Regression in R

Vaidehi Bulusu

Introduction

This notebook is a walkthrough of how to perform different kinds of regression in R. It contains explanations, demos and some practice questions. Feel free to experiment on your own with the dataset and techniques!

We will be looking at how to perform 2 main kinds of regression that you have learned in Econ 140:

1. Ordinary Least Squares (OLS) Regression
2. Instrumental Variables (IV) Regression (also called Two-Stage Least Squares, or TSLS)

Note that in the interest of time, we will not be explaining all the concepts in depth so you might also want to reference your lecture notes (and other course materials) to answer the questions. We will be using the `mexico.csv` dataset that you'll be using for Problem Set 2.

```
# upload the mexico.csv dataset: Session --> Set Working
# Directory (choose the folder that has the file you want
# to upload) --> run the cell below
mexico <- read.csv("mexico.csv")

# let's look at the first few rows
head(mexico, 10)
```

```
##    year municode    inc_m  ind_lang educ_years sales_hotel  logtemp logprecip
## 1  2000     1001 3367.053 0.0039165   9.087990      171550 5.171715  3.796707
## 2  2000     1002 2287.634 0.0012442   6.795107          40 5.140995  3.656000
## 3  2000     1003 1786.198 0.0027569   6.000360         687 5.169001  3.980089
## 4  2000     1005 2559.550 0.0015731   7.319565         361 5.147138  3.864694
## 5  2000     1006 3021.914 0.0003432   8.330784         378 5.149746  3.673766
## 6  2000     1007 2555.663 0.0000000   7.991003         574 5.113169  3.660473
## 7  2000     1011 3360.296 0.0019324   7.784682         729 5.171446  3.648598
## 8  2000     2001 4021.991 0.0806629   8.217420      152591 5.162446  2.744925
## 9  2000     2002 6460.129 0.0075334   9.065122      302796 5.315036  2.188445
## 10 2000     2003 4474.784 0.0415555   7.776525        7259 5.007053  3.305918
##    dist_us_km
## 1   600.69310
## 2   557.61430
## 3   616.01170
## 4   597.93450
## 5   573.91170
## 6   561.40970
## 7   576.39320
## 8   262.42290
## 9    90.35691
## 10   18.29244
```

Exploring Your Data

Before performing any kind of regression, you will first be doing some exploratory data analysis (or EDA) on your data. This involves tasks such as:

- Making sure that your data is the right format (e.g. if a variable `wages` is stored as a string, converting it to an numeric data type)
- Getting rid of unnecessary columns
- Creating dummy variables for categorical variables

We'll leave it to a data science class to teach you how to perform EDA, but in this class, you'll need to know how to do 2 main EDA tasks:

1. Data visualization, to visualize the relationship between your independent and dependent variable
2. Transforming your data so that you can fit a linear model (e.g. log and quadratic transformations, which you did in your problem sets and section assignments)

For our regression below, we'll be looking at the relationship between `sales_hotel` (our y variable) and `ind_lang` (our x variable). The question we are trying to answer is: do municipalities in which a higher percentage of people speak the indigenous language have more tourism (as measured by the total hotel sales for that municipality)?

Note on granularity: In this dataset, each row represents one municipality (you can verify this by looking at the `municode` column, which contains unique values - so each row corresponds to one municipality). This is also called the granularity of our dataset. It's important to keep this in mind as it affects how we interpret our coefficients.

Question: What do you think the relationship between `sales_hotel` and `ind_lang` would be (do you expect the relationship to be positive, negative or nothing at all)? We're not expecting a particular answer, we just want to get you thinking. You're free to come up with creative explanations!

Solution: One possibility is that there is a negative relationship between `sales_hotel` and `ind_lang`. Maybe places in which a higher proportion of people speak the native langauge are more rural and therefore have fewer high-rated hotels.

Data Visualization and Transformations

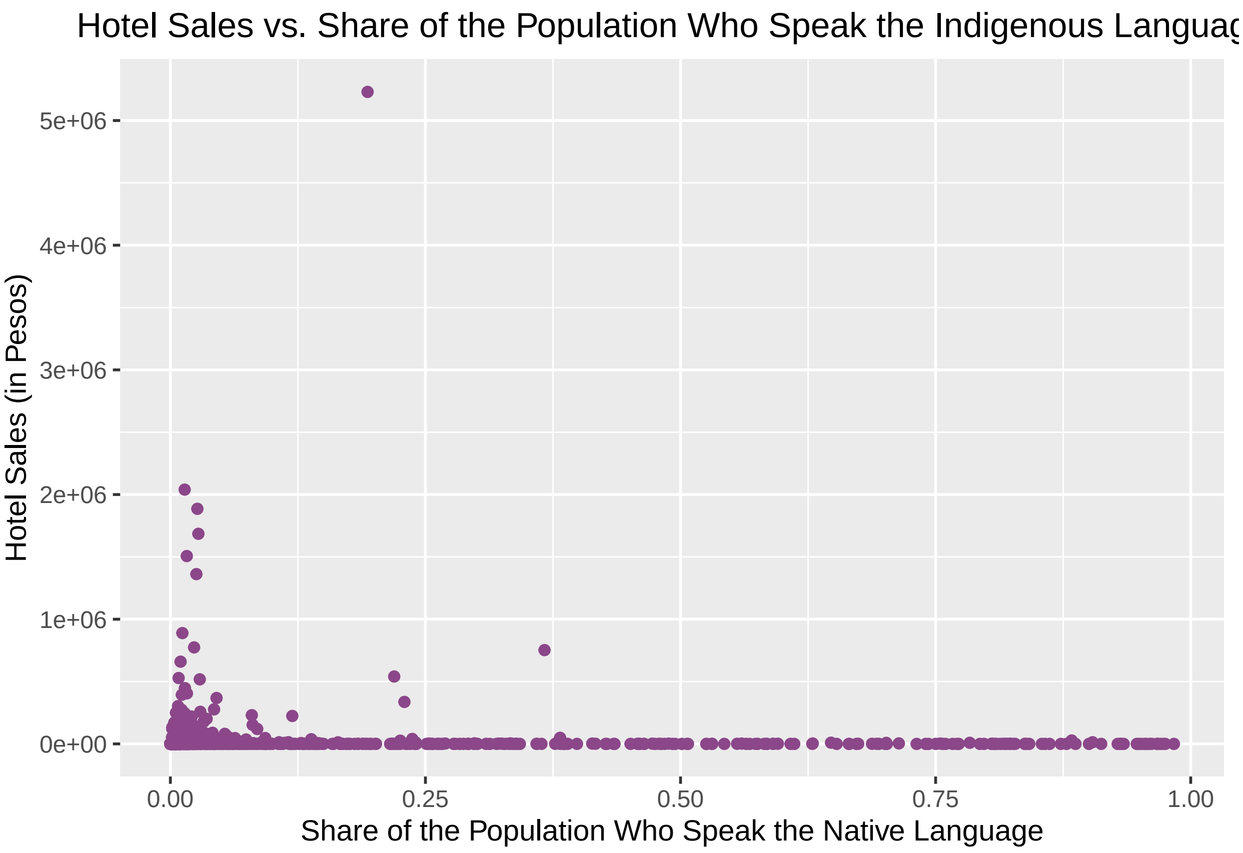
It's always a good idea to visualize your dependent variable against your independent variable before performing regression, to get a sense of their relationship. This will allow you to:

- Know what to expect, in terms of the sign and strength of the relationship between the variables
- Determine if a linear model is a good fit
  - If you see a non-linear relationship between x and y, you can transform your variables (e.g. log transform x) so that linear regression is more appropriate. If you don't do this and just go ahead and fit a linear model, you'll get biased coefficients. Can you think of why (hint: it's one of the 4 types of biases that you learned about)?
- Determine if there are outliers that you need to filter out before doing the regression

So, apart from allowing you to create pretty graphs, data visualization is a really useful tool (we'll talk more about this later).

Let's create a scatterplot to look at the relationship between `sales_hotel` and `ind_lang`.

```r
ggplot(data = mexico, aes(x = ind_lang, y = sales_hotel)) + geom_point(color = "orchid4") +
    labs(title = "Hotel Sales vs. Share of the Population Who Speak the Indigenous Language",
        x = "Share of the Population Who Speak the Native Language",
        y = "Hotel Sales (in Pesos)") + theme(plot.title = element_text(hjust = 0.5))
```



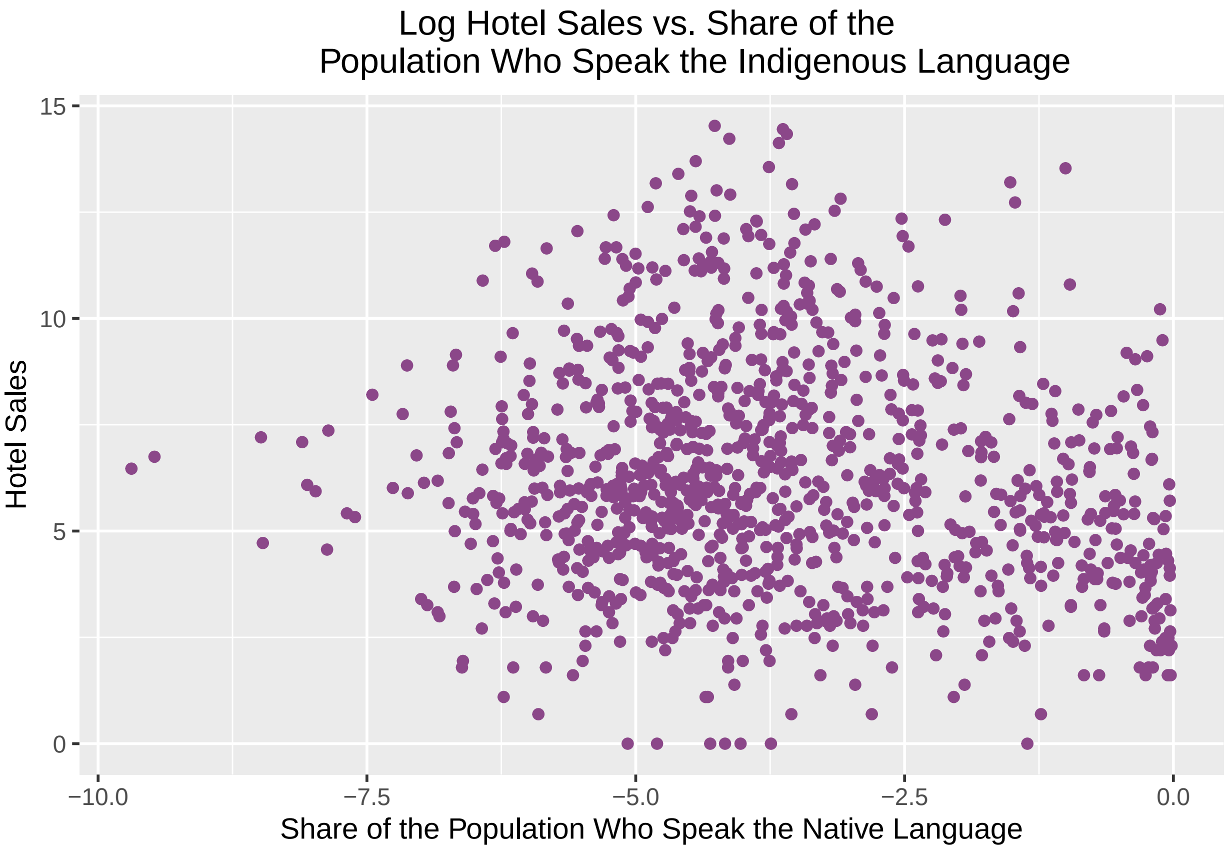
This scatterplot doesn't really show the association between the variables. There's that one outlier above and there isn't much variation in `sales_hotel`. Let's get rid of the outlier and take the log of both x and y.

Note: In reality, you may have to do quite a bit of trial and error to figure out which kind of transformation to use. To save time, we did all that behind the scenes.

```r
# remove the outlier
mexico <- filter(mexico, sales_hotel < (5e+06))

# log transform both variables
mexico$log_sales_hotel <- log(mexico$sales_hotel)
mexico$log_ind_lang <- log(mexico$ind_lang)
mexico <- filter(mexico, log_ind_lang != -Inf)

# create another scatterplot
ggplot(data = mexico, aes(x = log_ind_lang, y = log_sales_hotel)) +
    geom_point(color = "orchid4") + labs(title = "Log Hotel Sales vs. Share of the
         Population Who Speak the Indigenous Language",
    x = "Share of the Population Who Speak the Native Language",
    y = "Hotel Sales") + theme(plot.title = element_text(hjust = 0.5))
```



This scatterplot looks much better! There is a lot more variation in the data which makes the association between the variables much clearer. We can see that there is a somewhat negative relationship between the log transformed variables.

Question: Does the scatter plot align with what you expected to see? Again, we're not looking for a specific answer, we just want you to speculate!

Solution: We expected a negative relationship which is what we see in the scatterplot.

Ordinary Least Squares Regression

Now that we've done some necessary EDA, we can move ahead with our regression. The first type of regression we'll be looking at is ordinary least squares, or OLS, regression Recall that in OLS regression, we want to find the values of our coefficients (i.e. the intercept and slope coefficients) that minimizes the squared error. OLS regression can further be broken down into 2 types:

- Simple linear regression: the model only has one independent variable
- Multiple linear regression: the model has multiple independent variables

Simple Linear Regression

Recall that this is our simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

This is how we can perform simple linear regression in R:

```
model <- lm(y ~ x, data = df)
```

This line of code will estimate the coefficients for your data. This is called fitting the model. To see your results, write:

```
summary(model)
```

Practice: Regress `log_sales_hotel` on `log_ind_lang`. We’ve provided the skeleton code for you below.

```r
# fit a simple linear regression model
model_simple <- lm(log_sales_hotel ~ log_ind_lang, data = mexico)

# let's see our results
summary(model_simple)
```

```
##
## Call:
## lm(formula = log_sales_hotel ~ log_ind_lang, data = mexico)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5841 -1.7967 -0.3474  1.6114  8.1707
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.51240    0.18139  30.389  < 2e-16 ***
## log_ind_lang -0.21117    0.04425  -4.772 2.06e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.661 on 1104 degrees of freedom
## Multiple R-squared:  0.02021,    Adjusted R-squared:  0.01933
## F-statistic: 22.78 on 1 and 1104 DF,  p-value: 2.064e-06
```

When you display your results, you’ll get a table with a lot of information - but you just want to focus on a few things:

- Intercept
- Slope coefficients
- Hypothesis testing information (t-statistics and p-values)
- R^2

As you might have learned in the class, R^2 is primarily used for comparing across models rather than for evaluating the quality of a given model.

Question: Identify each of the 4 values we talked about in the regression table above. Note that in this model, we’re taking the log of both x and y: how would you interpret the coefficient

Soltuion:

- Intercept: 5.512
- Coefficient on `log_ind_lang`: -0.211 (21% decrease in `log_sales_hotel` per 1% increase in `log_ind_lang`)
- Hypothesis testing: The coefficient is statistically significant at the 5% level.
- R^2: 0.0202

Multiple Linear Regression

We don’t generally use a simple linear regression model as it gives us biased coefficients. Instead, we incorporate multiple independent variables and fit a multiple linear regression model.

Recall that this is our multiple linear regression model (with n independent variables):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_n X_{ni} + \epsilon_i$$

Question: What are some biases affecting simple linear regression coefficients?

Solution: Omitted variable bias is the main issue as we’re not accounting for confounding factors in simple linear regression.

Fitting a multiple linear regression model in R is very similar to fitting a simple linear regression model, you just have to include additional independent variables:

```
model <- lm(y ~ x_1 + x_2, data = df)
```

Displaying the results of your model is the same as before, you have to use `summary()`.

Practice: Look at the data description and choose at least 2 other independent variables that could make the coefficients from your previous regression less biased. Run a multiple linear regression model of `log_sales_hotel` against these variables (don’t forget to include our independent variable of interest - `log_ind_lang`).

```r
# fit a multiple linear regression model
model_multiple <- lm(log_sales_hotel ~ log_ind_lang + inc_m +
    educ_years, data = mexico)

# let's see our results
summary(model_multiple)
```

```
##
## Call:
## lm(formula = log_sales_hotel ~ log_ind_lang + inc_m + educ_years,
##     data = mexico)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3562 -1.4841 -0.0317  1.5588  7.1536
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.277e+00  3.554e-01  -3.591 0.000343 ***
## log_ind_lang  4.321e-02  3.935e-02   1.098 0.272398
## inc_m         7.601e-05  6.056e-05   1.255 0.209735
## educ_years    1.071e+00  5.694e-02  18.815  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.237 on 1102 degrees of freedom
## Multiple R-squared:  0.309,  Adjusted R-squared:  0.3071
## F-statistic: 164.3 on 3 and 1102 DF,  p-value: < 2.2e-16
```

Question: Have a look at the coefficients of your regression. Some interesting questions to think about: are the coefficients statistically significant? Has the R^2 improved? How did adding this variable/variables change the coefficients from the previous 2 regressions?

Solution: Only the coefficient on `educ_years` is statistically significant. The coefficient on `log_ind_lang` is no longer statistically significant. The R^2 improved significantly.

IV Regression

We might still be concerned about bias (e.g. omitted variable bias) after running a multiple linear regression model. In this case, we can use instrumental variables regression to uncover the causal effect of the independent variable (the endogenous variable) on the dependent variable.

Note: As you may remember from class, an endogenous variable is an independent variable that is correlated with the error term. In this case, we have an endogeneity problem and our coefficients are biased.

Question: Suppose you hypothesize that Z is a valid instrument. What are the conditions for a valid instrument? How would you test each of these conditions (if applicable)?

Solution: The instrument must be relevant (correlated with the endogenous variable) and exogenous (uncorrelated with the error term).

Instrumental variables regression is also called two-stage least sqaures (TSLS), because it is performed in 2 stages:

1. Regress X on Z, a valid instrument

$$X_i = \alpha_0 + \alpha_1 Z_i + \nu_i$$

2. Regress Y on the fitted values of X, which you get from stage 1 of TSLS

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \epsilon_i$$

In R, this is how you can perform instrumental variables regression:

`model <- ivreg(y ~ x | z, data = df)`

Note that this code does both stages of TSLS and gives you the results of the second stage. You can also perform TSLS manually, by running separate OLS models, but you have to be careful about standard errors (not recommended).

Practice: Choose a variable you think would be a valid instrument. Test the first assumption and use your intuition/understanding of economic theory to think about why the second condition is satisfied.

Solution: `inc_m` might be good instrument as we can expect it to be correlated with `log_ind_lang` and doesn't affect `sales_hotel` directly.

```r
# test the first condition
stage_1 <- lm(log_ind_lang ~ inc_m, data = mexico)

# let's see our results
summary(stage_1)
```

```
##
## Call:
## lm(formula = log_ind_lang ~ inc_m, data = mexico)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5282 -1.2348 -0.1321  1.2395  6.2680
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.724e+00  1.186e-01 -22.969   <2e-16 ***
## inc_m       -3.704e-04  4.126e-05  -8.977   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.747 on 1104 degrees of freedom
## Multiple R-squared:  0.06803,    Adjusted R-squared:  0.06719
## F-statistic: 80.59 on 1 and 1104 DF,  p-value: < 2.2e-16
```

Question: Does your instrument satisfy the first condition? [Hint: Go through lecture 13 slides.]

Solution: The F-statistic is far above 10 so the instrument does satisfy the instrument relevance condition.

Practice: Run an instrumental variables regression model with `log_sales_hotel`, `log_ind_lang` and your chosen instrument (go ahead with this even if your instrument wasn’t statistically significant).

```r
# fit your IV regression model
model_iv_1 <- ivreg(log_sales_hotel ~ log_ind_lang | inc_m, data = mexico)

# let's see our results
summary(model_iv_1)
```

```
##
## Call:
## ivreg(formula = log_sales_hotel ~ log_ind_lang | inc_m, data = mexico)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.6339430 -2.7182169 -0.0005822  2.7728805 11.6097869
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.2839     0.8751   0.324    0.746
## log_ind_lang  -1.6324     0.2359  -6.918 7.72e-12 ***
##
## Diagnostic tests:
##                    df1  df2 statistic p-value
## Weak instruments     1 1104     80.59  <2e-16 ***
## Wu-Hausman           1 1103     80.75  <2e-16 ***
## Sargan               0   NA        NA      NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.701 on 1104 degrees of freedom
## Multiple R-Squared: -0.8953, Adjusted R-squared: -0.8971
## Wald test: 47.87 on 1 and 1104 DF,  p-value: 7.721e-12
```

Exogenous Variables

You might want to include control variables in your instrumental variables regression model. These variables are also called exogenous variables as they are not correlated with the error term. Performing IV regression with control variables is similar to what we did before, with slight differences.

In this case, the following would be our TSLS regression:

1. Regress X on Z and the control variables (Ws):

$$X_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 W_{1i} + ... + \alpha_n W_{ni} + \nu_i$$

2. Regress Y on the estimated X and control variables:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_{1i} + ... + \beta_n W_{ni} + \epsilon_i$$

In R, we would run the following regression:

`model <- ivreg(y ~ x + w1 + w2 | w1 + w2 + z1)`

Practice: Choose 2 variables you think would be exogenous in the model. Run the previous regression but with control variables.

```r
# fitting our IV regression model
model_iv_2 <- ivreg(log_sales_hotel ~ log_ind_lang + logtemp +
    logprecip | logtemp + logprecip + inc_m, data = mexico)

# displaying our results
summary(model_iv_2)
```

```
##
## Call:
## ivreg(formula = log_sales_hotel ~ log_ind_lang + logtemp + logprecip |
##     logtemp + logprecip + inc_m, data = mexico)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -15.82230  -3.78326  -0.06842   3.98321  13.47160
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -25.7830     7.9286  -3.252  0.00118 **
## log_ind_lang  -2.8217     0.6801  -4.149 3.59e-05 ***
## logtemp        1.8859     0.7768   2.428  0.01535 *
## logprecip      2.7141     0.9714   2.794  0.00530 **
##
## Diagnostic tests:
##                    df1  df2 statistic  p-value
## Weak instruments    1 1102     21.82 3.35e-06 ***
## Wu-Hausman          1 1101     66.59 9.05e-16 ***
## Sargan              0   NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.152 on 1102 degrees of freedom
## Multiple R-Squared: -2.666,  Adjusted R-squared: -2.676
## Wald test: 10.81 on 3 and 1102 DF,  p-value: 5.327e-07
```

Question: Based on the results above, do you think the exclusion restriction holds for your instrumental variable of choice? [Hint: Go through lecture 14 slides.]

Solution: It seems that the exclusion restriction doesn't hold as the coefficient on `log_ind_lang` has changed significantly.

Conclusion

This brings us to the end of introduction to regression in R! We went through the main steps of performing linear regression in R, from data visualization and transformation to performing OLS and IV regressions. We hope this helps you in Econ 140 and beyond.