Introduction to Multivariate Regression & Econometrics

HED 612

Lecture 5

1. Prep

2. Prediction

3. Break [10 min]

4. Measures of Model Fit: R^2

5. Measures of Model Fit: Standard Error of the Regression

Prep

Open R Script; Install new Package!

We'll be using the California Schools Data today; no need to re-download!

1. Download the Lecture 5 PDF and R files for this week
   - Place all files in HED612_SP21 »> lectures »> lecture5
2. Open the RProject (should be in your main HED612_SP21 folder)
3. Once the RStudio window opens, open the Lecture 5 R script by clicking on:
   - file »> open file... »> [navigate to lecture 5 folder] »> lecture5.R
4. **Install the "Margins" package in Lecture5.R**
   - If R prompts you to install margins and "its dependencies", go ahead and click install
   - If R doesn't prompt you, install via line 5 `install.packages("margins")`

Homework Questions

- Problem Set #4
  - Questions or Concerns?
- General R Issues?
  - Be careful of $<$ / $>$ and $<=$ / $=>$

Class Overview

Last week:

- Introduced population linear regression model
- Estimating β_0 and β_1 via OLS Estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

Today:

- Prediction
- Measures of Model Fit
    - R^2
    - Standard Error of the Regression (SER)

Homework:

- HW#5 posted on D2L
- Stock and Watson Chapter 5

Next week:

- Using $\hat{\beta}_1$ for hypothesis testing about β_1

Prediction

General things we do in regression analysis

1. **Estimation** [Last Week]

- How do we choose estimates of β_0 and β_1 using sample data?
- OLS Estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared errors

2. **Prediction** [Today]

- What is the predicted value of Y for someone with a particular value of X?
- Example: What is the predicted income for a college graduate with an engineering degree?

3. **Hypothesis testing** [focus of the rest of the semester]

- Hypothesis testing and confidence intervals about β_1
- This is estimating causal effects!

General steps in prediction

1. Write out population linear regression model
   - $Y_i = \beta_0 + \beta_1 X_i + u_i$
   - Where Y= Annual Income; X= hours worked per week
2. Write out OLS prediction line without estimate values
   - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
3. Write out OLS prediction line with estimates [last week's example]
   - $\hat{Y}_i = \$7,477 + \$503 * X_i$
4. If it's helpful, sketch out the OLS prediction line
5. Calculate predicted values of Y ($\hat{Y}_i$) for selected values of X_i

Common Examples of Prediction

Early warning system for academic probation

- Using data from last year, run regression for relationship between number of class absences and semester GPA
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Where Y= GPA; X= number of absences
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - $\hat{Y}_i = 3.5 + (-0.5) * X_i$
- If a student has 4 class absences?
  - $\hat{Y}_i = 3.5 + (-0.5) * 4 = \hat{Y}_i = 3.5 + (-2) =$ 1.5
- Track attendance of current year students
  - Student receives academic probation warning if predicted to have a GPA less than 2.0

Common Examples of Prediction

Obama Administration College Rating System

- Rates colleges based on student outcomes to predict the "best value"
  - graduation rates
  - labor market participation and return rates
  - loan repayment/default rates
  - attendance cost
- Penalties for colleges with low performance
  - Less maximimum pell grants
  - Shared "risk" for loan default
  - Lose Title IV funding altogether
- Any red flags with *only* using the above student outcomes to rate colleges?
  - Different student populations: UA students vs Harvard students?
  - Different labor market participation: Liberal arts college vs technical school?
  - Obama Administration opted out of ratings system; opted instead for providing students with more information via College Score Cards for them to compare institutions themselves
  - Some accountability kept via gainful employment regulations for for-profits

Prediction Example: California School Data

RQ: What is the relationship between district-wide math test scores (Y) and district expenditures per student (X)?

- Do you think this relationship is positive? Negative? Why?

- Practical problem: our X is in dollars.
  - $\hat{\beta}_1$ will be the change in Y for a one-unit change ($1 change) in X
  - District wide test scores are not likely to change much for a $1 increase in per student spending
- Let's make our X in thousands of dollars
  - $\hat{\beta}_1$ will be the change in Y for a one-unit change ($1,000 change) in X

1. Write out the population regression model
2. OLS line without estimates
3. OLS line with estimates
   - Run regression in R

   -
     ```
     mod1 <-  lm(math_scr ~ expn_stu_000, data=caschool)
     ```

   -
     ```
     summary(mod1)
     ```
4. Interpret the values of $\hat{\beta}_0$ and $\hat{\beta}_1$
5. Draw OLS regression line on graph
6. Calculate predicted district-wide math test score for X=5 ($5,000 expenditures per student)

Student Exercise #1

RQ: What is the relationship between district-wide reading test scores (Y) and district expenditures per student (X)?

- Y= `read_scr`
- X= `expn_stu_000`

1. Write out the population regression model
2. OLS line without estimates
3. OLS line with estimates
    - Run regression in R
    - Assign this regression as `mod2`
4. Interpret the values of $\hat{\beta}_0$ and $\hat{\beta}_1$
5. Draw OLS regression line on graph (if it is helpful to you)
6. Calculate predicted district-wide reading score for X=11 ($11,000 expenditures per student)

`margins` and `prediction` packages in R

- These packages do a lot of useful things; we'll learn how to use functionality little by little

- For today:
  - general introduction
  - use them to calculate predicted values

- We need to run a regression and save it as an "object"
- Then we can run `margins` and `prediction` functions on our regression model!

Break [10 min]

Measures of Model Fit: R^2

R^2

What does R^2 measure?

- R^2, “the coefficient of determination”, measures the fraction of the variance in Y that is explained by X (and is not already explained by sample mean, $\bar{Y}$)
  - In other words, how much better is the regression model in predicting values of Y than $\bar{Y}$
  - Draw scatterplot

Interpreting R^2

- R^2 ranges from 0 to 1
  - $R^2 = 0$: the model does not explain any variation in Y
  - $R^2 = 1$: the model explains 100% of the variation in Y
  - $R^2 = 0.135$: the model explains 13.5% of the variation in Y

Components of R^2

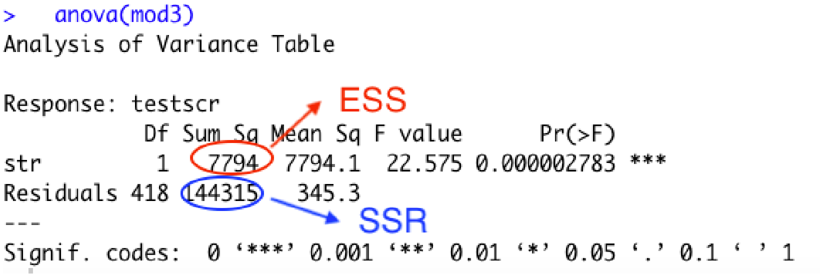
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- $\hat{u}_i = Y_i - \hat{Y}_i$

- Total Sum of Squares (TSS) $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
  - A measure of the total variance in Y, relative to $\bar{Y}$
- Explained Sum of Squares (ESS) $= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$
  - Difference between predicted values $\hat{Y}_i$ and sample mean $\bar{Y}$
  - Measures amount of variation in Y explained by X
- Sum of Squared Residuals (SSR) $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
  - Difference between actual value observed and predicted value by our regression line
  - Measures amount of variation in Y not explained by X

- Show in scatter plot: What is the effect of student teacher ratio (X) on district test scores?
  - TSS = ESS + SSR
  - Total variation in Y (TSS) equals variation explained by the model (ESS) plus variation not explained by model (SSR)

R^2 in R

- $R^2 = \frac{\textbf{variance in Y that is explained by X}}{\textbf{total variance in Y}} = \frac{ESS}{TSS}$
- Note that in R:
    - `summary(MODEL)` will only give you the R^2
    - `anova(MODEL)` will give you ESS and SSR

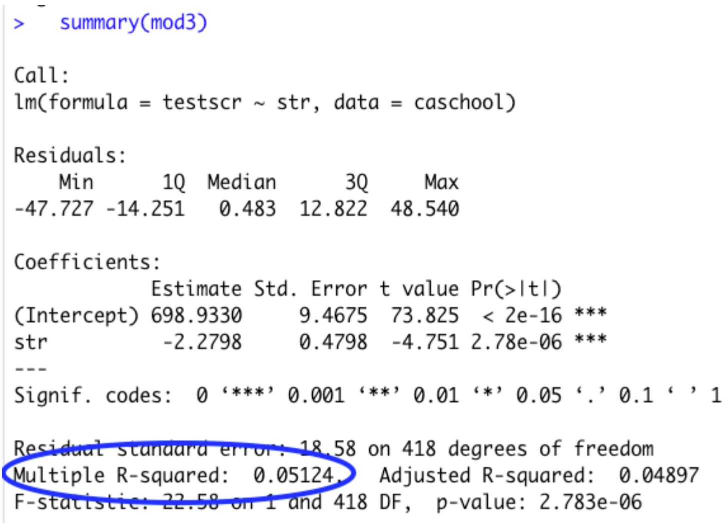


```
>   anova(mod3)
Analysis of Variance Table

Response: testscr
           Df Sum Sq Mean Sq F value      Pr(>F)    
str         1   7794  7794.1  22.575 0.000002783 ***
Residuals 418 144315   345.3                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- `mod3 <- lm(testscr ~ str, data=caschool)`
- `anova(mod3)`
    - ESS= 7794
    - SSR = 144315
    - TSS = 7794 + 144315 =152109
    - R^2 = 7794/152109 = 0.051
    - Student teacher ratio explains 5.1% of the variation in district test scores (not already explained by $\bar{Y}$)

R^2 in R via `summary(MODEL)`



```
>   summary(mod3)

Call:
lm(formula = testscr ~ str, data = caschool)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.727 -14.251   0.483  12.822  48.540 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
str          -2.2798     0.4798  -4.751 2.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.58 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
```

Student Exercise #2

RQ1: What is the relationship between district-wide math test scores (Y) and district expenditures per student (X)?

- Y= `math_scr`
- X= `expn_stu_000`

1. Run regression in R (did this already in exercise #1; `mod1`)
2. Using `summary(mod1)` :
   - What is R^2? Interpret R^2.
3. Using `anova(mod1)` :
   - Show that $R^2 =$ ESS/TSS

RQ2: What is the relationship between district-wide math test scores (Y) and student-teacher ratio (X)?

- Y= `math_scr`
- X= `str`

1. Run regression in R (need to do this as `mod4`)
2. Using `summary(mod4)` :
   - What is R^2? Interpret R^2.
3. Using `anova(mod4)` :
   - Show that $R^2 =$ ESS/TSS

Student Exercise #2 [SOLUTIONS]

RQ1: What is the relationship between district-wide math test scores (Y) and district expenditures per student (X)?

1. Run regression in R (did this already in exercise #1; `mod1`)
2. Using `summary(mod1)` :
   - What is R^2? *Measures the fraction of the variance in Y that is explained by X (and is not already explained by sample mean)*
   - R^2 = 0.02402; *District expenditures per student explains 2.4% of the variation in district math test scores*
3. Using `anova(mod1)` :
   - ESS= 3540
   - SSR = 143831
   - TSS = 3540 + 143831 =147371
   - R^2 = 3540/147371 = 0.024

RQ2: What is the relationship between district-wide math test scores (Y) and student-teacher ratio (X)?

1. Run regression in R (need to do this as `mod4`)
2. Using `summary(mod4)` :
   - Interpret R^2 = 0.03824; *District student-teacher ratio explains 3.8% of the variation in district math test scores*
3. Using `anova(mod4)` :
   - ESS= 5636
   - SSR = 141735
   - TSS = 5636 + 141735 =147371
   - R^2 = 5636/147371 = 0.03824

Student Exercise #2

Based on the R^2 of Model 1 vs R^2 of Model 4:

- Does district expenditures per student or student teacher ratio do a better job at predicting math test scores?

Measures of Model Fit: Standard Error of the Regression

Standard Error of the Regression

Standard Error of the Regression (SER): an estimator of the standard deviation of the residuals $\hat{u_i}$

- In other words, SER is, on average, how far away an actual observed value of Y is from the predicted value of Y for a random observation

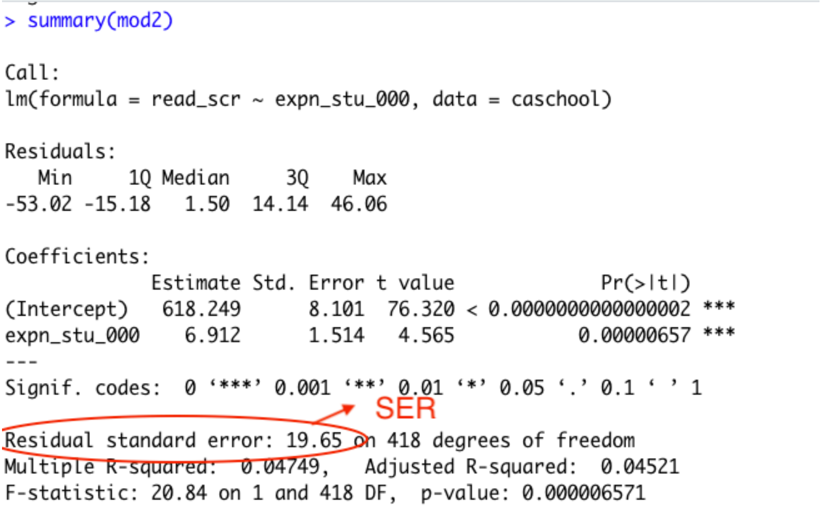
- Sample standard deviation, $s_Y = \hat{\sigma_Y}$
  - Average distance between a random observation Y_i an the sample mean $\bar{Y}$
  - $s_Y = \hat{\sigma_Y} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}$

- Standard error of the regression (SER or $s_{\hat{u}}$ or $\hat{\sigma_{u_i}}$)
  - Average distance between a random observation Y_i and the value predicted by the OLS regression $\hat{Y_i}$
  - SER $= \sqrt{\frac{SSR}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{u_i})^2}{n-2}}$
  - SSR $=$ sum of squared residuals
  - Why (n-2)? Because we lose one degree of freedom for calculating sample mean $\bar{Y}$ and another for the independent variable X

Standard Error of the Regression (SER)

RQ: What is the relationship between district-wide reading test scores (Y) and district expenditures per student (X)?

- In R, SER is “Residual Standard Error”



```
> summary(mod2)

Call:
lm(formula = read_scr ~ expn_stu_000, data = caschool)

Residuals:
   Min     1Q Median     3Q    Max 
-53.02 -15.18   1.50  14.14  46.06 

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)    
(Intercept)   618.249      8.101  76.320 < 0.0000000000000002 ***
expn_stu_000    6.912      1.514   4.565           0.00000657 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.65 on 418 degrees of freedom
Multiple R-squared:  0.04749,	Adjusted R-squared:  0.04521 
F-statistic: 20.84 on 1 and 418 DF,  p-value: 0.000006571
```

Calculating Standard Error of the Regression (SER)

RQ: What is the relationship between district-wide math test scores (Y) and district expenditures per student (X)?

- SER $= \sqrt{\frac{SSR}{n-2}}$
- `anova(mod1)`
  - SSR= 161369
  - n = 420 (there's 420 observations/districts in the California School Data)

$$\text{SER} = \sqrt{\frac{161369}{420-2}} = \sqrt{\frac{161369}{428}} = \sqrt{377.0303} = 19.648$$

Standard Error of the Regression (SER) cont.

SER $= 19.648$

- Average distance between reading test score of a random district and the reading test score predicted by the OLS regression is 19.65
- SER is always in the same unit as the dependent variable
- A “higher” SER means that our predicted OLS line is often wrong by a lot!

SER vs Standard Deviation

- Standard Deviation of `read_scr` $= 20.1$
  - Average distance between reading test score of a random district and the sample mean of reading test scores is 20.1

Does predicting reading test scores based on district expenditures per student result in dramatic improvements in predictions in comparison to just taking the sample mean of math test scores?

Comments on SER and R^2

R^2

- Low R^2 tell us that the model does not explain much of the variation in the dependent variable; there are other factors (not included in the model) that explain most of the variation in the dependent variable
- Low R^2 is not necessarily bad and a high R^2 is not necessarily good (e.g., predicting provery from having a cell phone has a very high R^2)

SER

- High SER tells us that our predictions will often be wrong by a lot; but that does not necessarily mean our model is bad given some things are hard to predict and we often are interested in "averages"!

Econometrics Research & Model Fit

- Usually we're most concerned with estimating β_1, which is the "causal effect" of X on Y
- Models with high R^2 and low SER may still be wrong about β_1

Student Exercise #3

RQ1: What is the relationship between district-wide math test scores (Y) and district expenditures per student (X)?

- Y= `math_scr`
- X= `expn_stu_000`

1. Run regression in R (did this already in exercise #1 & 2; `mod1`)
2. Using `summary(mod1)` :
   - What is SER? Interpret SER.
3. Using `anova(mod1)` :
   - Show that SER $= \sqrt{\frac{SSR}{n-2}}$

RQ2: What is the relationship between district-wide math test scores (Y) and student-teacher ratio (X)?

- Y= `math_scr`
- X= `str`

1. Run regression in R (already did this; `mod4`)
2. Using `summary(mod4)` :
   - Interpret SER.
3. Using `anova(mod4)` :
   - Show that SER $= \sqrt{\frac{SSR}{n-2}}$

Student Exercise #3 [SOLUTIONS]

RQ1: What is the relationship between district-wide math test scores (Y) and district expenditures per student (X)?

2. Using `summary(mod1)` :
   - What is SER? **SER is on average how far away an actual observed value of Y is from the predicted value of Y for a random observation**
   - Interpret SER. **18.55; Average distance between math test score of a random district and the math test score predicted by the OLS regression is 18.55**
3. Using `anova(mod1)` :
   - Show that SER $= \sqrt{\frac{SSR}{n-2}} = \sqrt{\frac{143831}{418}} = \sqrt{344.0933}$ =18.549

RQ2: What is the relationship between district-wide math test scores (Y) and student-teacher ratio (X)?

2. Using `summary(mod4)` :
   - Interpret SER **18.41; Average distance between math test score of a random district and the math test score predicted by the OLS regression is 18.41**
3. Using `anova(mod4)` :
   - Show that SER $= \sqrt{\frac{SSR}{n-2}} = \sqrt{\frac{141735}{418}} = \sqrt{339.0789} =$ 18.414