# Hypothesis Tests and Confidence Intervals in Multiple Regression (SW Ch. 7)

**Ercan Karadas**

**Econometrics (ECON 3112)**

**Belk College of Business, UNCC**

December 3, 2018

# Outline

Hypothesis Tests and Confidence Intervals for a Single Coefficient

Tests of Joint Hypotheses

Testing Single Restrictions Involving Multiple Coefficients

Model Specification for Multiple Regression

Analysis of the Test Score Data Set

# Hypothesis Tests and Confidence Intervals for a Single Coefficient

- Hypothesis tests and confidence intervals for a single coefficient in multiple regression follow the same logic and recipe as for the slope coefficient in a single-regressor model.
- From CLT, we have this important result

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{\text{var}\left(\widehat{\beta}_i\right)}} \sim N(0,1), \qquad \text{for all } i$$

- Therefore, hypothesis on $\beta_i$ can be tested using the usual $t$-statistic:

$$H_0 : \beta_i = \beta_{i,0} \text{ vs. } H_1 : \beta_i \neq \beta_{i,0}$$

where $\beta_{i,0}$ is the hypothesized value under the null. To conclude this hypothesis, compare the $t$-actual, $t^a = \frac{\widehat{\beta}_i - \beta_{i,0}}{SE\left(\widehat{\beta}_i\right)}$, to the critical $t$.

- And C.I.s can be constructed as before: $\widehat{\beta}_i \pm 1.96 \times SE(\widehat{\beta}_i)$

# Example: Test Score - Class Size data

**(1) Regression Output with a Single Regressor:**

$$\widehat{\text{Test Score}} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR$$

**(2) Regression Output with Two Regressors:**

$$\widehat{\text{Test Score}} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.65} \times PctEL$$

where $PctEL$ is a new regressor denoting the percentage of English learners in a district.

- The coefficient on $STR$ in (2) is the effect on TestScores of a unit change in $STR$, *holding constant the percentage of English Learners in the district*.
- Notice that the coefficient on $STR$ falls by one-half. This is because of the omitted variable bias that we discussed in Chapter 6.

# Example: Test Score - Class Size data (Cont'd)

## (2) Regression Output with Two Regressors:

$$\widehat{\text{Test Score}} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.65} \times PctEL$$

- Let's test the significance of the coefficient of $PctEL$:
  - Hypothesis: $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$
  - Construct the $t$-statistic

$$t^a = \frac{\widehat{\beta}_2 - \beta_{2,0}}{SE(\widehat{\beta}_2)} = \frac{-0.65 - 0}{0.031} = -20.97$$

  - The 5% 2-sided critical value is 1.96, so we reject the null at the 5% significance level.
- The 95% C.I. for the coefficient on $PctEL$ ($\beta_2$) is

$$\{-0.65 \pm 1.96 \times 0.031\} = (-0.71,\ -0.59)$$

# R Codes for Estimating Multiple Regression

```
#Prepare variables
library("AER")
data(CASchools)
TestScore <- (CASchools$read + CASchools$math)/2
STR <- CASchools$students/CASchools$teachers
PctEL <- CASchools$english

# Estimate multiple regression and save in myReg:
myReg <- lm(TestScore ~ STR + PctEL)

# An alternative way to estimate multiple regression on R:
myReg <- lm(TestScore ~ STR + english, data = CASchools)
```

# R Output (Homoskedasticity-only)

```
> summary(myReg)  # this command produces the output below

Call:
lm(formula = TestScore ~ STR + PctEL)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 686.03224    7.41131  92.566  < 2e-16 ***
STR          -1.10130    0.38028  -2.896  0.00398 **
PctEL        -0.64978    0.03934 -16.516  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

# R Output (Heteroskedasticity-robust)

```
# To produce heteroskedasticity-robust output:
> library("car")
> S(myReg, vcov. = hccm) # this replaces "summary" function

Call: lm(formula = TestScore ~ STR + PctEL)
Standard errors computed by hccm

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 686.0322     8.8122   77.85   <2e-16 ***
STR          -1.1013     0.4371   -2.52   0.0121 *
PctEL        -0.6498     0.0313  -20.76   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05

Residual standard deviation: 14.46 on 417 degrees of freedom
Multiple R-squared: 0.4264
F-statistic: 220.1 on 2 and 417 DF,  p-value: < 2.2e-16
```

# Tests of Joint Hypotheses

- Sometimes the null hypothesis puts restriction on more than one coefficients jointly.
- To see an example of this, let's add the variable $Expn$, which measure the expenditures per pupil, to the population regression

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

- And suppose that we are interested in testing the hypothesis that "school resources don't matter" vs. "they do".
- The null and alternative hypothesis can be written formally as:

$$H_0 : \beta_1 = 0 \textbf{ and } \beta_2 = 0$$
$$H_1 : \textbf{either } \beta_1 \neq 0 \textbf{ or } \beta_2 \neq 0 \textbf{ or both}$$

- An equivalent way of expressing the same test:

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \text{At least one of the coefficients is nonzero}$$

- A **joint hypothesis** specifies a value for two or more coefficients, that is, it **imposes** a restriction on two or more coefficients.
- In general, a joint hypothesis will involve $q$ restrictions. In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.
- A "common sense" idea is to reject $H_0$ if either of the individual $t$-statistics exceeds 1.96 in absolute value.
- But this "one at a time" test is NOT valid
- Because the resulting test rejects $H_0$ too often under the null hypothesis (more than 5%)!
- In another word, even if you reject $\beta_1 = 0$ and/or $\beta_2 = 0$ at 5% you will end up rejecting $\beta_1 = \beta_2 = 0$ more than 5% of the time.
- Let's see why that is the case...

# Why can't we just test the coefficients one at a time?

- We will calculate the probability of incorrectly rejecting the null, which is $\alpha$, using the "common sense" test based on the two individual $t$-statistics.
- To simplify the calculations, suppose $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are independently distributed and let $t_1$ and $t_2$ be the $t$-statistics:

$$t_1 = \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)} \text{ and } t_2 = \frac{\widehat{\beta}_2 - 0}{SE(\widehat{\beta}_2)}$$

- The 'one at a time' test is:

$$\text{Reject } H_0 : \beta_1 = \beta_2 = 0 \text{ if } |t_1| > 1.96 \text{ and/or } |t_2| > 1.96$$

- What is the probability that this 'one at a time' test rejects $H_0$, when $H_0$ is actually true? (It should be 5%.)

# The 'one at a time' test rejects $H_0$ too often

- The probability of incorrectly rejecting the null hypothesis using the 'one at a time' test is

$$
\begin{aligned}
&= P\left[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96\right] \\
&= 1 - P\left[|t_1| \leq 1.96 \text{ and } |t_2| \leq 1.96\right] \\
&= 1 - P\left[|t_1| \leq 1.96\right] \times P\left[|t_2| \leq 1.96\right] \\
&= 1 - (0.95)^2 \\
&= 0.0975 = 9.75\%
\end{aligned}
$$

  which is not the desired 5%.

- This result implies that the 'one at a time' test rejects the null hypothesis far too often!

- The size of a test is the actual rejection rate under the null hypothesis.
  - The size of the "one at a time" test is NOT 5%!!
  - In fact, its size depends on the correlation between $t_1$ and $t_1$ (and thus on the correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$).
- There are two solutions:
  1) Use a different critical value in this procedure - not 1.96 (this is the "Bonferroni method" - see SW App. 7.1). This method is rarely used in practice, so we won't cover it in this course.
  2) Use a different test statistic designed to test both $\beta_1$ and $\beta_2$ at once: the $F$-statistic. This is common practice.
- The distribution of $F$-statistic depends on the assumptions on the error terms $u_i$

Case 1. $u_i \sim N(0, \sigma_u^2)$ and $i.i.d.$ (normal + homoskedastic + i.i.d.)

Case 2. $u_i \sim (0, \sigma_u^2)$ and $n$ is large (homoskedastic + large sample)

Case 3. $n$ is large (large sample)

# Case 1: F-statistic when $u_i \sim N(0, \sigma_u^2)$ and $i.i.d.$

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

- **Test of interest:** "school resources don't matter"

  $H_0 : \beta_1 = \beta_2 = 0$

  $H_1$ : At least one of the coefficients is nonzero

- **Run two regressions**
  - **Restricted Regression Model:**

    $$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i, \quad R_r^2$$

  - **Unrestricted Regression Model:**

    $$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i, \quad R_u^2$$

- **F-test:**

  $$F_{h.o.n} = \frac{\left(R_u^2 - R_r^2\right)/q}{\left(1 - R_u^2\right)/(n-k-1)} \sim F_{q,n-k-1}$$

  $q$= # of restrictions under $H_0$; $k$ = # of regressors in unrest. model

$$F_{h.o.n} = \frac{\left(R_u^2 - R_r^2\right)/q}{\left(1 - R_u^2\right)/(n-k-1)} \sim F_{q,n-k-1}$$

- This test statistics is called **homoskedasticity-only $F$ under Normality (h.o.n)**.
- **Basic Idea:** Does relaxing the $q = 2$ restrictions that constitute the null hypothesis improves the fit of the regression by enough that this improvement is unlikely to be the result merely of random sampling variation if the null hypothesis is true.
- **Decision Rule:**

$$F_{h.o.n} > F^{\alpha}_{q,n-k-1} \implies \text{Reject } H_0$$

where we read the critical values $F^{\alpha}_{q,n-k-1}$ from Table 5A-C.

- The bigger the difference between the restricted and unrestricted $R^2$s - the greater the improvement in fit by adding the variables in question - the larger is the $F$, and therefore more likely to reject $H_0$.

# $F^{0.05}_{n_1,n_2}$: Critical Values at $\alpha$ =5% ($n_1 = q$, $n_2 = n - k - 1$)

**TABLE 5B** Critical Values for the $F_{n_1,n_2}$ Distribution—5% Significance Level

| Denominator Degrees of Freedom ($n_2$) | Numerator Degrees of Freedom ($n_1$) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 161.40 | 199.50 | 215.70 | 224.60 | 230.20 | 234.00 | 236.80 | 238.90 | 240.50 | 241.90 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.39 | 19.40 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 |
| 26 | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 |
| 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 | 2.20 |
| 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 | 2.19 |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 |
| 90 | 3.95 | 3.10 | 2.71 | 2.47 | 2.32 | 2.20 | 2.11 | 2.04 | 1.99 | 1.94 |
| 120 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 | 1.91 |
| ∞ | **3.84** | **3.00** | **2.60** | **2.37** | **2.21** | **2.10** | **2.01** | **1.94** | **1.88** | **1.83** |

This table contains the 95th percentile of the distribution $F_{n_1,n_2}$ which serves as the critical values for a test with a 5% significance level.

# Example: Test Score - Class Size data

**(1) Restricted Regression Output:**

$$\widehat{\text{TestScore}} = \underset{(1.0)}{664.7} - \underset{(0.032)}{0.671}\, PctEL, \quad R_r^2 = 0.4149$$

**(2) Unrestricted Regression Output:**

$$\widehat{\text{TestScore}} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29}\, STR + \underset{(1.59)}{3.87}\, Expn - \underset{(0.032)}{0.656}\, PctEL, \quad R_u^2 = 0.4366$$

Therefore,

$$F_{h.o.n} = \frac{\left(R_u^2 - R_r^2\right)/q}{\left(1 - R_u^2\right)/(n-k-1)} = \frac{(0.4366 - 0.4149)/2}{(1-0.4366)/(420-3-1)} = 8.01$$

Suppose we choose $\alpha = 0.05$. Then we read the critical value of $F$ statistics from Table 5B for $F_{2,416}^{0.05}$ as 3.0 (Note that the denominator degrees of freedom is too big therefore we actually take $F_{2,\infty} = 3.0$) Since $8.01 > 3.0$ we reject $H_0$: school resources do matter!

# Case 2: F-statistic when $u_i \sim (0, \sigma_u^2)$ and $n$ is Large

- We are relaxing the Normality assumption. But $u_i$s are still assumed to be homoskedastic and we require large sample
- Then

$$F_{h.o.} = \frac{\left(R_u^2 - R_r^2\right)/q}{\left(1 - R_u^2\right)/(n-k-1)} \sim F_{q,\infty}$$

- This test statistics is called **homoskedasticity-only $F$ (h.o.)**.
- **Decision Rule:**

$$F_{h.o.} > F_{q,\infty}^{\alpha} \implies \text{Reject } H_0$$

- The bigger the difference between the restricted and unrestricted $R^2$s - the greater the improvement in fit by adding the variables in question - the larger is the $F$, and therefore more likely to reject $H_0$.

# Case 3: F-statistic when $n$ is Large

- This is the most general version of $F$ statistics. We don't require neither Normality nor homoskedasticity of $u_i$s, but we still need large samples
- Under these assumptions (for $q = 2$) :

$$F_{ht.r.} = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\widehat{\rho}_{t_1,t_2} t_1 t_2}{1 - \widehat{\rho}_{t_1,t_2}^2}\right) \sim F_{q,\infty}$$

  where $t_1$ and $t_2$ are $t$-statistics we computed before; and $\widehat{\rho}_{t_1,t_2}$ is an estimator of the correlation between the two $t$-statistics.
- This test statistics is called **Heteroskedasticity-robust $F$ (ht.r.)**.
- **Decision Rule:**

$$F_{ht.r.} > F_{q,\infty}^{\alpha} \implies \text{Reject } H_0$$

- For the previous example, we will see that R computes $F_{ht.r.} = 5.26$ which is greater than $F_{2,416}^{0.05} = 3.0$, so we reject $H_0$ with Heteroskedasticity-robust standard errors as well.

# R Codes for Tests of Joint Hypothesis

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

$$H_0 : \beta_1 = \beta_2 = 0, \text{ vs. } H_1 : \beta_i \neq 0, \text{ for at least one } i = 1, 2$$

```
# Prepare Data
STR <- CASchools$students/CASchools$teachers
TestScore <- (CASchools$read + CASchools$math)/2
PctEL <- CASchools$english
Expn <- CASchools$expenditure/1000

# Estimate the regression:
myReg <- lm(TestScore ~ STR + PctEL + Expn)

# Perform the test (homoskedasticity-only):
linearHypothesis(myReg, c("STR=0", "Expn=0"))

# Perform the test (heteroskedasticity-robust):
linearHypothesis(myReg, c("STR=0", "Expn=0"), vcov = hccm)
```

# R Output (homoskedasticity-only)

```
> linearHypothesis(myReg, c("STR=0", "Expn=0"))

Linear hypothesis test

Hypothesis:
STR = 0
Expn = 0

Model 1: restricted model
Model 2: TestScore ~ STR + PctEL + Expn

  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1    418 89000
2    416 85700  2    3300.3 8.0101 0.000386 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
```

# R Output (heteroskedasticity-robust)

```
> linearHypothesis(myReg, c("STR=0", "Expn=0"), vcov = hccm)

Linear hypothesis test

Hypothesis:
STR = 0
Expn = 0

Model 1: restricted model
Model 2: TestScore ~ STR + PctEL + Expn
Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    418
2    416  2 5.2617 0.005537 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
```

# Use of Different $F$-statistics

- In general Heteroskedasticity-robust $F$ and homoskedasticity-only $F$ are different
- Unless you have a good reason (economic theory and/or intuition), you should use Heteroskedasticity-robust $F$ statistics.
- However, since computation of Heteroskedasticity-robust $F$ requires some information (i.e. $\widehat{\rho}_{t_1,t_2}$) that is not reported in default regression output, we will not use this in our examples and problem sets.
- R will give homoskedasticity-only $F$ statistics unless you specifically ask it to compute Heteroskedasticity-robust $F$

# Overall Significance Test

- In default regression output, an $F$-statistics is always included. But it is important to keep in mind that it corresponds to a very particular null-hypothesis: **all slope coefficients are zero**
- For example, the reported $F$-statistics in the default regression output for

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

corresponds to the following test:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_i \neq 0, \text{ for at least one } i = 1, 2, 3$$

- Another important point to keep in mind is that: the reported $F$-statistics in the default regression is homoskedasticity-only $F$. To get heteroskedasticity-robust $F$ use $S(myReg, vcov. = hccm)$.

# Testing Single Restrictions Involving Multiple Coefficients

- Consider the following model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- And suppose we are interested in testing

$$H_0 : \beta_1 = \beta_2$$
$$H_1 : \beta_1 \neq \beta_2$$

- This null imposes a single restriction ($q = 1$) on multiple coefficients (in this example, on two coefficients)
- Don't confuse this with

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \beta_i \neq 0, \text{ for at least one } i = 1, 2$$

  This second test is a joint hypothesis with multiple restrictions ($q = 2$ in this case)

# Two Methods

There are two methods for testing single restrictions on multiple coefficients:

1) **Rearrange (transform) the regression:** Rearrange the regressors so that the restriction becomes a restriction on a single coefficient in an equivalent regression.
2) **Perform the test directly:** Some software, including R and STATA, lets you test restrictions using multiple coefficients directly.

# Method 1: Rearrange (transform) the regression

- **Original Model and Test**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$H_0 : \beta_1 = \beta_2 \text{ vs. } H_1 : \beta_1 \neq \beta_2$$

- Add and subtract $\beta_2 X_{1i}$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + [\beta_2 X_{1i} - \beta_2 X_{1i}] + \beta_3 X_{3i} + u_i \\ &= \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + \beta_3 X_{3i} + u_i \\ &= \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + \beta_3 X_{3i} + u_i \end{aligned}$$

  where $\gamma_1 = \beta_1 - \beta_2$ and $W_i = X_{1i} + X_{2i}$

- **Transformed Model and Test**

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + \beta_3 X_{3i} + u_i$$

$$H_0 : \gamma_1 = 0 \text{ vs. } H_1 : \gamma_1 \neq 0$$

- These two regressions have the same $R^2$, $\widehat{u}_i$ and $\widehat{Y}_i$
- To test $H_0 : \gamma_1 = 0$, use $t$-test as usual ...

# Method 1: Example

- Original Test:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

$$H_0 : \beta_1 = \beta_2 \text{ vs. } H_1 : \beta_1 \neq \beta_2$$

- Transformed Test:

$$TestScore_i = \beta_0 + \gamma_1 STR_i + \beta_2 W_i + \beta_3 PctEL_i + u_i$$

$$H_0 : \gamma_1 = 0 \text{ vs. } H_1 : \gamma_1 \neq 0$$

where $\gamma_1 = \beta_1 - \beta_2$ and $W_i = STR_i + Expn_i$

# Method 1: R Implementation

```
TestScore <- (CASchools$read + CASchools$math)/2
STR <- CASchools$students/CASchools$teachers
PctEL <- CASchools$english
Expn <- CASchools$expenditure/1000   # in million dollars
W <- STR + Expn  # compute the new variable
# run the regression:
myReg <- lm(TestScore ~ STR + W + PctEL)
summary(myReg) # display the regression output:

Call:
lm(formula = TestScore ~ STR + W + PctEL)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 649.57795   15.20572  42.719  < 2e-16 ***
STR          -4.15430    1.17676  -3.530 0.000461 ***
W             3.86790    1.41212   2.739 0.006426 **
PctEL        -0.65602    0.03911 -16.776  < 2e-16 ***
```

The estimated coefficient of $STR$ is highly significant, so we reject $H_0 : \gamma_1 = 0$, equivalently we reject $H_0 : \beta_1 = \beta_2$.

# Method 2: Perform the test directly in R

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

$$H_0 : \beta_1 = \beta_2 \text{ vs. } H_1 : \beta_1 \neq \beta_2$$

R commands to perform this test:

```
TestScore <- (CASchools$read + CASchools$math)/2
STR <- CASchools$students/CASchools$teachers
PctEL <- CASchools$english
Expn <- CASchools$expenditure/1000

library("car")  # this package is required:

# estimate the regression:
myReg <- lm(TestScore ~ STR + PctEL + Expn)

# this command performs the test:
linearHypothesis(myReg, c(0,1,-1,0), vcov = hccm)
```

# R Output

```
> linearHypothesis(myReg, c(0,1,-1,0), vcov = hccm)
Linear hypothesis test

Hypothesis:
STR - Expn = 0

Model 1: restricted model
Model 2: TestScore ~ STR + Expn + PctEL

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    417
2    416  1 8.6366 0.003478 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
```

Since $p$-value is 0.003478, which is smaller than any desirable significance level, we reject $H_0$.

# Model Specification for Multiple Regression

- We want to get an unbiased estimate of the effect on test scores of changing class size, holding constant factors outside the school committee's control - such as outside learning opportunities (museums, etc), parental involvement in education (reading with mom at home?), etc.
- If we could run an experiment, we would randomly assign students (and teachers) to different sized classes. Then $STR_i$ would be independent of all the things that go into $u_i$, so $E(u_i|STR_i) = 0$ and the OLS estimator on $STR_i$ would be an unbiased.
- But with observational data, $u_i$ might depend on additional factors (museums, parental involvement, knowledge of English etc) that can cause omitted variable bias.

- The general conditions for omitted variable (OV) bias in multiple regression are similar to those for a single regressor: **OV bias** is the bias in the OLS estimator that arises when one or more included regressors are correlated with an omitted variable.
- For omitted variable bias to arise, two things must be true:
  1) At least one of the included regressors must be correlated with the omitted variable (say $Z$).
  2) The omitted variable, $Z$, must be a determinant of the dependent variable, $Y$.
- OV bias $\implies E\left(u_i|X_{1i},\ldots,X_{ki}\right) \neq 0$ (LSA # doesn't hold)
- To correct OV bias:
  - If you can observe those factors (e.g. $PctEL$), then include them in the regression.
  - But usually you can't observe all these omitted causal factors (e.g. parental involvement in homework). In this case, you can include **control variables**.

# Control Variables

- A **control variable W** is a variable that is correlated with, and controls for, an omitted causal factor in the regression of $Y$ on $X$, but which itself does not necessarily have a causal effect on $Y$.
- As we discussed above omitting "outside learning opportunities" from a test score regression might cause omitted variable bias
- To control for omitted income-related determinants of test scores, like outside learning opportunities, we augment the regression of test scores on $STR$ and $PctEL$ with the percentage of students receiving a free or subsidized school lunch ($LchPct$):

$$\widehat{\text{TestScore}} = \underset{(5.6)}{700.2} - \underset{(0.27)}{1.00}\, STR - \underset{(0.033)}{0.122}\, PctEL - \underset{(0.024)}{0.547}\, LchPct$$

- In this regression
  - which variable is the variable of interest?
  - which variables are control variables?
  - do they have causal components? What do they control for?

# Control Variables: Example

$$\widehat{\text{TestScore}} = \underset{(5.6)}{700.2} - \underset{(0.27)}{1.00}\, STR - \underset{(0.033)}{0.122}\, PctEL - \underset{(0.024)}{0.547}\, LchPct$$

- STR is the variable of interest
- $PctEL$ probably has a direct causal effect (school is tougher if you are learning English!). But it is also a control variable: immigrant communities tend to be less affluent and often have fewer outside learning opportunities, and $PctEL$ is correlated with those omitted causal variables. $PctEL$ is both a possible causal variable and a control variable.
- $LchPct$ might have a causal effect (eating lunch helps learning); it also is correlated with and controls for income-related outside learning opportunities. $LchPct$ is potentially both a possible causal variable and a control variable. But in this particular regression we will see that it is not a causal variable!

# What makes an effective control variable?

Three interchangeable statements about what makes an effective control variable:

- An effective control variable is one which, when included in the regression, makes the error term uncorrelated with the variable of interest.
- Holding constant the control variable(s), the variable of interest is "as if" randomly assigned.
- Among individuals (entities) with the same value of the control variable(s), the variable of interest is uncorrelated with the omitted determinants of $Y$

# Control variables need not be causal

Control variables need not be causal, and their coefficients generally do NOT have a causal interpretation. For example consider the same regression:

$$\widehat{\text{TestScore}} = \underset{(5.6)}{700.2} - \underset{(0.27)}{1.00}\, STR - \underset{(0.033)}{0.122}\, PctEL - \underset{(0.024)}{0.547}\, LchPct$$

Does the coefficient on LchPct have a causal interpretation?

- If so, then we should be able to boost test scores (by a lot! Do the math!) by simply eliminating the school lunch program, so that $LchPct = 0$!
- Therefore, $LchPct$ is not a causal variable in this model
- Does it make sense to treat the coefficient on the variable of interest $STR$ as causal, but not the coefficient on the control variable $LchPct$?

# Formal statement of what makes an effective control variable

- Recall that OV bias implies that LSA #1 $E\left(u_i|X_{1i},\ldots,X_{ki}\right)$ does not hold.
- If LSA #1 doesn't hold, then what does?
- We need a mathematical statement of what makes an effective control variable. This condition is **conditional mean independence:** given the control variable, the mean of $u_i$ doesn't depend on the variable of interest
- Formal Statement: Let $X$ denote the variable of interest and $W$ denote the control variable(s). $W$ is an effective control variable if conditional mean independence holds:

$$E\left(u|X,W\right)=E\left(u|W\right)$$

- If $W$ is a control variable, then conditional mean independence replaces LSA #1 - it is the version of LSA #1 which is relevant for control variables.

# Implications for variable selection and "model specification"

- Identify the variable of interest
- Think of the omitted causal effects that could result in omitted variable bias
- Include those omitted causal effects if you can or, if you can't, include variables correlated with them that serve as control variables.
- The control variables are effective if the conditional mean independence assumption plausibly holds (if $u$ is uncorrelated with $STR$ once the control variables are included). This results in a "base" or "benchmark" model.

# Implications for variable selection and "model specification"

- Also specify a range of plausible alternative models, which include additional candidate variables.
- Estimate your base model and plausible alternative specifications ("sensitivity checks").
  - Does a candidate variable change the coefficient of interest ($\beta_1$)?
  - Is a candidate variable statistically significant?
  - Use judgment, not a mechanical recipe...
  - Don't just try to maximize $R^2$!

# Digression about measures of fit...

It is easy to fall into the trap of maximizing the $R^2$ and $\overline{R}^2$, but this loses sight of our real objective, an unbiased estimator of the class size effect.

- A high $R^2$ (or $\overline{R}^2$) means that the regressors explain the variation in $Y$ quite well.
- A high $R^2$ (or $\overline{R}^2$) does NOT mean that you have eliminated omitted variable bias.
- A high $R^2$ (or $\overline{R}^2$) does NOT mean that you have an unbiased estimator of a causal effect ($\beta_1$).
- A high $R^2$ (or $\overline{R}^2$) does NOT mean that the included variables are statistically significant - this must be determined using hypotheses tests.

# Analysis of the Test Score Data Set

1) Identify the variable of interest: $STR$
2) Think of the omitted causal effects that could result in omitted variable bias:
   - whether the students know English
   - outside learning opportunities
   - parental involvement
   - teacher quality (if teacher salary is correlated with district wealth)
   - ... there is a long list!

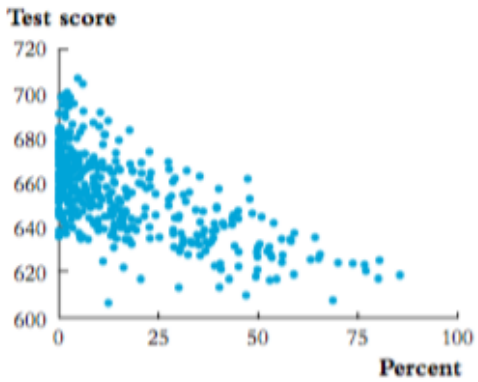
3) Include those omitted causal effects if you can or, if you can't, include variables correlated with them that serve as control variables.
   - Many of the omitted causal variables are hard to measure, so we need to find control variables. These include $PctEL$ (both a control variable and an omitted causal factor) and measures of district wealth.

4) Also specify a range of plausible alternative models, which include additional candidate variables.
   - It isn't clear which of the income-related variables will best control for the many omitted causal factors such as outside learning opportunities, so the alternative specifications include regressions with different income variables.

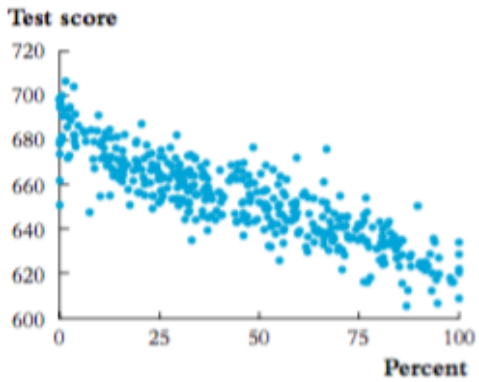
5) Estimate your base model and plausible alternative specifications ("sensitivity checks").

# Scatterplots

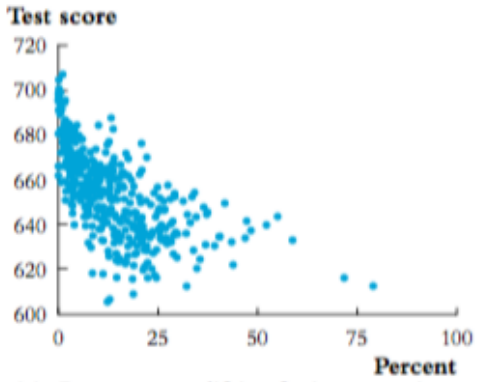
**FIGURE 7.2** Scatterplots of Test Scores vs. Three Student Characteristics




(a) Percentage of English language learners




(b) Percentage qualifying for reduced price lunch




(c) Percentage qualifying for income assistance

The scatterplots show a negative relationship between test scores and (a) the percentage of English learners (correlation $= -0.64$), (b) the percentage of students qualifying for a reduced price lunch (correlation $= -0.87$); and (c) the percentage qualifying for income assistance (correlation $= -0.63$).

# Tabular Presentation of Results

**TABLE 7.1** Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

**Dependent variable: average test score in the district.**

| Regressor | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Student–teacher ratio ($X_1$) | −2.28** | −1.10* | −1.00** | −1.31* | −1.01* |
| | (0.52) | (0.43) | (0.27) | (0.34) | (0.27) |
| Percent English learners ($X_2$) | | −0.650** | −0.122** | −0.488** | −0.130** |
| | | (0.031) | (0.033) | (0.030) | (0.036) |
| Percent eligible for subsidized lunch ($X_3$) | | | −0.547* | | −0.529* |
| | | | (0.024) | | (0.038) |
| Percent on public income assistance ($X_4$) | | | | −0.790** | 0.048 |
| | | | | (0.068) | (0.059) |
| Intercept | 698.9** | 686.0** | 700.2** | 698.0** | 700.4** |
| | (10.4) | (8.7) | (5.6) | (6.9) | (5.5) |
| **Summary Statistics** | | | | | |
| $SER$ | 18.58 | 14.46 | 9.08 | 11.65 | 9.08 |
| $\overline{R}^2$ | 0.049 | 0.424 | 0.773 | 0.626 | 0.773 |
| $n$ | 420 | 420 | 420 | 420 | 420 |

These regressions were estimated using the data on K–8 school districts in California, described in Appendix (4.1). Heteroskedasticity-robust standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

# Discussion of Empirical Results

- Controlling for these student characteristics cuts the effect of the STR on test scores approximately in half.
  - This estimated effect is not very sensitive to which specific control variables are included in the regression. In all cases the coefficient on the STR remains statistically significant at the 5% level.
  - In the four specifications with control variables, regressions (2) through (5), reducing the STR by one student per teacher is estimated to increase average test scores by approximately 1 point, holding constant student characteristics.
- The student characteristic variables are potent predictors of test scores.
  - The STR alone explains only a small fraction of the variation in test scores (the $\overline{R}^2$ in column (1) is 0.049).
  - The $\overline{R}^2$ jumps, however, when the student characteristic variables are added. For example, the $\overline{R}^2$ in the base specification, regression (3), is 0.773. The signs of the coefficients on the student demographic variables are as expected...

# Discussion of Empirical Results

- The control variables are not always individually statistically significant.
  - In specification (5), the hypothesis that the coefficient on the percentage qualifying for income assistance is zero is not rejected at the 5% level (the t-statistic is -0.82).
  - Because adding this control variable to the base specification (3) has a negligible effect on the estimated coefficient for the studentteacher ratio and its standard error, and because the coefficient on this control variable is not significant in specification (5), this additional control variable is redundant, at least for the purposes of this analysis.