0.0207596525  -7.528   0.0000000000000738 ***
## income       0.0000028680  0.0000006303   4.550   0.0000056434561990 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.707 on 2268 degrees of freedom
## Multiple R-squared:  0.034,Adjusted R-squared:  0.03272
## F-statistic:  26.6 on 3 and 2268 DF,  p-value: < 0.00000000000000022

RSSideol <- sum(ols4$residuals^2)
RSSideol

## [1] 6611.636
```

Finally, we calculate the SE for ideology:

```
SEideol <- Se/sqrt(RSSideol)
SEideol

## [1] 0.02992328
```

Once the $SE(B_j)$ is known, the t-test for the ideology coefficient can be calculated. The t value is the ratio of the estimated coefficient to its standard error.

$$t = \frac{B_j}{SE(B_j)} \tag{1.2}$$

This can be calculated using R.

```
ols3$coef[5]/SEideol

##     ideol
## -34.64245
```

As we see, the result is statistically significant, and therefore we reject the null hypothesis. Also note that the results match those from the R output for the full model, as was shown earlier.

1.4 Summary

The use of multiple regression, when compared to simple bivariate regression, allows for more sophisticated and interesting analyses. The most important feature is the ability of the analyst (that's you!) to statistically control for the effects of all other IVs when estimating any B. In essence, we "clean" the estimated relationship between any X and Y of the influence of all other Xs in the model. Hypothesis testing in multiple regression requires that we identify the independent variation in each X, but otherwise the estimated standard error for each B is analogous to that for simple regression.

So, maybe it's a little more complicated. But look at what we can observe! Our estimates from the examples in this chapter show that age, income and education are all related to political ideology, but even when we control for their effects, ideology retains a potent influence on the perceived risks of climate change. Politics matters.