Advanced Data Analysis I

Testing hypotheses about model parameters

PA 541 Week 5

Michael D. Siciliano

Department of Public Administration

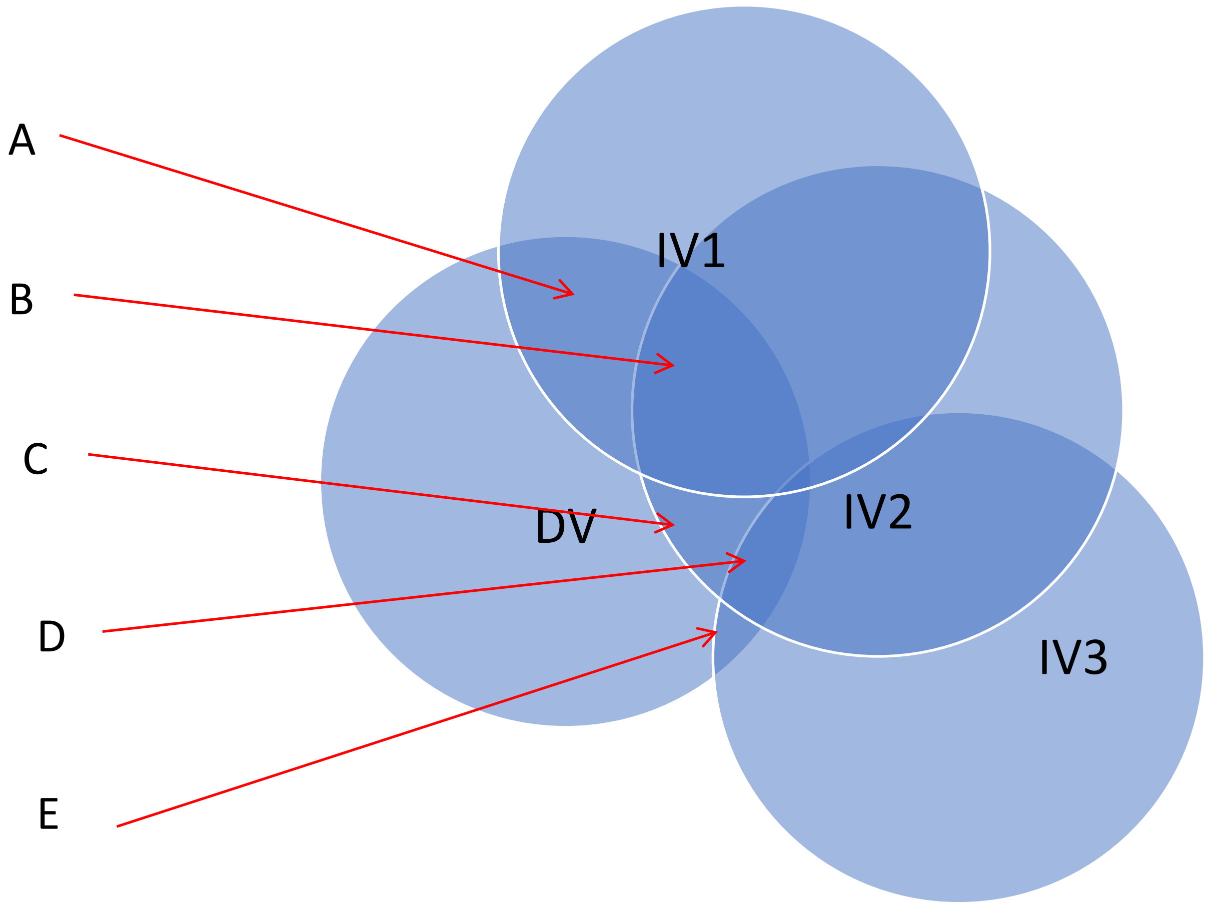
College of Urban Planning and Public Affairs

Today’s topics

- Review omitted variable bias (and quickly review in-class exercise).
- Statistical inference
- Practical significance versus statistical significance
    - Identifying the “importance” of a predictor
- Testing one-sided alternatives
- Testing the equality of regression coefficients
- Testing multiple linear restrictions

Starter Question

- Think back to your, perhaps painful, memories of our Venn Diagram last week and our discussion of omitted variable bias?
- Within this context, why is running an experiment so useful in understanding the relationship between X and Y.



Let’s REVIEW Week 4

Interpreting Beta Coefficients

Model Specification

Causality

Why Multiple Regression?

Endogeneity

P-values

Omitted Variable Bias

Including Irrelevant Predictors

Ceteris Paribus

Bias Toward Zero

Partialling Out

It is always easier to justify a coefficient after the fact….

- Consider this example (Watts 2011)
  - Watts discusses the fact that many people don’t see what sociology or other social sciences can tell us about the world that an intelligent person couldn’t figure out on their own. Here is an example.
  - Paul Lazarsfeld was writing about the *The American Soldier,* a published study on 600,00 servicemen during and after WWII.
  - He listed several findings that were claimed to be representative of the report.
    - Ex. Men from rural backgrounds were usually in better spirits during their army life than soldiers from city backgrounds.
    - Why?

- Aha, you might have said, that makes perfect sense.
    - Rural men in the 1940s were accustomed to harsher living standards and more physical labor than city men, so naturally they had an easier time adjusting.
    - Why did we need a vast and expensive study to tell me what I could have figured out on my own?

- However, each of the findings that Lazarsfeld listed was in fact the exact opposite of what the study had actually found.
    - It was actually city men and not rural men who were happier during their army life.
    - Of course, had one been given this finding originally he/she could have just as easily reconciled it by saying:
        - City men are more used to working in crowded conditions and in corporations with chains of command, strict standards of clothing, defined social norms, may have seen more violence, etc…
        - That finding is obvious, why did I need a big expensive study to tell me that?

Review Omitted Variable Slides

From last week

Let’s look at an example from the text

- **<u>Omitting relevant variables: the simple case</u>**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

True model (contains x_1 and x_2)

$$y = \alpha_0 + \alpha_1 x_1 + w$$

Estimated model (x_2 is omitted)

Example from the text…

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Again, here is the ‘true’ model

- **Omitted variable bias**

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

If x_1 and x_2 are correlated, assume a linear regression relationship between them

$$\Rightarrow \quad y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$$

$$= \underbrace{\boxed{(\beta_0 + \beta_2\delta_0)}}_{\text{If y is only regressed on } x_1 \text{ this will be the estimated intercept}} + \underbrace{\boxed{(\beta_1 + \beta_2\delta_1)}}_{\text{If y is only regressed on } x_1\text{, this will be the estimated slope on } x_1} x_1 + \underbrace{\boxed{(\beta_2 v + u)}}_{\text{error term}}$$

- **<u>Conclusion:</u> All estimated coefficients will be biased**

Omitted variable bias occurs when **both** of the following conditions are met

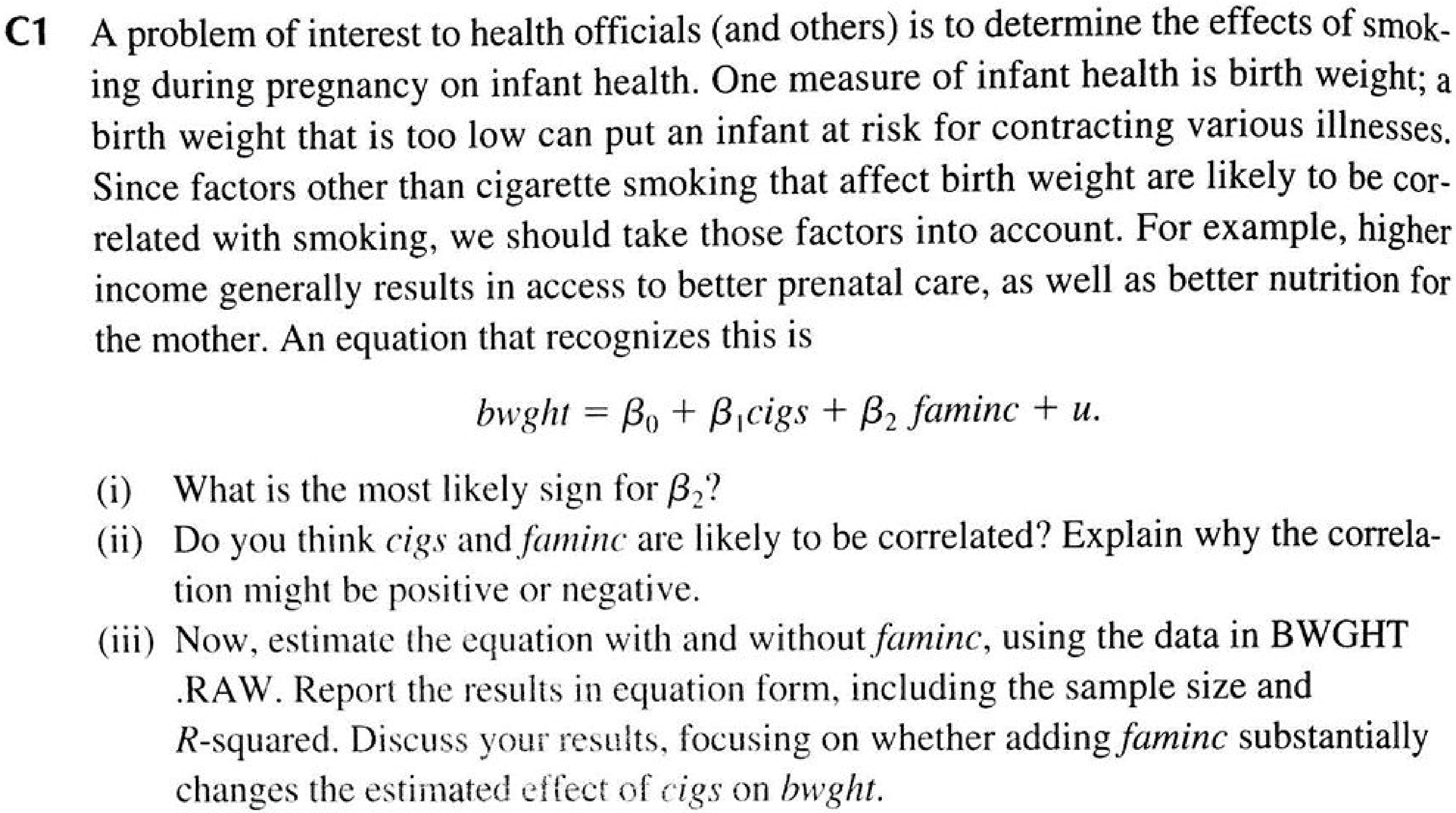
1. The omitted variable affects the dependent variable.

$$\beta_2 \neq 0$$

2. The omitted variable is correlated with the included independent variable

$$\delta_1 \neq 0$$

Let’s look at part of the in-class exercise



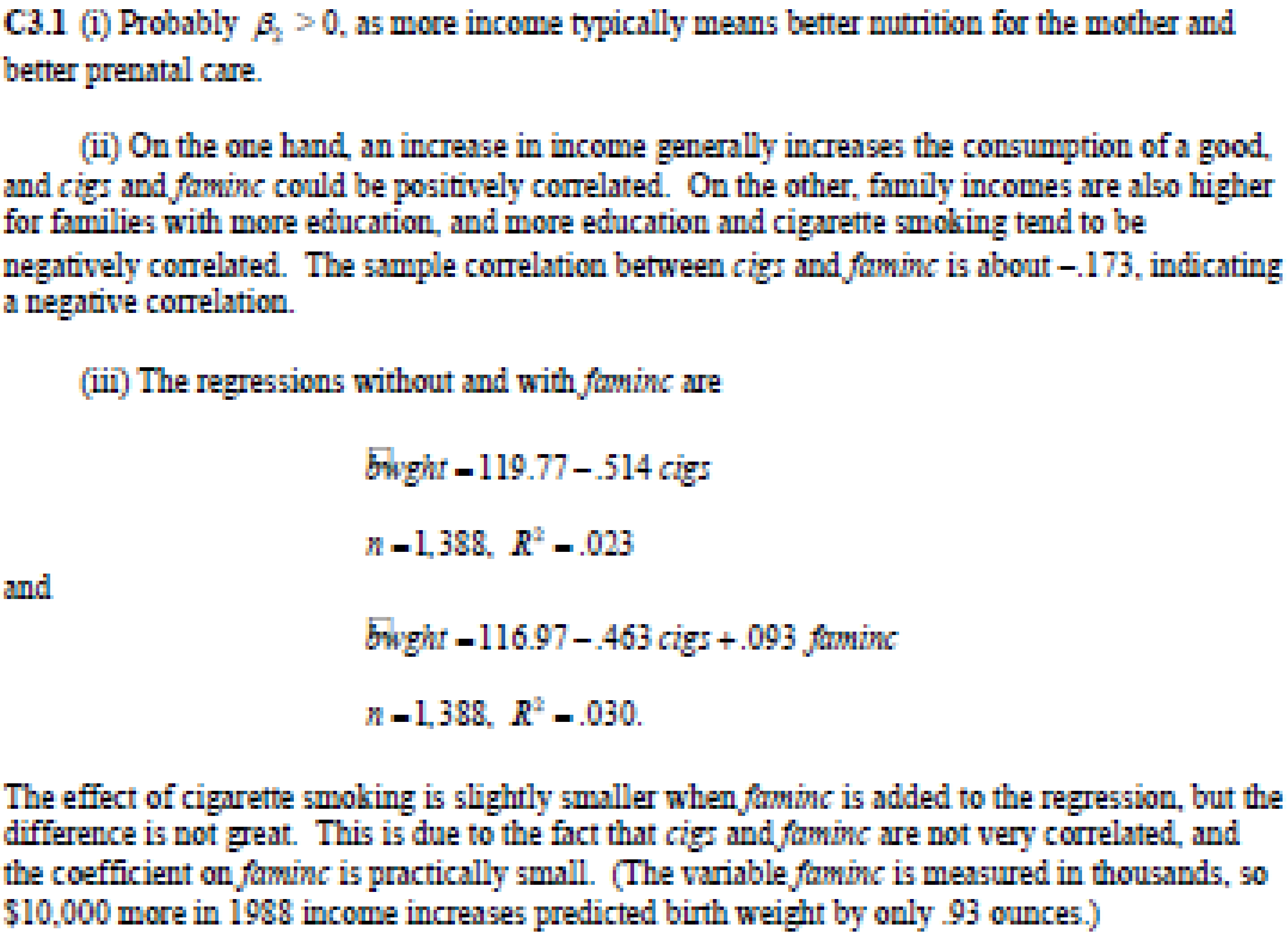
C1 A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is

$$bwght = \beta_0 + \beta_1 cigs + \beta_2\, faminc + u.$$

(i) What is the most likely sign for β_2?

(ii) Do you think *cigs* and *faminc* are likely to be correlated? Explain why the correlation might be positive or negative.

(iii) Now, estimate the equation with and without *faminc*, using the data in BWGHT .RAW. Report the results in equation form, including the sample size and *R*-squared. Discuss your results, focusing on whether adding *faminc* substantially changes the estimated effect of *cigs* on *bwght*.



C3.1 (i) Probably $\beta_2 > 0$, as more income typically means better nutrition for the mother and better prenatal care.

(ii) On the one hand, an increase in income generally increases the consumption of a good, and *cigs* and *faminc* could be positively correlated. On the other, family incomes are also higher for families with more education, and more education and cigarette smoking tend to be negatively correlated. The sample correlation between *cigs* and *faminc* is about –.173, indicating a negative correlation.

(iii) The regressions without and with *faminc* are

$$\widehat{bwght} = 119.77 - .514\, cigs$$

$$n = 1{,}388,\ R^2 = .023$$

and

$$\widehat{bwght} = 116.97 - .463\, cigs + .093\, faminc$$

$$n = 1{,}388,\ R^2 = .030.$$

The effect of cigarette smoking is slightly smaller when *faminc* is added to the regression, but the difference is not great. This is due to the fact that *cigs* and *faminc* are not very correlated, and the coefficient on *faminc* is practically small. (The variable *faminc* is measured in thousands, so $10,000 more in 1988 income increases predicted birth weight by only .93 ounces.)

Answer to C1

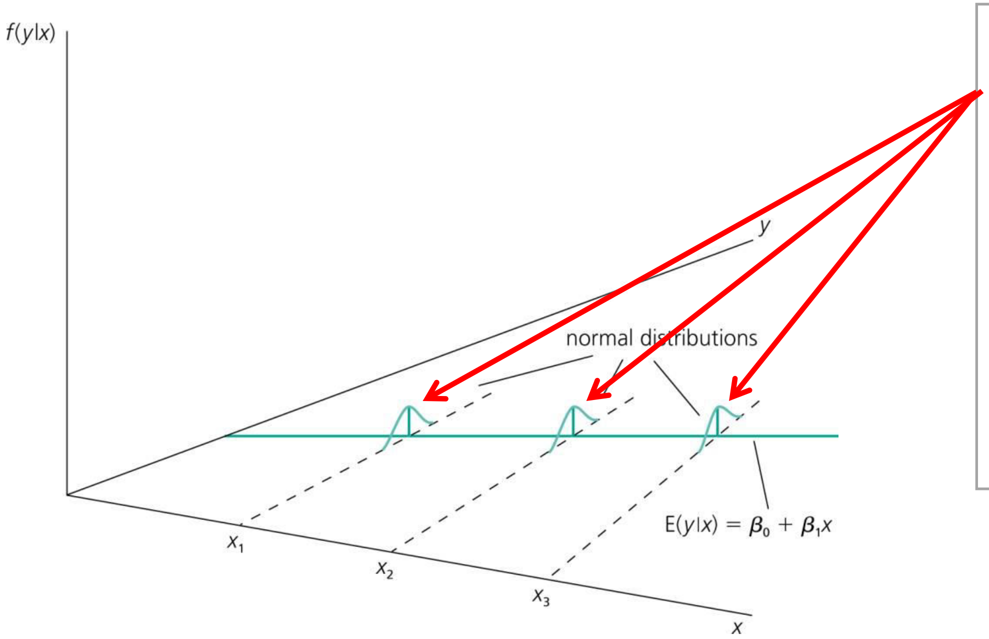
Statistical Inference

Statistical Inference

- **Statistical inference in the regression model**
  - Hypothesis tests about population parameters
  - Construction of confidence intervals
- **Sampling distributions of the OLS estimators**
  - We already know their expected values and their variances
  - However, for hypothesis tests we need to know their distribution
  - In order to derive their distribution we need additional assumptions
  - Assumption about distribution of errors: **normal distribution**
    - **Note: only really need this assumptions for small samples**

- **Assumption MLR.6 (Normality of error terms)**

$$u_i \sim N(0, \sigma^2) \quad \text{independently of} \quad x_{i1}, x_{i2}, \ldots, x_{ik}$$



It is assumed that the unobserved factors are normally distributed around the population regression function.

The form and the variance of the distribution does not depend on any of the explanatory variables.

It follows that:

$$y|\mathbf{x} \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$$

- **Discussion of the normality assumption**
  - The normality of the error term can be examined through the residuals
    - We did this by looking at a histogram of the residuls as well as a qq-plot.
  - At least the distribution should be close to normal
    - Small deviations from normality are not a problem
  - In many cases, normality is questionable or impossible by definition (i.e. the distribution can take on only take on certain values)

- **Discussion of the normality assumption (cont.)**
  - Examples where normality cannot hold:
    - Wages (nonnegative; also: minimum wage)
    - Number of arrests (takes on a small number of integer values)
    - Unemployment (indicator variable, takes on only 1 or 0)
  - In some cases, normality can be achieved through transformations of the dependent variable (e.g. use log(wage) instead of wage)
    - We will be discussing transformations in the coming weeks and models for non-normal variables later in the semester.

- **Terminology**

$$\underbrace{MLR.1 - MLR.5}_{\text{Gauss-Markov assumptions}} \qquad \underbrace{MLR.1 - MLR.6}_{\text{Classical linear model (CLM) assumptions}}$$

- **Theorem 4.1 from Wooldridge (Normal sampling distributions)**
  Under assumptions MLR.1 – MLR.6:

$$\widehat{\beta}_j \sim N(\beta_j, Var(\widehat{\beta}_j)) \qquad \frac{\widehat{\beta}_j - \beta_j}{sd(\widehat{\beta}_j)} \sim N(0,1)$$

The estimators are normally distributed around the true parameters

The standardized estimators follow a standard normal distribution

- **Testing hypotheses about a single population parameter**
- **Theorem 4.1 (t-distribution for standardized estimators)**

Under assumptions MLR.1 – MLR.6:

$$\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \sim t_{n-k-1}$$

If the standardization is done using the estimated standard deviation (= standard error), the normal distribution is replaced by a t-distribution

Note: The t-distribution is close to the standard normal distribution if n-k-1 is large.

- **Null hypothesis**

$$H_0 : \ \beta_j = 0$$

The population parameter is equal to zero, i.e. after controlling for the other independent variables, there is no effect of x_j on y

- **t-statistic (or t-ratio)**

$$t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$$

The t-statistic will be used to test the above null hypothesis. The farther the estimated coefficient is away from zero, the less likely it is that the null hypothesis holds true. But what does „far“ away from zero mean?

This depends on the variability of the estimated coefficient, i.e. its standard deviation. The t-statistic measures how many estimated standard deviations the estimated coefficient is away from zero.

- **Distribution of the t-statistic if the null hypothesis is true**

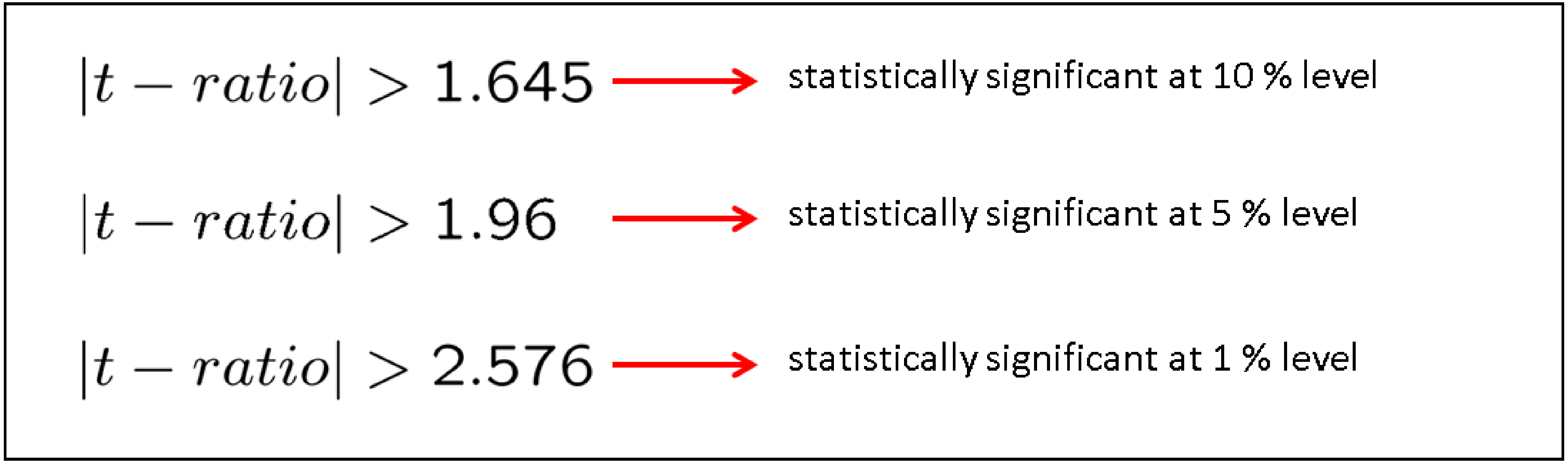
$$t_{\widehat{\beta}_j} = \widehat{\beta}_j / se(\widehat{\beta}_j) = (\widehat{\beta}_j - \beta_j)/se(\widehat{\beta}_j) \sim t_{n-k-1}$$

- **Goal: Define a rejection rule so that, if it is true, H_0 is rejected only with a small probability (= significance level, e.g. 5%)**

```
> lm1 = lm (prestige ~ education + income, data=prest)
> summary(lm1)

Call:
lm(formula = prestige ~ education + income, data = prest)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.4040  -5.3308   0.0154   4.9803  17.6889 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.8477787  3.2189771  -2.127   0.0359 *  
education    4.1374444  0.3489120  11.858  < 2e-16 ***
income       0.0013612  0.0002242   6.071 2.36e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.81 on 99 degrees of freedom
Multiple R-squared: 0.798,	Adjusted R-squared: 0.7939 
F-statistic: 195.6 on 2 and 99 DF,  p-value: < 2.2e-16
```

- Again, if we return to our model from last week, we see that we have an estimate for the coefficient, an estimate for the st. error and from those we produce the t-statistic that allows us to test for significance at a specified level.

- **Statistically significant variables in a regression**
  - If a regression coefficient is different from zero in a two-sided test, the corresponding variable is said to be <u>statistically significant</u>
  - If the number of degrees of freedom is large enough so that the normal approximation applies, the following rules of thumb apply:



$|t - ratio| > 1.645 \longrightarrow$ statistically significant at 10 % level

$|t - ratio| > 1.96 \longrightarrow$ statistically significant at 5 % level

$|t - ratio| > 2.576 \longrightarrow$ statistically significant at 1 % level

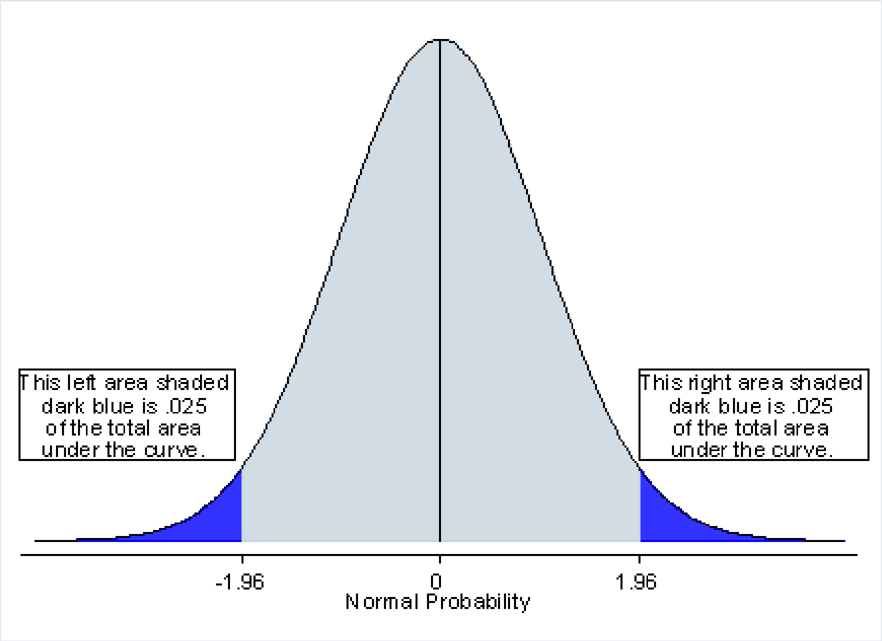
Practical versus statistical inference

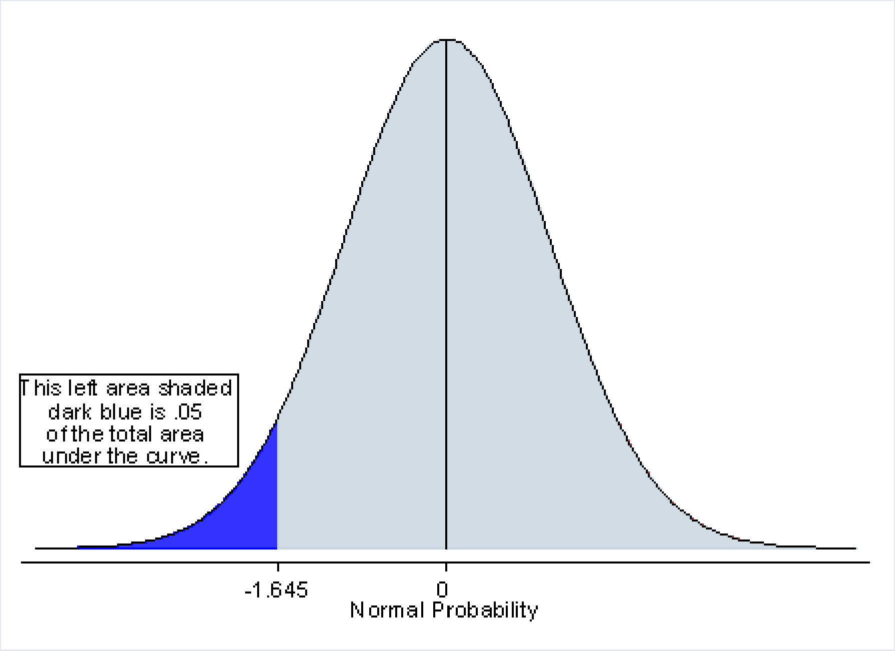
- Roughly speaking, a coefficient is significant if it is more than 2 standard errors away from zero.
    - When it is statistically significant, then we can be fairly confident that the direction of its effect is stable, and not just an artifact of the sample.

- To help assess **practical** significance we need to know the range of our variable and its unit of measurement.
    - Remember, the coefficient tells us the impact on the dependent variable when the independent variable changes by 1.

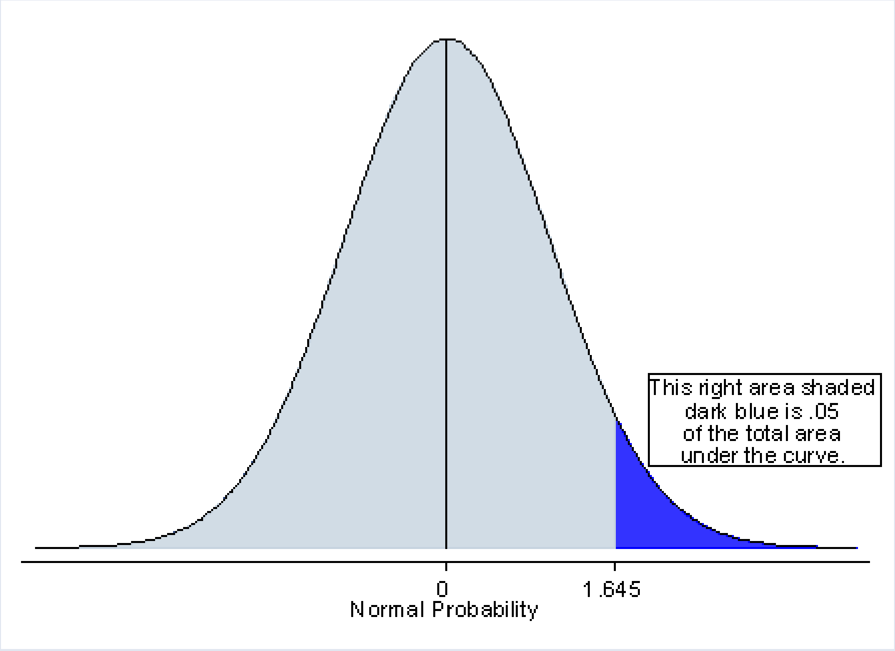
- For certain variables 1 could be more than the entire range (say if the variable is the fraction of something) or 1 could be a very small increase (say if we measured GDP in dollars as opposed to millions of dollars).
- Thus, we can’t look only at the coefficient to get a sense of its importance. We need to look at the range, and then we need to consider:
    - What size gain (or decrease) in Y makes a significant impact?
    - What portion of the range of X, do we need to move to obtain that size gain in Y.
    - From a policy standpoint, is it practical to think we could move X that far?

Testing one-sided alternatives





- **Two-tailed test** – if using a 5% significance level, a two-tailed test allots half of your alpha in each tail.
- **One-tailed test** – if using a 5% significance level, a one-tailed test allots all of your alpha to testing significance in one direction.



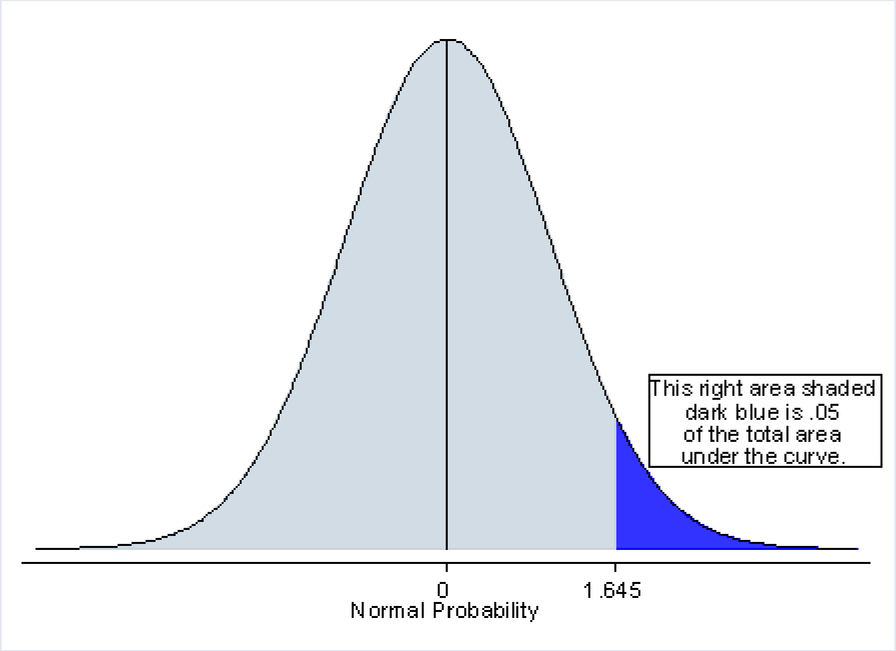
When should you use a one tailed test?

- Using a one-tailed test clearly provides more power to detect an effect. Because of this, you may be tempted to use it if you have a hypothesis regarding the direction of the effect of X on Y.

- However, one-tailed tests are rarely used.
  - It precludes your ability to find significance in the other direction.
  - Readers/reviewers are skeptical if they see a one-tail test that is just on the border of significance.
    - You should never choose to do a one-tailed test after you observed a non-significant two-tailed test.

Obtaining a one-tailed test in R

```
> lm1=lm(salary ~ roe, data=ceo)
> summary(lm1)

Call:
lm(formula = salary ~ roe, data = ceo)

Residuals:
    Min      1Q  Median      3Q     Max 
-1160.2  -526.0  -254.0   138.8 13499.9 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   963.19     213.24   4.517 1.05e-05 ***
roe            18.50      11.12   1.663   0.0978 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 1367 on 207 degrees of freedom
Multiple R-squared: 0.01319,	Adjusted R-squared: 0.008421 
F-statistic: 2.767 on 1 and 207 DF,  p-value: 0.09777
```



- Statistical packages, like R, report two-tailed tests. Because t-distribution is symmetric about zero, we can derive one-tailed p-values from the two-tailed test.
- We can divide the p-value by 2, assuming we hypothesized a positive coefficient for *roe*. Thus, the significance of a one-tailed test would be .0489.

Hypothesis Testing About Multiple Coefficients

1. Testing the equality of regression coefficients
2. Testing exclusion restrictions

1. Testing the equality of regression coefficients

Testing the equality of regression coefficients

- When might we care that two regression coefficients in our model are equal?
  - Perhaps, as Wooldridge did, we want to know if time spent at different types of higher ed institutions provides the same return.
  - Perhaps we want to compare hours of professional development in different programs and their impact on job performance.
  - Maybe I want to compare if time spent working on R has the same impact on your grade as time spent reading Wooldridge.
  - You may have a model with different variables for race and you want to test the equivalence of the effect of Black or Hispanic.
  - There may be plenty of reasons you care about the equality of two regression coefficients.

Let’s revisit the example from the reading this week…

- **<u>Testing hypotheses about a linear combination of parameters</u>**
- **Example: Return to education at 2 year vs. at 4 year colleges**

Years of education at 2 year colleges

Years of education at 4 year colleges

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

Test $H_0 : \beta_1 - \beta_2 = 0$ against $H_1 : \beta_1 - \beta_2 < 0$

A possible test statistic would be:

$$t = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{se(\widehat{\beta}_1 - \widehat{\beta}_2)}$$

The difference between the estimates is normalized by the estimated standard deviation of the difference. The null hypothesis would have to be rejected if the statistic is „too negative“ to believe that the true difference between the parameters is equal to zero.

- **Impossible to compute with standard regression output because**

$$se(\widehat{\beta}_1-\widehat{\beta}_2) = \sqrt{\widehat{Var}(\widehat{\beta}_1-\widehat{\beta}_2)} = \sqrt{\widehat{Var}(\widehat{\beta}_1)+\widehat{Var}(\widehat{\beta}_2)-2\widehat{Cov}(\widehat{\beta}_1,\widehat{\beta}_2)}$$

Usually not available in regression output

- **Alternative method**

Define $\theta_1 = \beta_1 - \beta_2$ and test $H_0 : \theta_1 = 0$ against $H_1 : \theta_1 < 0$

$$\log(wage) = \beta_0 + \boxed{(\theta_1 + \beta_2) jc} + \beta_2 univ + \beta_3 exper + u$$

$$= \beta_0 + \theta_1 jc + \beta_2 \boxed{(jc + univ)} + \beta_3 exper + u$$

Insert into original regression

a new regressor (= total years of college)

- **Estimation results**

Total years of college

$$\widehat{\log}(wage) = \underset{(.021)}{1.472} - \underset{(.0069)}{.0102}\, jc + \underset{(.0023)}{.0769}\, totcoll + \underset{(.0002)}{.0049}\, exper$$

$$n = 6,763, \;\; R^2 = .222$$

Fail to reject the null at the 5% level, but we do reject the null at the 10% level

$$t = -.0102/.0069 = -1.48$$

$$p - value = P(t - ratio < -1.48) = .070$$

$$-.0102 \pm 1.96(.0069) = (-.0237, .0003)$$

- **This method works <u>always</u> for single linear hypotheses**

Let’s look at another example and two ways to make the comparison.

- The Data: Duncan Dataset available in the ‘car’ package (similar to our prestige data from last week)
- This data frame contains the following columns:
  - *Type* - Type of occupation. A factor with the following levels: *prof*, professional and managerial; *wc*, white-collar; *bc*, blue-collar.
  - *Income* - Percent in occupation earning $3500 or more in 1950.
  - *Education* - Percent in occupation in 1950 who were high-school graduates.
  - *Prestige* - Percent of raters in NORC study rating occupation as excellent or good in prestige

Question, does the percentage of highly paid employees and the percentage of highly educated employees equally effect the prestige of a job?

Here is the model with income and education

```
> mod.duncan <- lm(prestige ~ income + education, data=Duncan)
> summary(mod.duncan)

Call:
lm(formula = prestige ~ income + education, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.538  -6.417   0.655   6.605  34.641 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.06466    4.27194  -1.420    0.163    
income       0.59873    0.11967   5.003 1.05e-05 ***
education    0.54583    0.09825   5.555 1.73e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared: 0.8282,	Adjusted R-squared:  0.82 
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

- We cannot make any claim about the equality of the coefficients on income and education based on this output.

Approach 1- Wooldridge version

```
> Duncan$inced = Duncan$income + Duncan$education
>
> mod3 = lm(prestige ~ income + inced, data=Duncan)
> summary(mod3)

Call:
lm(formula = prestige ~ income + inced, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.538  -6.417   0.655   6.605  34.641 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.06466    4.27194  -1.420    0.163    
income       0.05290    0.20251   0.261    0.795    
inced        0.54583    0.09825   5.555 1.73e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared: 0.8282,	Adjusted R-squared:  0.82 
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

Here we see no evidence that income’s effect is different from education’s effect

- Again, here is the approach used by Wooldridge.
  - Create a combined term (income + education).
  - Check the significance of the coefficient on the variable of interest (income).

Approach 2 – Use an R function

```
> library("car")
> mod.duncan <- lm(prestige ~ income + education, data=Duncan)
> linearHypothesis(mod.duncan, "income = education")
Linear hypothesis test

Hypothesis:
income - education = 0

Model 1: restricted model
Model 2: prestige ~ income + education

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     43 7518.9                           
2     42 7506.7  1    12.195 0.0682 0.7952
```

Same value as in previous slide

- The linearHypothesis() function allows you to set and test a variety of restrictions.
- Here we see that we get identical results to our previous method.

2. Testing exclusion restrictions

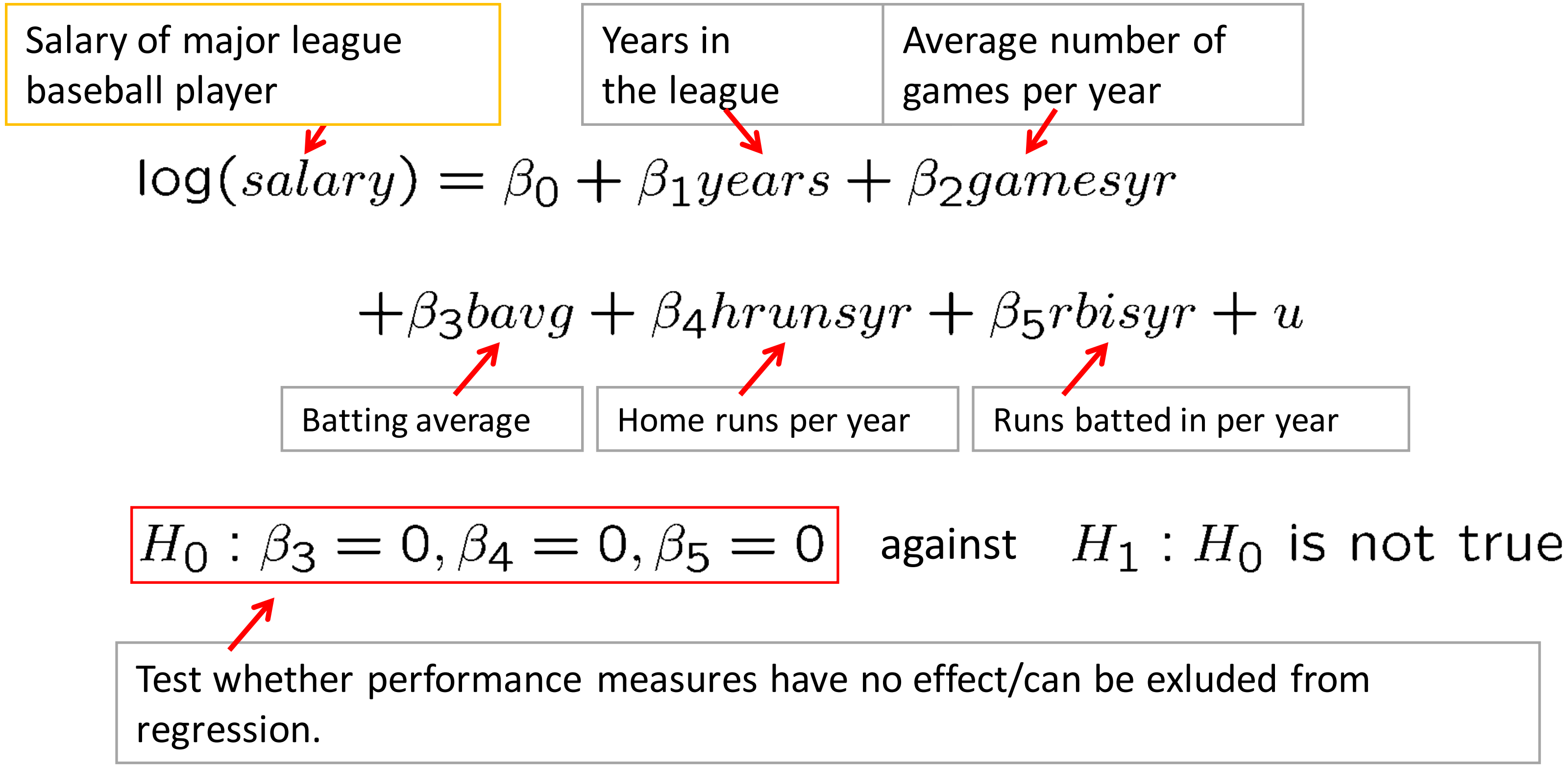
Testing exclusion restrictions

- There will be times when you wish to test multiple hypotheses about the population parameters.
- You may want to know if a restricted version of the model is “just as good” as the unrestricted version.
  - These are known as nested models and can be easily tested.
- Variables of interest to you may be highly correlated such that no single individual predictor is significant. In this case you may wish to test their joint significance.
  - If found jointly significant you can think about dropping a variable or creating a composite variable.

- Let’s first look at the example from the Wooldridge text and then I’ll walk us through another example with a separate dataset.

- **<u>Testing multiple linear restrictions: The F-test</u>**

- **Testing exclusion restrictions**



Salary of major league baseball player

Years in the league

Average number of games per year

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr$$

$$+\beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

Batting average

Home runs per year

Runs batted in per year

$\boxed{H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0}$ against $H_1 : H_0$ is not true

Test whether performance measures have no effect/can be exluded from regression.

- **Estimation of the unrestricted model**

$$\widehat{\log}(salary) = \underset{(0.29)}{11.19} + \underset{(.0121)}{.0689}\ years + \underset{(.0026)}{.0126}\ gamesyr$$

$$+ \underset{(.00110)}{.00098}\ bavg + \underset{(.0161)}{.0144}\ hrunsyr + \underset{(.0072)}{.0108}\ rbisyr$$

None of these variables is statistically significant when tested individually

$$n = 353,\ SSR = 183.186,\ R^2 = .6278$$

Idea: How would the model fit be if these variables were dropped from the regression?

- **Estimation of the restricted model**

$$\widehat{\log}(salary) = \underset{(0.11)}{11.22} + \underset{(.0125)}{.0713}\ years + \underset{(.0013)}{.0202}\ gamesyr$$

$$n = 353,\ SSR = 198.311,\ R^2 = .5971$$

The sum of squared residuals necessarily increases, but is the increase statistically significant?

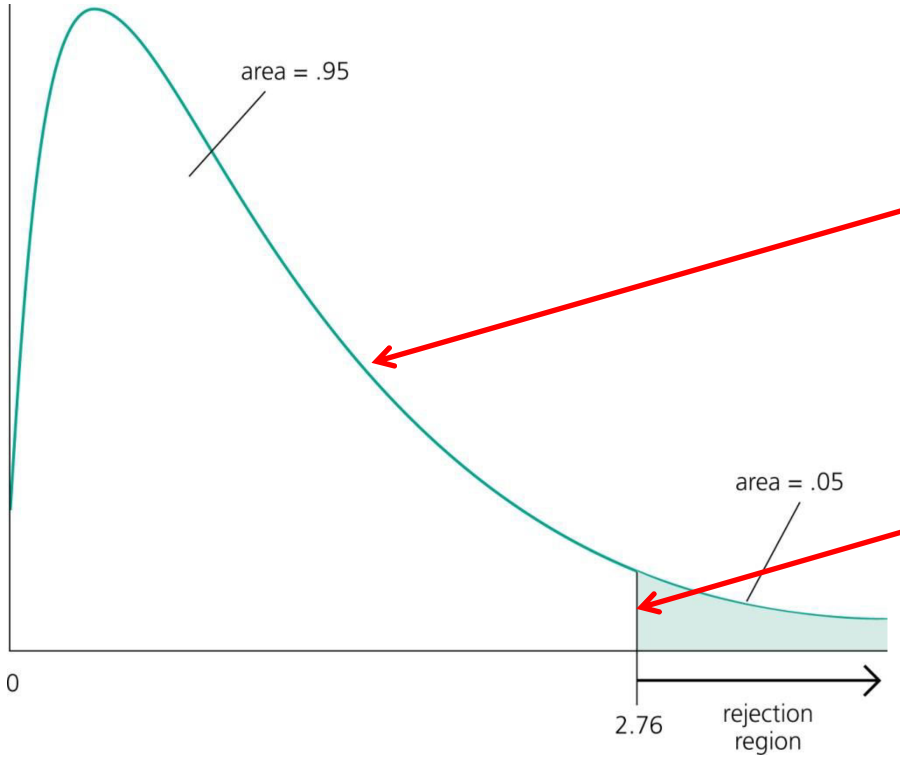
- **Test statistic**

Number of restrictions

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$

The relative increase of the sum of squared residuals when going from H_1 to H_0 follows a F-distribution (if the null hypothesis H_0 is correct)

- **Rejection rule (Figure 4.7)**

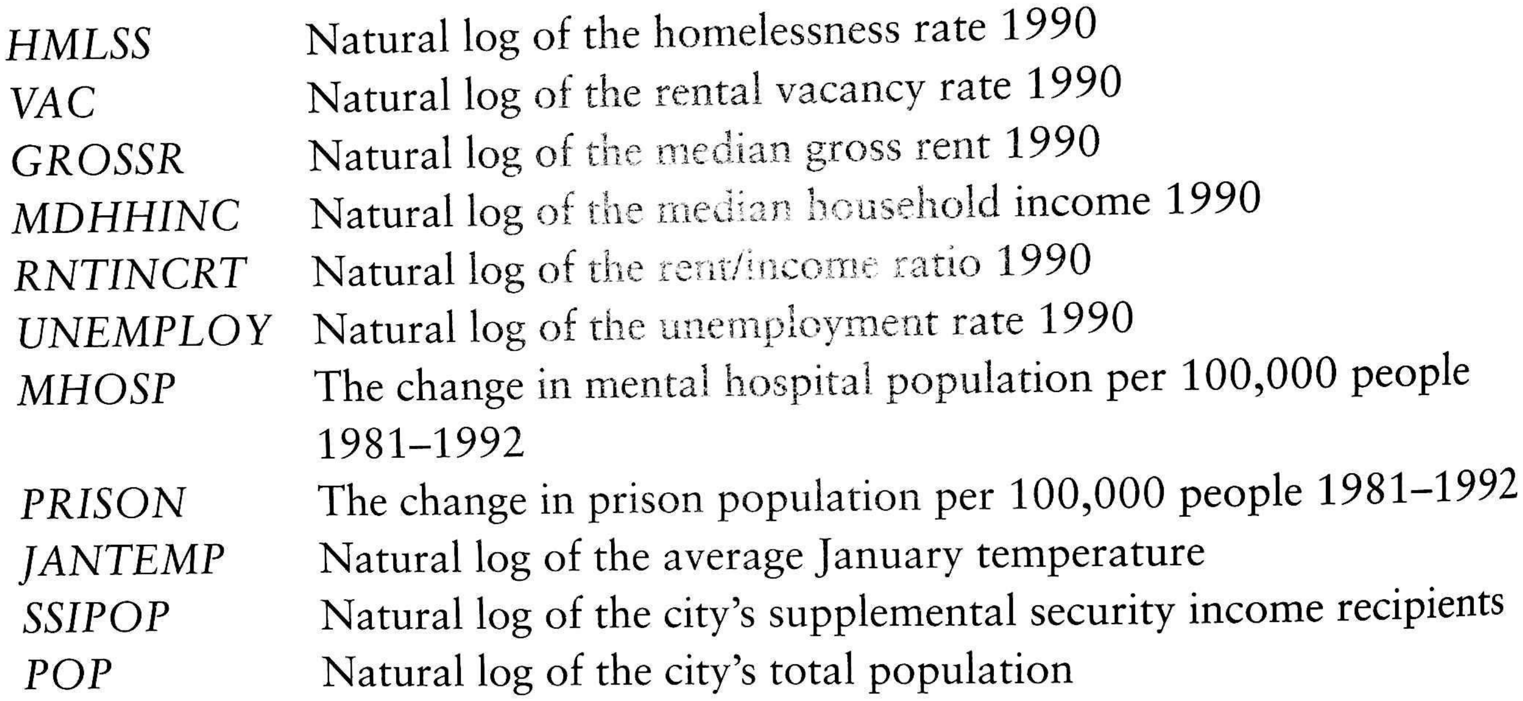


A F-distributed variable only takes on positive values. This corresponds to the fact that the sum of squared residuals can only increase if one moves from H_1 to H_0.

Choose the critical value so that the null hypothesis is rejected in, for example, 5% of the cases, although it is true.

Let’s look at another example…

- The rise of homelessness is often attributed to the release of a large number of individuals from mental hospitals during the 1980s. Housing economists, however, have painted a more complex picture, pointing to the housing market conditions and economic circumstances as significant contributors to the problem. This dataset will allow us to explore some of these factors based on data in 273 urban areas. Saved as “homeless1.dta”



HMLSS	Natural log of the homelessness rate 1990
VAC	Natural log of the rental vacancy rate 1990
GROSSR	Natural log of the median gross rent 1990
MDHHINC	Natural log of the median household income 1990
RNTINCRT	Natural log of the rent/income ratio 1990
UNEMPLOY	Natural log of the unemployment rate 1990
MHOSP	The change in mental hospital population per 100,000 people 1981–1992
PRISON	The change in prison population per 100,000 people 1981–1992
JANTEMP	Natural log of the average January temperature
SSIPOP	Natural log of the city’s supplemental security income recipients
POP	Natural log of the city’s total population

Here is our ‘full’ model

```
> mod1 = lm (hmlss ~ grossr + ssipop + unemploy +
+              mhosp , data=hl)
> summary(mod1)

Call:
lm(formula = hmlss ~ grossr + ssipop + unemploy + mhosp, data = hl)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5407 -0.3831  0.1081  0.4628  1.9459

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.121700   1.920051  -5.272 2.81e-07 ***
grossr        2.038924   0.270464   7.539 7.56e-13 ***
ssipop        0.015699   0.129272   0.121    0.903
unemploy      0.114595   0.206387   0.555    0.579
mhosp         0.005785   0.002744   2.108    0.036 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7781 on 265 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared: 0.1913,	Adjusted R-squared: 0.1791
F-statistic: 15.67 on 4 and 265 DF,  p-value: 1.597e-11
```

In this model we are using median gross rent, the number of SSI recipients in the city, the unemployment rate, and the change in the mental hospital population.

Note: all IVs and DVs are logged. When this is the case we can interpret the coefficient as the percentage change in Y given a 1% change in X

- Now, we may be surprised that unemployment and the number of individuals receiving SSI was not significant.
  - We can test their joint significance with the F-test.

```
> mod2 = lm(hmlss ~ grossr + mhosp, data=hl)
> summary(mod2)

Call:
lm(formula = hmlss ~ grossr + mhosp, data = hl)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5286 -0.3945  0.1012  0.4740  2.0093 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.975760   1.521663  -6.556 2.85e-10 ***
grossr       1.974443   0.254728   7.751 1.92e-13 ***
mhosp        0.005774   0.002729   2.115   0.0353 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.776 on 267 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared: 0.1896,	Adjusted R-squared: 0.1835 
F-statistic: 31.24 on 2 and 267 DF,  p-value: 6.468e-13
```

- Take a few minutes to you to calculate the F-statistic given the output on following slide.
- Also, look up the critical value in an F-table. Note, you will need to know the numerator and denominator degrees of freedom.

```
> anova(mod1)
Analysis of Variance Table

Response: hmlss
           Df  Sum Sq Mean Sq F value    Pr(>F)    
grossr      1  34.925  34.925 57.6809 5.299e-13 ***
ssipop      1   0.082   0.082  0.1362   0.71236    
unemploy    1   0.255   0.255  0.4210   0.51701    
mhosp       1   2.691   2.691  4.4445   0.03595 *  
Residuals 265 160.452   0.605                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(mod2)
Analysis of Variance Table

Response: hmlss
           Df  Sum Sq Mean Sq F value   Pr(>F)    
grossr      1  34.925  34.925 57.9956 4.56e-13 ***
mhosp       1   2.695   2.695  4.4748  0.03532 *  
Residuals 267 160.786   0.602                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

An alternative approach in R

```
> mod1 = lm (hmlss ~ grossr + ssipop + unemploy + mhosp , data=hl)

> mod2 = lm(hmlss ~ grossr + mhosp, data=hl)

> anova(mod2, mod1)#note, we get the exact same F-value and same result

Analysis of Variance Table

Model 1: hmlss ~ grossr + mhosp
Model 2: hmlss ~ grossr + ssipop + unemploy + mhosp
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    267 160.79
2    265 160.45  2   0.33375 0.2756 0.7593
```

- We used anova() before to get or sums of squares for our regression and to look at the overall f-test for model significance.
- When given a sequence of objects (such as to lm objects), anova tests the models against one another in the order specified.
- Here we see it is comparing our restricted model against our unrestricted model.

In class exercises

In-class Exercise: Analysis of Global COVID-19 Data

- See the end of this week’s script.
- Read in the '**cov2.csv**' data. This dataset contains information for 190 countries on a number of variables related to COVID-19 cases, deaths, and vaccinations. The data come from *https://ourworldindata.org/coronavirus*.
- The data also contains information on the population, percentage of the population over 65, gdp per capita, as well as information on a number of relevant health factors. The data are current as February 1st, 2021.
- See the associated codebook for a full description of each variable along with its associated source.