ECON 340
Economics Research Methods

Div Bhagia

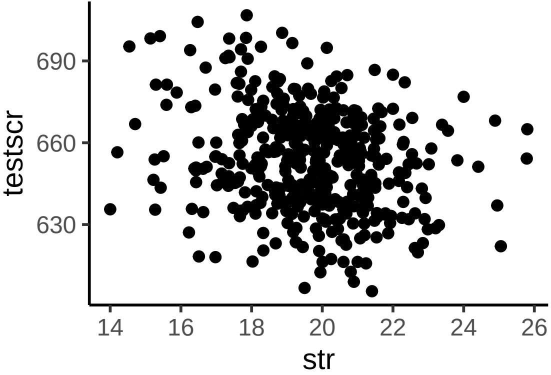
Lecture 21: Regression Analysis in R

Housekeeping

```r
rm(list=ls())
library(tidyverse)
library(stargazer)
#setwd("~/Dropbox (CSU Fullerton)/Econ340_R")
data <- read.csv("caschool.csv")
```

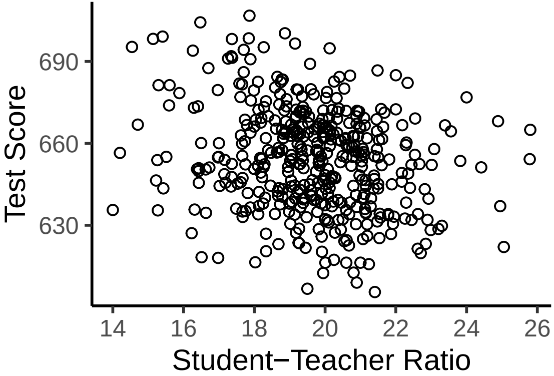
Scatterplot

```
ggplot(data, aes(x=str, y=testscr)) +
  geom_point() +
  theme_classic()
```



Scatterplot

```
ggplot(data, aes(x=str, y=testscr)) +
  geom_point(shape=1) + theme_classic() +
  labs(x="Student-Teacher Ratio", y="Test Score")
```



Linear Regression

- `lm()` is a function used to fit linear regression models
- Syntax: `lm(y ~ x1 + x2 + ... , data = mydata)`
- Useful to store it as an object

```
model <- lm(testscr ~ str, data)
```

- Apply `summary()` function to the stored result from output

Regression Output

```r
summary(model)
```

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

Regression Output

- Fitted model:

$$\widehat{testscr} = 698.93 - 2.28 \cdot str$$

- $R^2 = 0.05$ implies that 5% of variation in test scores explained by student teacher ratio
- Standard errors (deviations):

$$SE_{\hat{\beta}_0} = 9.47, \quad SE_{\hat{\beta}_1} = 0.48$$

Regression Output

- Often interested in testing the hypothesis: $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$
- Corresponding t-value:

$$t_0 = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} = \frac{-2.28}{0.48} = -4.75$$

- p-value: $p = 2Pr(Z > t_0)$
- If $p < \alpha$, coefficient significant at α% level of significance

Confidence Intervals

- $(1-\alpha)\%$ confidence interval is given by: $\hat{\beta}_1 \pm z_{\alpha/2} \cdot SE_{\hat{\beta}_1}$
- Note that $z_{0.025} = 1.96$, so the 95% confidence interval:

$$-2.28 \pm 1.96 \cdot 0.48$$

```r
confint(model)
```

```
##                  2.5 %     97.5 %
## (Intercept) 680.32313 717.542779
## str          -3.22298  -1.336637
```

Predicted and Residual Values

```r
data$yhat <- predict(model)
data$uhat <- residuals(model)
```

Should the average of `testscr` and `yhat` be the same?

```r
mean(data$testscr)
mean(data$yhat)
```

What should be the average of `uhat`?

```r
mean(data$uhat)
```

Predicted and Residual Values

What is the predicted value when `str`=21?

```
data %>% select(testscr, str, yhat, uhat) %>%
  filter(str==21)
```

```
##   testscr str     yhat      uhat
## 1   616.3  21 651.057 -34.75699
```

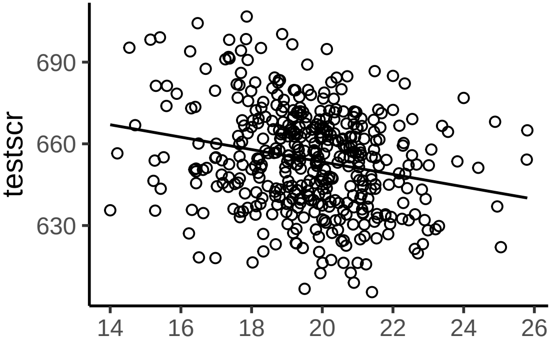
Remember:

$$\widehat{testscr} = 698.93 - 2.28 \cdot str$$

Note that: $\hat{u}_i = Y_i - \hat{Y}_i$

Plotting the Fitted Line

```
ggplot(data, aes(x=str, y=testscr)) +
  geom_point(shape=1) + theme_classic() +
  geom_line(aes(y=yhat))
```



Output using Stargazer

```r
stargazer(model, type="text",
          keep.stat = c("n", "adj.rsq"))
```

	Dependent variable:
	testscr
str	-2.280^{***}
	(0.480)
Constant	698.933^{***}
	(9.467)
Observations	420
Adjusted R^2	0.049

Output from Multiple Models

```
model1 <- lm(math_scr ~ str, data)
model2 <- lm(read_scr ~ str, data)
stargazer(model1, model2, type="text",
          keep.stat = c("n", "adj.rsq"))
```

Output from Multiple Models

	Dependent variable:	
	math_scr	read_scr
	(1)	(2)
str	-1.939^{***}	-2.621^{***}
	(0.476)	(0.504)
Constant	691.417^{***}	706.449^{***}
	(9.382)	(9.941)
Observations	420	420
Adjusted R^2	0.036	0.059
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01	

Multiple Regression Model

```r
model3 <- lm(testscr ~ str + comp_stu, data)
stargazer(model, model3, type="text",
          keep.stat = c("n", "adj.rsq"))
```

- Note: Use the adjusted R^2 to compare two models with different number of variables

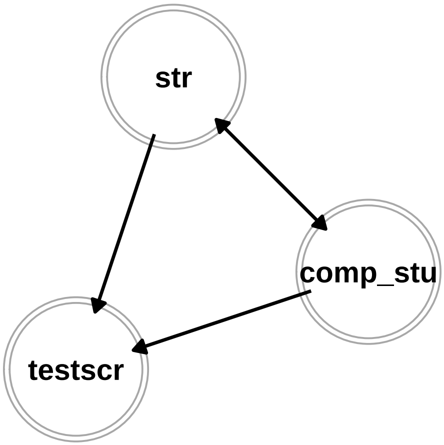
Multiple Regression Model

	Dependent variable:	
	testscr	
	(1)	(2)
str	-2.280^{***}	-1.593^{***}
	(0.480)	(0.493)
comp_stu		65.160^{***}
		(14.351)
Observations	420	420
Adjusted R^2	0.049	0.092
Note:	*p<0.1; **p<0.05; ***p<0.01	

Omitted Variable Bias

- Negative coefficient on `str` smaller in magnitude after controlling for `comp_stu`
- Lower `comp_stu` $\rightarrow$ Lower `testscr`
- Lower `comp_stu` $\leftrightarrow$ Higher `str`
- So `comp_stu` explains some of the relationship between `str` and `testscr`

Omitted Variable Bias



Next Class

- For the next class download and load `acs2019` dataset from the Dropbox folder
- We will continue with linear regression in R
- Come prepared so we can start quickly