Regression Interpretation and Diagnostics

Hisam Sabouni, PhD

April 2020

Overview

We've gone through all of the mathematics of deriving the optimal solution to ordinary least squares estimator. Furthermore, we were able to derive the sample properties of our estimator to show that OLS is an unbiased. We also were able to derive the variance of OLS which we used introduce hypothesis testing. Today we're going to take a step back from the details of the math for a moment and focus in on the following topics:

1. Interpretation of regression models
2. Data scaling of regression models
3. Regression diagnostics
4. Collected data from the internet (web-scraping; see other lecture)

Interpretation

So far in this class we have relied on using simulations to derive properties of our estimators. Simulations are great because we can create a true underlying data generating process, take a sample, and see if we can recover the true model using just a sample. While useful mathematically this distances us as econometricians from the underlying interpretation of the estimated relationships. So we are going to go through a couple of examples and walk through the interpretation.

Lets get started by going back to the data on test scores we have covered in the past in the AER package. The dataset contains data on test performance, school characteristics and student demographic backgrounds for school districts in California.

The relationship we looked at previously related the ideas of classroom size to performance on test scores:

$$TotalScore = ReadScore + MathScore$$

$$Student - to - Teacher(STR) = \frac{\text{Total Enrollment}}{\text{Number of Teachers}}$$

The simple specification we looked at was as follows:

$$TotalScore = \hat{\beta}_0 + \hat{\beta}_1 STR + \epsilon_i$$

How do we interpret this model?

What does $\hat{\beta}_0$ represent? What does $\hat{\beta}_1$ represent? Whenever you see a regression you need to think about the units of the dependent and independent variables. The units are key to interpreting the model. Now that we know the underlying mathematics of OLS and that we can implement OLS on our own, we are now allowed to use the built in function in R to estimate a regression:

```
##
## Call:
## lm(formula = totalScore ~ STR, data = CASchools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.453 -28.501   0.965  25.644  97.081
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1397.8659    18.9350  73.825  < 2e-16 ***
## STR           -4.5596     0.9597  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.16 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Our estimated model is as follows:

$$TotalScore = 1397.8659 - 4.5596 STR + \hat{\epsilon}$$

We can interpret this base model as follows:

Is our model more informative than a simple model that includes only an intercept term? (Hint: think about R^2).

What is the intercept? 1397.8659. What does the intercept mean? Well the intercept is the expected value of Y whenever all of the independent varibles are set to 0.

What is it the estimated partial effect of STR on $TotalScore$? The partial effect of STR on the average score in a district is measuring how the average score ina district changes as the student to teacher ratio changes (→ derivative).

$$\frac{\partial TotalScore}{\partial STR} = -4.5596$$

As the student to teacher ratio increases we expect the averrage total score for the school district to decline becasuse there is negative and statistically significant from zero sign on $\hat{\beta}_1$. Given our estimate we expect the average total score per districts to decline by about 4.5 points as the student to teacher ratio increases by one unit. You can see this by modifying the derived derivative above as follows:

$$\partial TotalScore = -4.5596 \ \partial STR$$

Now while the interpretation here is pretty clear is there any room for improvement? Intuitively its straight forward to understand the idea of a student to teacher ratio but perhaps it would be easier to interpret our model interms of percentages. That is it may be more helpful to ask the question of what happens to the average score in a district if the student to teacher ratio increases by ___%. To estimate the model and get the interpretation in terms of percentages we don't need any additional data. We can simply respecify our model in terms of natural logarithms. Recall the derivative of a logarithm:

$$\frac{\partial log(f(x))}{\partial x} = \frac{1}{f(x)}\frac{\partial f(x)}{\partial x}$$

If we estimate our model using the natural log of the student to teacher ratio, our model becomes:

$$TotalScore = \hat{\beta}_0 + \hat{\beta}_1 log(STR) + \epsilon_i$$

The partial effect of STR on average test score per district becomes:

$$\frac{\partial TotalScore}{\partial STR} = \hat{\beta}_1 \frac{1}{STR}$$

or,

$$\partial TotalScore = \hat{\beta}_1 \frac{1}{STR} \partial STR$$

If we multiply both sides of the derivative by 100, this interpretation is now in terms of percentages!

$$100 \partial TotalScore = \hat{\beta}_1 100 \frac{1}{STR} \partial STR$$

$$\partial TotalScore = \frac{\hat{\beta}_1}{100} \partial STR\%$$

Lets go back to the data and see how this changes our estimated parameters. Again, note we are not changing ANY of the data. We are simply changing the scaling of our independent variable:

```
##
## Call:
## lm(formula = totalScore ~ logSTR, data = CASchools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.031 -28.561   0.968  25.620  97.199
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1571.67      55.18  28.481  < 2e-16 ***
## logSTR        -88.59      18.55  -4.775 2.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.15 on 418 degrees of freedom
## Multiple R-squared:  0.05173,    Adjusted R-squared:  0.04946
## F-statistic:  22.8 on 1 and 418 DF,  p-value: 2.489e-06
```

The intercept of our model is now 1571.67. As the independent variable $log(STR)$ changes we expect the average test score per district to decline. In otherwords, a one-percent increase in the student to teacher ratio leads to a decline of 0.8859 points. A more economical example would be if the student to teacher ratio increases by 10% the expected average test score is expected to decline by 8.85 points.

```
##         2
## -8.858759
```

What we've estimated here is a level-log regression model. Typically we have been working with level-level models. At this link there is a nice summary of the interpretation of various combinations: https://sites.google.com/site/curtiskephart/ta/econ113/interpreting-beta.

What's another usefule transformation that can be done to help with model interpretation?

Diagnostics

When estimating a model it is pretty important to have a good specification. That is, as the econmetrician you need to decide what is included in your regression and what is omitted. Your choices will have consequences. When covering the mathematical derivation of OLS we discussed the idea of multicollinearity. We stated that in order for us to solve for the optimal $\hat{\beta}$ that minimizes the sum of squares of the residuals we would need to invert our matrix of independent variables, as a result, we could not have any variables that were linear combinations of other variables. This is informative. We can not ever include variables that are perfect linear combinations of other variables. A simple example to think of this is known as a ‘dummy trap’. We typically think of indpendent variables as continuous variables, we can also have categorical variables, or dummy variables, in our model. For instance in our data on schools we create a categorical variable for the grades that a school serves:

```
## 
## KK-06 KK-08 
##    61   359
```

In our sample, there are 61 schools that serve grades kindergarten through grade six and 359 schools that serve kindergarten through grade eight. Lets create a new column kk_06 which will take on a value of 1 if a school serves kk through grade six and 0 otherwise. We will create the second column called kk_08 which will take on a value of 1 if a school serves kk through grade eight and 0 otherwise.

Can we include both of these dummy variables in our regression? If not, why?

```
## 
## Call:
## lm(formula = totalScore ~ logSTR, data = CASchools)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -95.031 -28.561   0.968  25.620  97.199 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1571.67      55.18  28.481  < 2e-16 ***
## logSTR        -88.59      18.55  -4.775 2.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 37.15 on 418 degrees of freedom
## Multiple R-squared:  0.05173,	Adjusted R-squared:  0.04946 
## F-statistic:  22.8 on 1 and 418 DF,  p-value: 2.489e-06
```

Clearly, we have an issue of multicollinearity! We can only control for one category at at time for one category necessarily defines the other category.

What about the in-between case, when we leave out a variable that we just so happen to think isn't all that important. For instance, in our CASchools dataset there is a column called 'english' which contains data on the percentage of english learners. When we leave out a variable like english from our model we are essentially allowing the variation that could be attributed to that variable in the error, or, residual term. Recall that one of the core assumptions of OLS is that the residuals should be uncorrelated with any of the indpendent variables: $E(\epsilon|X) = 0$.

If X does happen to be correlated with a term in the residual our estimate of $\hat{beta}$ will be biased. Lets walk through the math than get back to our example:

Suppose we think the true relationship is specficied as $Y = X\beta + \epsilon$

We've shown that our optimal solution for OLS is $\hat{\beta} = (X'X)^{-1}X'Y$

If the true relationship is actually $Y = X\beta + Z\gamma + \omega$, that is contains ommitted variables Z from the model, then our OLS solution is actually giving us:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + Z\gamma + \omega)$$

$$\hat{\beta} = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'\omega$$

$$\hat{\beta} = \beta + (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'\omega$$

Now assuming that ω is entirely uncorrelated with X, if we take the expectation conditional on the sample we observe:

$$E(\hat{\beta}|X) = \beta + (X'X)^{-1}E(X'Z|X)\gamma$$

$$E(\hat{\beta}|X) = \beta + \text{bias}$$

If any of our indpendent variables happen to be correlated with a variable we happen to omit from our model our estimate of $\hat{\beta}$ will be biased! The extent of the bias is going to depend on how strongly correlated the omitted variable is with the variable we are controlling for.

For example, in the regression we ran earlier:

$$TotalScore = \hat{\beta}_0 + \hat{\beta}_1 log(STR) + \epsilon_i$$

By choosing to omit the variable that accounts for the proportion of english learners in a given school we may have omitted variable bias in our model. We can check if this is problematic by checking the correlation between 'english' and 'log(STR)'.

```
## [1] -0.2274364
```

```
## [1] 0.1908914
```

Here we see that there is a strong positive correlation between porportion of english learners and the log of the student to teacher ratio. This indicates that our original estimate of $\hat{\beta}_1$ is biased as we omitted an important variable. In absolute terms this indicates that our estimate of $|\hat{\beta}_1|$ is too large. By including the additional control we should expect for the effect of student to teacher ratios to decline in absolute terms: $|\hat{\beta}_1|$.

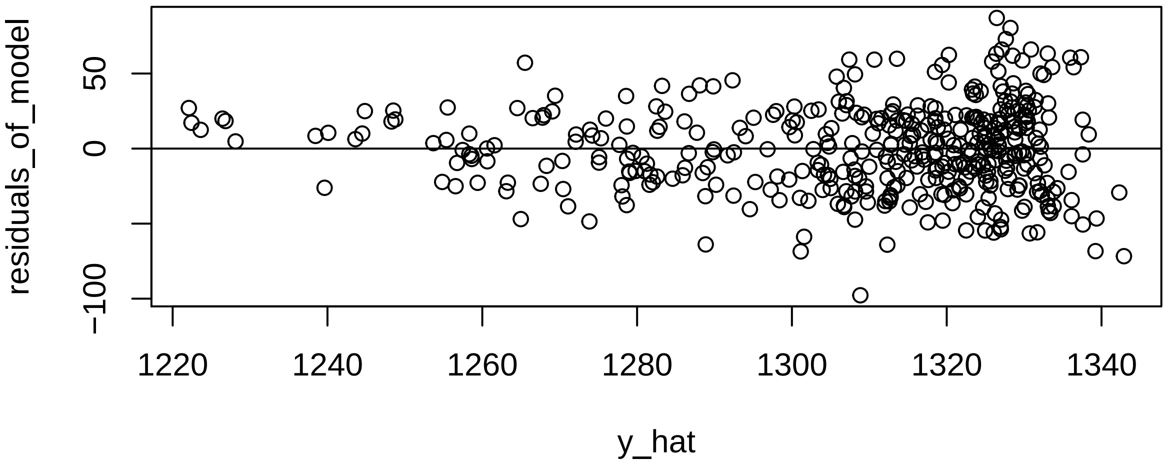
```
## 
## Call:
## lm(formula = totalScore ~ logSTR + english, data = CASchools)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -97.74 -20.29  -0.70  19.67  87.03 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1454.35507   43.56048  33.387  < 2e-16 ***
## logSTR       -42.23397   14.71894  -2.869  0.00432 ** 
## english       -1.29918    0.07875 -16.498  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 28.93 on 417 degrees of freedom
## Multiple R-squared:  0.4262, Adjusted R-squared:  0.4235
## F-statistic: 154.9 on 2 and 417 DF,  p-value: < 2.2e-16
```

As you can see nailing down a proper specification is pretty important! The degree of omitted variable bias in this case was so large that it cut the magnitude of $\hat{\beta}$ in half.

So, as of now you should be aware of three new things: (1) dummy, or, categorical variables, (2) what happens when we have perfect multicollinearity, and (3) what happens when we have ommitted variable bias. Lets take a look back at the interpretation of our model now that we have added in a new variable. What is the partial effect of the natural logarithm of the student to teacher ratio? What is the partial effect of the percent of english learners?

Given the interpreation, think about our discussion of the Gauss Markov assumptions. The standard errors R is reporting assume no correlation between the residuals and the indpendent variables. The standard errors R is reporting allso assume homoscedasticity We can check if we are satisfying homoscedasticity (constant variance) in the residuals by simply plotting out the residuals against our predicted values of the depdendent variable.



Interestingly as the predicted test score increases we seem to have increased error. This would be a direct violation of the assumption of homoscedasticity. As a result, we should use our robust estimator which allows for variation in the variance of the errors in the variance-covariance matrix of the residual terms.

```
## (Intercept)      logSTR     english
## 52.16679712 17.41653247  0.06194277
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value Pr(>|t|)
## (Intercept) 1454.355068   52.166797  27.8789  < 2e-16 ***
## logSTR       -42.233969   17.416532  -2.4249  0.01573 *
## english       -1.299178    0.061943 -20.9739  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

While each of our explanatory variables is still statistically significant by using the corrected standard errors we definitely see our standard errors inflate.

Now lets jump to collecting data from anywhere on the internet.