Econometrics Lecture 5

EC2METRIE

Dr. Anna Salomons

Utrecht School of Economics (U.S.E.)

12 December 2016

This class

- **Multicollinearity**
- **Heteroskedasticity**
- **Studenmund**
  - Ch 8: Multicollinearity
  - Ch 10: Running your own regression project (as background material for the project; not discussed in lecture or tutorial)

Assumptions 1-4

OLS is unbiased estimator of parameters β_k if assumptions 1-4 hold:

1. **Population model is linear in parameters** (and the error term is additive).
2. **Error term has a zero population mean**: $E(\varepsilon_i) = 0$.
3. **All independent variables are uncorrelated with the error term**: $Corr(\varepsilon_i, X_i) = 0$.
4. **No perfect (multi)collinearity** between independent variables.

Assumptions 5-6

OLS is unbiased estimator of σ^2 (and hence of $Var(\widehat{\beta_k})$) if **assumptions 1-4 hold, as well as 5-6**:

5. **No serial correlation**: $Corr(\varepsilon_i, \varepsilon_j) = 0$.
6. **No heteroskedasticity**: error term has constant variance, $Var(\varepsilon_i) = \sigma^2$ (where σ^2 is a constant).

Perfect vs imperfect (multi)collinearity

- **Perfect (multi)collinearity**: violation of assumption 4
- **Imperfect (multi)collinearity**: does not violate any assumptions, but may still be a concern.

Perfect (multi)collinearity: definition

- **Definition** of perfect (multi)collinearity: **perfect linear relationship** between 2 or more independent variables.
- In other words, the variation in one independent variable can be completely explained by movements in one or more other independent variables

A perfect linear relationship between two independent variables

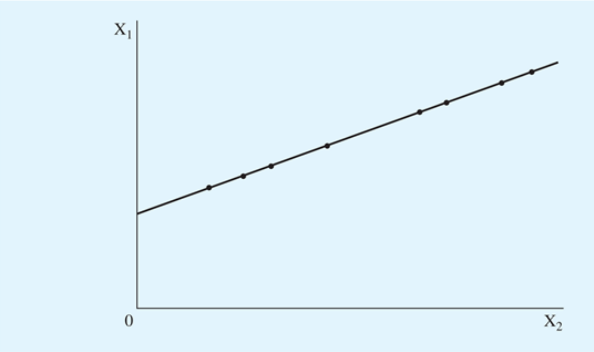


Figure 8.1 Perfect Multicollinearity

Examples of perfectly (multi)collinear independent variables

- Income in Euros (X_{1i}) and income in thousands of Euros (X_{2i}): $X_{1i} = \frac{X_{2i}}{1000}$
- Dummy for female gender ($female_i$) and dummy for male gender ($male_i$): $female_i + male_i = 1$
- Dummies for Western, Eastern, Northern and Southern region: $west_i + east_i + north_i + south_i = 1$

These variables are perfectly collinear because there is a **perfect linear relationship between them**.

Perfect (multi)collinearity: diagnosis

Diagnosis:

- 2 independent variables: **correlation coefficient = 1 or -1**
- $>$2 independent variables: **R-squared from auxiliary regression** of one independent variable on the others **is 1**.

Perfect (multi)collinearity: consequences

- **OLS** estimates of β_k are **biased (violation of assumption 4)**- in fact, OLS is incapable of generating estimates of the regression coefficients
- The partial effect of each of the collinear variables on the dependent variable cannot be calculated because the perfectly collinear variables cannot be distinguished from each other.
  - You cannot "hold all the other independent variables in the equation constant" if every time one variable changes, another changes in an identical manner!
- This is why **Stata produces an error message** if perfectly (multi-collinear) are included

Perfect (multi)collinearity: solution

- **Solution to perfect (multi)collinearity: omit one of the collinear variables**
- It **doesn't matter which one**- they are essentially identical, anyway
  - We saw this in last week's tutorial with dummies for regions.
  - For further proof, see tutorial question 2c.

Stata example

```
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------
yrsmarr         float  %9.0g                  years married

. gen monthsmarr=yrsmarr*12

. reg naffairs monthsmarr yrsmarr
note: yrsmarr omitted because of collinearity

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  1,   599) =   21.67
       Model |  227.929033     1  227.929033           Prob > F      =  0.0000
    Residual |  6301.1525    599  10.5194533           R-squared     =  0.0349
-------------+------------------------------           Adj R-squared =  0.0333
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.2434

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  monthsmarr |    .009219   .0019805     4.65   0.000     .0053294    .0131087
     yrsmarr |  (omitted)
       _cons |   .5512198   .2351106     2.34   0.019     .0894785    1.012961
------------------------------------------------------------------------------
```

Imperfect multicollinearity: definition

- **Imperfect multicollinearity** occurs when two or more explanatory variables are **imperfectly linearly related**
- Also called **multicollinearity** (note: this means imperfect, not perfect, multicollinearity!)

An imperfect linear relationship between two independent variables

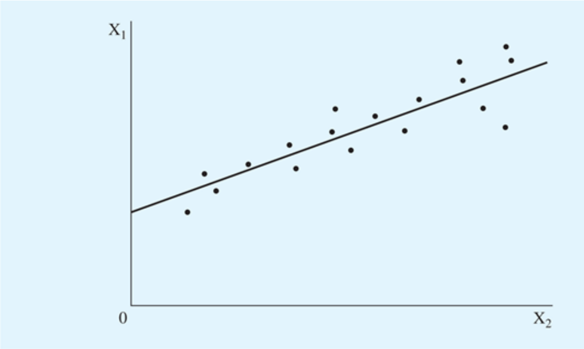


Figure 8.2 Imperfect Multicollinearity

Imperfect multicollinearity: consequences

- Estimates of β_k and σ^2 will **remain unbiased**: **none of the 6 OLS assumptions is violated**!
- BUT the **variances** (and hence standard errors) **of the estimates will increase**. This makes it more difficult to reject the null hypothesis that a particular independent variable has (cet. par.) no impact on the dependent variable.

Imperfect multicollinearity: further consequences

- Estimates can also become very sensitive to changes in specification (e.g. adding a variable; changes in the number of observations).
- The estimation of the coefficients of nonmulticollinear variables will be largely unaffected.

Example of multicollinearity

Explaining the **number of extramarital affairs** for people who cheat (i.e. at least one affair in the past year), by using how **highly people rate the quality of their marriage**, and **how many years they've been married**:

$$naffairs_i = \beta_0 + \beta_1 ratemarr_i + \beta_2 yrsmarried_i + \varepsilon_i$$

We then also **add age** to the equation:

$$naffairs_i = \beta_0 + \beta_1 ratemarr_i + \beta_2 yrsmarried_i + \beta_3 age_i + \varepsilon_i$$

Example of multicollinearity: summary statistics

variable name	storage type	display format	value label	variable label
naffairs	byte	%9.0g		number of affairs within last year
ratemarr	byte	%9.0g		5 = vry hap marr, 4 = hap than avg, 3 = avg, 2 = smewht unhap, 1 = vry unhap
yrsmarr	float	%9.0g		years married
age	float	%9.0g		in years

```
. sum naffairs ratemarr yrsmarr age if naffairs>0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
naffairs	150	5.833333	4.255934	1	12
ratemarr	150	3.446667	1.212555	1	5
yrsmarr	150	9.531947	5.187217	.125	15
age	150	33.41	8.614618	17.5	57

Example of multicollinearity: estimates of the two models

```
. reg naffairs ratemarr yrsmarr if naffairs>0

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  2,   147) =    8.98
       Model |  293.802634     2  146.901317           Prob > F      =  0.0002
    Residual |  2405.0307    147  16.3607531           R-squared     =  0.1089
-------------+------------------------------           Adj R-squared =  0.0967
       Total |  2698.83333   149  18.1129754           Root MSE      =  4.0448

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ratemarr |  -.6971104   .2782571    -2.51   0.013    -1.247011   -.1472094
     yrsmarr |   .1876488   .0650449     2.88   0.005     .0591049    .3161928
       _cons |   6.447382   1.279535     5.04   0.000     3.918721    8.976042
------------------------------------------------------------------------------

. reg naffairs ratemarr yrsmarr age if naffairs>0

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  3,   146) =    5.98
       Model |  295.133205     3   98.377735           Prob > F      =  0.0007
    Residual |  2403.70013   146  16.4636995           R-squared     =  0.1094
-------------+------------------------------           Adj R-squared =  0.0911
       Total |  2698.83333   149  18.1129754           Root MSE      =  4.0575

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ratemarr |  -.6874021   .2812124    -2.44   0.016    -1.243175   -.1316291
     yrsmarr |   .2095873   .1010581     2.07   0.040     .0098616    .4093131
         age |  -.0170266   .0598926    -0.28   0.777    -.1353952    .1013419
       _cons |   6.773664   1.721855     3.93   0.000     3.370683    10.17665
------------------------------------------------------------------------------
```

Example of multicollinearity

Estimates of the original model:

$$\underset{\text{(standard errors)}}{\widehat{naffairs_i}} = 6.45 \underset{(0.278)}{-0.70}\, ratemarr_i + \underset{(0.065)}{0.19}\, yrsmarried_i$$

When we **add the age of the respondent to the equation**:

$$\underset{\text{(standard errors)}}{\widehat{naffairs_i}} = 6.45 \underset{(0.281)}{-0.69}\, ratemarr_i + \underset{(0.101)}{0.21}\, yrsmarried_i - \underset{(0.060)}{0.02}\, age_i$$

This **increases the standard error on yrsmarried subtantially**.
Why? Because of **multicollinearity**!

Why does multicollinearity increase standard errors?

Recall the formula for the **standard error of the estimated parameter, for instance of $\widehat{\beta}_{yrsmarr}$**:

$$se(\widehat{\beta}_{yrsmarr}) = \sqrt{\frac{\widehat{\sigma^2}}{\left(1 - R^2_{yrsmarr}\right) TSS_{yrsmarr}}}$$

- $R^2_{yrsmarr}$ is the R^2 from a regression of *yrsmarr* on all other independent variables (*age* and *ratemarr*)
- $se(\widehat{\beta}_{yrsmarr})$ is higher when $R^2_{yrsmarr}$ is higher
- Multicollinearity increases $R^2_{yrsmarr}$: there is less unique variation in the variable *yrsmarr* because a large part of the variation is explained by variation in the variables *age* and *ratemarr*.

Why does multicollinearity increase standard errors?

Formula for the **standard error of** $\widehat{\beta}_{yrsmarr}$:

$$se(\widehat{\beta}_{yrsmarr}) = \sqrt{\frac{\widehat{\sigma^2}}{\left(1 - R^2_{yrsmarr}\right) TSS_{yrsmarr}}}$$

- A higher $R^2_{yrsmarr}$ makes it more difficult to find the partial effect of years of marriage on the number of affairs, i.e. the effect holding contant the age of the respondent and how the respondent rates their marriage.
- Therefore, the standard error of $\widehat{\beta}_{yrsmarr}$ is higher.

Interlude: the determinants of the standard error revisited

The standard error for the estimated coefficient on years of marriage is:

$$se(\widehat{\beta}_{yrsmarr}) = \sqrt{\frac{\widehat{\sigma^2}}{\left(1 - R^2_{yrsmarr}\right) TSS_{yrsmarr}}}$$

The standard error of $\widehat{\beta}_{yrsmarr}$ is larger:

- When $R^2_{yrsmarr}$ is larger (this is where multicollinearity has an effect)
- When $TSS_{yrsmarr}$ is smaller
- When $\widehat{\sigma^2}$ is larger

See lecture of week 2 for explanation on each of these.

Multicollinearity: diagnosis

Almost always have some correlation between independent variables: i.e. it's **not a matter of whether there is any multicollinearity, but how much**. To diagnose this:

- Calculate the **correlations between independent variables**: higher correlations imply more multicollinearity
- Estimate an **auxiliary regression** of each independent variable on the other independent variables: a higher R_k^2 indicates more multicollinearity (a higher **variance inflation factor** $\frac{1}{1-R_k^2}$).

Example of multicollinearity: diagnosis

```
. corr yrsmarr age ratemarr if naffairs>0
(obs=150)

             |  yrsmarr      age ratemarr
-------------+---------------------------
     yrsmarr |   1.0000
         age |   0.7607   1.0000
    ratemarr |  -0.1883  -0.0658   1.0000

. reg yrsmarr age ratemarr if naffairs>0

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  2,   147) =  109.29
       Model |  2397.10137     2  1198.55068           Prob > F      =  0.0000
    Residual |  1612.07456   147  10.9664936           R-squared     =  0.5979
-------------+------------------------------           Adj R-squared =  0.5924
       Total |  4009.17593   149   26.907221           Root MSE      =  3.3116

------------------------------------------------------------------------------
     yrsmarr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4525661   .0315608    14.34   0.000     .3901946    .5149376
    ratemarr |  -.5938604   .2242242    -2.65   0.009     -1.03698   -.1507411
       _cons |  -3.541448   1.374602    -2.58   0.011    -6.257981   -.8249143
------------------------------------------------------------------------------
```

Example of multicollinearity: diagnosis

```
. reg yrsmarr age ratemarr if naffairs>0

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  2,   147) =  109.29
       Model |  2397.10137     2  1198.55068           Prob > F      =  0.0000
    Residual |  1612.07456   147  10.9664936           R-squared     =  0.5979
-------------+------------------------------           Adj R-squared =  0.5924
       Total |  4009.17593   149   26.907221           Root MSE      =  3.3116

------------------------------------------------------------------------------
     yrsmarr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4525661   .0315608    14.34   0.000     .3901946    .5149376
    ratemarr |  -.5938604   .2242242    -2.65   0.009     -1.03698   -.1507411
       _cons |  -3.541448   1.374602    -2.58   0.011    -6.257981   -.8249143
------------------------------------------------------------------------------

. reg yrsmarr ratemarr if naffairs>0

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =    5.44
       Model |  142.157285     1  142.157285           Prob > F      =  0.0210
    Residual |  3867.01864   148  26.1285043           R-squared     =  0.0355
-------------+------------------------------           Adj R-squared =  0.0289
       Total |  4009.17593   149   26.907221           Root MSE      =  5.1116

------------------------------------------------------------------------------
     yrsmarr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ratemarr |   -.805545   .3453524    -2.33   0.021    -1.488004   -.1230863
       _cons |   12.30839   1.261364     9.76   0.000     9.815782      14.801
------------------------------------------------------------------------------
```

Example of multicollinearity: diagnosis

- **Years of marriage and age are strongly correlated,** $r_{yrsmarr,age} = 0.76$- they are (imperfectly) collinear.
- The R^2 of an **auxiliary regression** of years of marriage on age and the quality of the marriage is 0.5979.
  - In contrast, the R^2 of an auxiliary regression of years of marriage on only the quality of the marriage is 0.0355: this shows that the model without age had much less multicollinearity than the model with age.

Imperfect (multi)collinearity: solutions

3 different solutions:

1. **Do nothing**:

    a Multicollinearity will not necessarily increase standard errors enough to makes estimates statistically insignificant and/or change the estimated coefficients to make them differ from expectations.

    b The deletion of a multicollinear variable that belongs in an equation (according to economic theory) will cause omitted variable bias.

Imperfect multicollinearity: solutions

2. If there are (many) insignificant estimates, consider if you can **drop a redundant variable**:

   a Viable strategy when two variables measure essentially the same thing.

   b Always use theory as the basis for this decision! I.e. never drop if a variable if this is not justified by economic theory.

Imperfect multicollinearity: solutions

3. **Increase the sample size** (i.e. gather more data):

   a This is frequently impossible but a useful alternative if feasible.

   b The idea is that the larger sample will reduce the standard errors of the estimated coefficients (since it increases TSS_k), counteracting the impact of multicollinearity.

Sports economics

Sports economics is a sub-field of economics which analyzes the design and outcomes of sports competitions, labor markets for athletes, economic effects of large sports events, etc. Some interesting findings come out of this literature e.g.:

- Are Big-Time Sports a Threat to Student Achievement?
  https://www.aeaweb.org/articles?id=10.1257/app.4.4.254
- Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior
  http://qje.oxfordjournals.org/content/early/2011/03/21/qje.qjr001
- Game Theory and Major League Sports
  http://www.nber.org/digest/oct09/w15347.html
- Work and Play: International Evidence of Gender Equality in Employment and Sports
  http://jse.sagepub.com/content/5/3/227.abstract

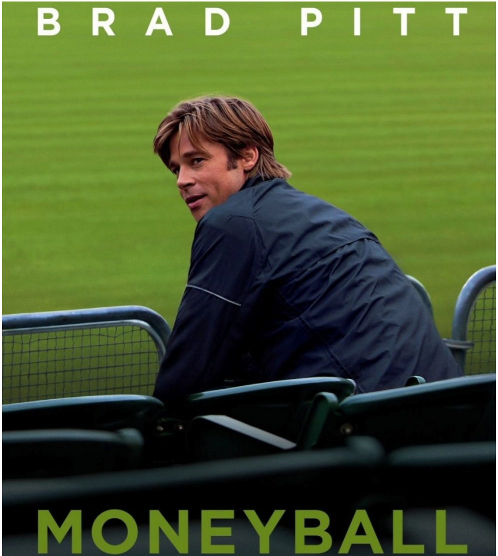
From the Journal of Economic Perspectives (2006)

An Economic Evaluation of the *Moneyball* Hypothesis

Jahn K. Hakes and Raymond D. Sauer

In his 2003 book *Moneyball*, financial reporter Michael Lewis made a striking claim: the valuation of skills in the market for baseball players was grossly inefficient. The discrepancy was so large that when the Oakland Athletics hired an unlikely management group consisting of Billy Beane, a former player with mediocre talent, and two quantitative analysts, the team was able to exploit this inefficiency and outproduce most of the competition, while operating on a shoe-string budget.

Another example of multicollinearity: Moneyball



Performance pay in major league baseball

- Estimate the following model, relating the **log salary of major league baseball players** to the number years they played in the major league (measuring their experience) and how many games they play per year (measuring their hours worked):

$$\ln salary_i = \beta_0 + \beta_1 years_i + \beta_2 gamesyr_i + \varepsilon_i$$

- We then **add the performance variables** *batting average*, *number of home runs per year* and *number of in runs per year*:

$$\begin{aligned} \ln salary_i &= \beta_0 + \beta_1 years_i + \beta_2 gamesyr_i + \beta_3 bavg_i \\ &\quad + \beta_4 hrunsyr_i + \beta_5 rbisyr_i + \varepsilon_i \end{aligned}$$

Summary statistics

variable name	storage type	display format	value label	variable label
salary	float	%9.0g		1993 season salary
lsalary	float	%9.0g		log(salary)
years	byte	%9.0g		years in major leagues
gamesyr	float	%9.0g		games per year in league
bavg	float	%9.0g		career batting average
hrunsyr	float	%9.0g		home runs per year
rbisyr	float	%9.0g		runs batted in per year

```
. sum salary lsalary years gamesyr bavg hrunsyr rbisyr
```

Variable	Obs	Mean	Std. Dev.	Min	Max
salary	353	1345672	1407352	109000	6329213
lsalary	353	13.49218	1.182466	11.5991	15.66069
years	353	6.325779	3.880142	1	20
gamesyr	353	90.07604	36.13248	5.5	159
bavg	353	258.9858	38.4224	111	625
hrunsyr	353	7.119053	6.796919	0	31.42857
rbisyr	353	35.05021	22.82877	.5	97.625

Estimates of model without performance variables

```
. reg lsalary years gamesyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  2,   350) =  259.32
       Model |  293.864058     2  146.932029           Prob > F      =  0.0000
    Residual |  198.311477   350  .566604221           R-squared     =  0.5971
-------------+------------------------------           Adj R-squared =  0.5948
       Total |  492.175535   352  1.39822595           Root MSE      =  .75273

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |    .071318    .012505     5.70   0.000     .0467236    .0959124
     gamesyr |   .0201745   .0013429    15.02   0.000     .0175334    .0228156
       _cons |    11.2238    .108312   103.62   0.000     11.01078    11.43683
------------------------------------------------------------------------------
```

Estimates of model with performance variables

```
. reg lsalary years gamesyr bavg hrunsyr rbisyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  5,   347) =  117.06
       Model |  308.989208     5  61.7978416           Prob > F      =  0.0000
    Residual |  183.186327   347  .527914487           R-squared     =  0.6278
-------------+------------------------------           Adj R-squared =  0.6224
       Total |  492.175535   352  1.39822595           Root MSE      =  .72658

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   .0688626   .0121145     5.68   0.000     .0450355    .0926898
     gamesyr |   .0125521   .0026468     4.74   0.000     .0073464    .0177578
        bavg |   .0009786   .0011035     0.89   0.376    -.0011918     .003149
     hrunsyr |   .0144295    .016057     0.90   0.369    -.0171518    .0460107
      rbisyr |   .0107657    .007175     1.50   0.134    -.0033462    .0248776
       _cons |   11.19242   .2888229    38.75   0.000     10.62435    11.76048
------------------------------------------------------------------------------
```

The OLS estimates for the performance variables *bavg*, *hrunsyr* and *rbisyr* are **individually statistically insignificant**. Does this automatically mean **performance is not important for wages?** No- **standard errors could be inflated due to multicollinearity**.

Diagnosing multicollinearity: correlations among explanatory variables

```
. corr years gamesyr bavg hrunsyr rbisyr
(obs=353)
```

	years	gamesyr	bavg	hrunsyr	rbisyr
years	1.0000				
gamesyr	0.5624	1.0000			
bavg	0.1973	0.3191	1.0000		
hrunsyr	0.3802	0.6138	0.1906	1.0000	
rbisyr	0.4871	0.8487	0.3291	0.8907	1.0000

Diagnosing multicollinearity: some auxiliary regressions

```
. reg rbisyr years gamesyr bavg hrunsyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  4,   348) = 1469.33
       Model |   173190.97     4  43297.7426           Prob > F      =  0.0000
    Residual |  10254.7377   348   29.467637           R-squared     =  0.9441
-------------+------------------------------           Adj R-squared =  0.9435
       Total |  183445.708   352   521.15258           Root MSE      =  5.4284

------------------------------------------------------------------------------
      rbisyr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |  -.1049775   .0903352    -1.16   0.246    -.2826491    .0726941
     gamesyr |   .2979218   .0116611    25.55   0.000     .2749867    .3208569
        bavg |   .0408626   .0079482     5.14   0.000       .02523    .0564953
     hrunsyr |   1.998371   .0540007    37.01   0.000     1.892162     2.10458
       _cons |   -15.9307   1.981683    -8.04   0.000    -19.82828   -12.03312
------------------------------------------------------------------------------

. reg hrunsyr years gamesyr bavg rbisyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  4,   348) =  603.96
       Model |  14214.1793     4  3553.54483           Prob > F      =  0.0000
    Residual |  2047.55686   348  5.88378407           R-squared     =  0.8741
-------------+------------------------------           Adj R-squared =  0.8726
       Total |  16261.7362   352  46.1981141           Root MSE      =  2.4257

------------------------------------------------------------------------------
     hrunsyr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   .0601011   .0403154     1.49   0.137    -.0191914    .1393937
     gamesyr |  -.0964951   .0071638   -13.47   0.000    -.1105849   -.0824053
        bavg |  -.0165526   .0035756    -4.63   0.000    -.0235851   -.0095202
      rbisyr |   .3990134   .0107823    37.01   0.000     .3778068    .4202201
       _cons |   5.732154   .9139532     6.27   0.000     3.934587    7.529721
------------------------------------------------------------------------------
```

Solution for multicollinearity

- **In this example, we would not want to exclude any of the collinear variables** since economic theory tells us they should all have an impact on salary- the signs on the coefficients are also as we would expect.
- Hence, our "solution" would be to **do nothing**.
- However, we can of course perform an **F-test for joint significance of the performance variables** to examine whether performance matters for pay in major league baseball.

F-test shows that performance does matter for pay

```
. reg lsalary years gamesyr bavg hrunsyr rbisyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  5,   347) =  117.06
       Model |  308.989208     5  61.7978416           Prob > F      =  0.0000
    Residual |  183.186327   347  .527914487           R-squared     =  0.6278
-------------+------------------------------           Adj R-squared =  0.6224
       Total |  492.175535   352  1.39822595           Root MSE      =  .72658

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   .0688626   .0121145     5.68   0.000     .0450355    .0926898
     gamesyr |   .0125521   .0026468     4.74   0.000     .0073464    .0177578
        bavg |   .0009786   .0011035     0.89   0.376    -.0011918     .003149
     hrunsyr |   .0144295    .016057     0.90   0.369    -.0171518    .0460107
      rbisyr |   .0107657    .007175     1.50   0.134    -.0033462    .0248776
       _cons |   11.19242   .2888229    38.75   0.000     10.62435    11.76048
------------------------------------------------------------------------------

. test bavg hrunsyr rbisyr

 ( 1)  bavg = 0
 ( 2)  hrunsyr = 0
 ( 3)  rbisyr = 0

       F(  3,   347) =    9.55
            Prob > F =    0.0000
```

Summary: multicollinearity

- **Disease** = imperfect multicollinearity
- **Consequence** = estimates of β_k and of $Var(\widehat{\beta_k})$ remain unbiased (since no OLS assumption has been violated), but the estimated $Var(\widehat{\beta_k})$ is larger (i.e. larger standard errors)
- **Diagnosis** = examine correlations among regressors; estimate auxiliary regressions
- **Solution** = do nothing (only drop one of the highly correlated variables if economic theory justifies this)

Assumptions 1-4

OLS is unbiased estimator of parameters β if assumptions 1-4 hold:

1. **Population model is linear in parameters** (and the error term is additive).
2. **Error term has a zero population mean**: $E(\varepsilon_i) = 0$.
3. **All independent variables are uncorrelated with the error term**: $Corr(\varepsilon_i, X_k) = 0$.
4. **No perfect (multi)collinearity** between independent variables.

Assumptions 5-6

OLS is unbiased estimator of $Var(\widehat{\beta})$ if **assumptions 1-4 hold, as well as 5-6**:

5. **No serial correlation**: $Corr(\varepsilon_i, \varepsilon_j) = 0$.
6. **No heteroskedasticity**: error term has constant variance, $Var(\varepsilon_i) = \sigma^2$ (where σ^2 is a constant).

Hypothesis testing

- **When assumptions 1-6 are met, we can perform hypothesis tests about a single population parameter using the t-test**. Specifically, under H_0 the test statistic follows a t-distribution:
  - For large sample sizes.
  - For small sample sizes, if we additionally assume normality of the error term ε.
- **When assumptions 1-6 are met, we can perform hypothesis tests about multiple population parameters using the F-test**. Specifically, under H_0 the test statistic follows an F-distribution:
  - For large sample sizes.
  - For small sample sizes, if we additionally assume normality of the error term ε.

Heteroskedasticity: violation of assumption 6

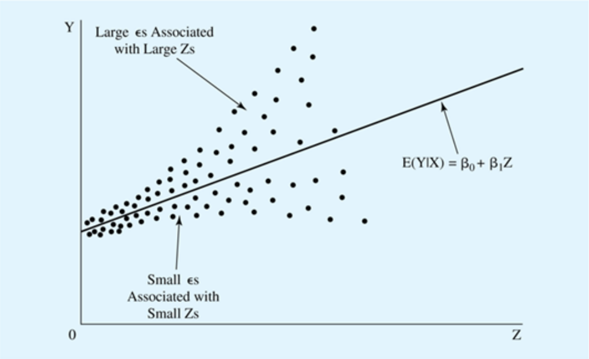
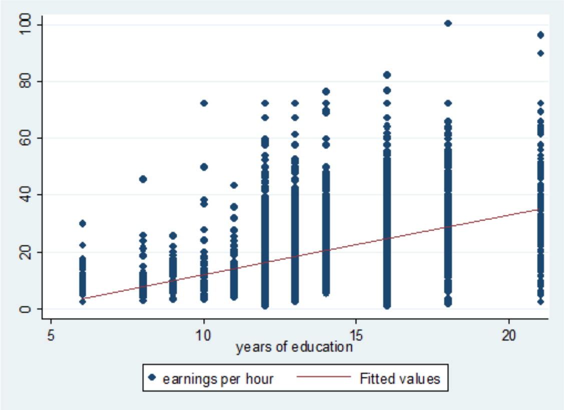


Figure 4.2 An Error Term Whose Variance Increases as Z Increases (Heteroskedasticity)

Heteroskedasticity: violation of assumption 6



Heteroskedasticity: consequences

Assumption 6: error term has constant variance, $Var(\varepsilon_i) = \sigma^2$.
Consequences of violation of this assumption:

- OLS is still an unbiased estimator of β_k (since assumptions 1-4 are not violated)
- But since $Var(\widehat{\beta}_k)$ depends on σ^2, **it is a biased estimator of** $Var(\widehat{\beta}_k)$
- **t-statistics are incorrect** since these depend on σ^2
- **F-statistics are incorrect** since these depend on σ^2

Heteroskedasticity: consequences

- If t- and F-statistics are incorrect, we **cannot perform hypothesis tests**!
- **Without hypothesis tests, we cannot perform inference** about the population from a sample, which is the aim of applied econometric analysis!
- Therefore, we need to know how to diagnose heteroskedasticity and then solve the problem if we find any.

Heteroskedasticity: a closer look

Assumption 6: **homoskedasticity**, which means the error term has constant variance, $Var(\varepsilon_i) = \sigma^2$.

- A constant error variance implies that σ^2 does not depend on any of the independent variables $X_1, X_2, .., X_k$
- OLS estimates the error variance σ^2 as the residual sum of squares divided by the number of degrees of freedom:

$$\widehat{\sigma^2} = \frac{\sum e_i^2}{n-k-1} = \frac{e_1^2 + e_2^2 + ... + e_n^2}{n-k-1}$$

- e_i^2 gives the contribution of the i^{th} residual to the estimated error variance

Heteroskedasticity: a closer look

Combining insights from the previous slide, we get that **under homoskedasticity**:

- σ^2 does not depend on any of the independent variables $X_1, X_2, .., X_k$
- $\widehat{\sigma^2}$ does not depend on any of the independent variables $X_1, X_2, .., X_k$
- e_i^2 does not depend on any of the independent variables $X_1, X_2, .., X_k$: we can **use this to diagnose heteroskedasticity**.

Heteroskedasticity: diagnosis with the Breusch-Pagan test

1. **Estimate the model** :

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

2. **Predict residuals** e_i from the estimated model $Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + e_i$, **and square them** (e_i^2)
3. **Regress squared residuals** e_i^2 **on independent variables** from the original model

$$e_i^2 = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + \nu_i$$

4. Test whether **the independent variables have a jointly significant impact** on e_i^2: if they do (i.e. H_0 is rejected), we have **heteroskedasticity**.

$$\begin{aligned} H_0 &: \delta_1 = \delta_2 = 0 && \textit{(homoskedasticity)} \\ H_A &: H_0 \text{ not true} && \textit{(heteroskedasticity)} \end{aligned}$$

Heteroskedasticity: diagnosis with the White test

Note: only step 3 is different from Breusch-Pagan test– on the exam, the White test will not be asked.

1. **Estimate the model** : $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$
2. **Predict residuals** e_i from the estimated model, **and square them**
3. **Regress squared residuals e_i^2 on independent variables, their squares and interaction terms:**

$$\begin{aligned} e_i^2 &= \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{1i}^2 + \delta_4 X_{2i}^2 \\ &\quad + \delta_5 X_{1i} X_{2i} + \nu_i \end{aligned}$$

4. Test whether **the independent variables have a jointly significant impact** on e_i^2: if they do (i.e. H_0 is rejected), we have **heteroskedasticity**.

$$\begin{aligned} H_0 &: \delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = 0 \qquad \textit{(homosk.)} \\ H_A &: H_0 \text{ not true} \qquad \textit{(heterosk.)} \end{aligned}$$

Example of the Breusch-Pagan test

We want to examine the **relationship between economic development**, measured as log gdp per capita, **workers' education level,** and **entrepreneurship** (measured as the fraction of the working age population in self-employment).

Because we expect entrepreneurship to be nonlinearly related to development, we estimate the following model:

$$\ln gdp_i = \beta_0 + \beta_1 educ_i + \beta_2 selfemp_i + \beta_3 selfemp_i^2 + \varepsilon_i$$

Example of the Breusch-Pagan test

```
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
lnreggdp        float  %8.0g                  Log of gdp per capita for 547 different regions
                                                in 35 countries
yearsed         float  %8.0g                  Years of education
self_emp        float  %8.0g                  Percentage of self-employed in the working age
                                                population
self_emp2       float  %9.0g                  self_emp squared

. sum lnreggdp yearsed self_emp self_emp2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    lnreggdp |       547    8.982509    1.168632    6.08819   11.87397
     yearsed |       547    6.914256    2.889793   1.390702   12.83251
    self_emp |       547    21.77006    17.29819          0   77.34605
   self_emp2 |       547     772.616    1045.851          0   5982.412
```

Example of the Breusch-Pagan test: steps 1 & 2

Estimates of the model:

```
. reg lnreggdp yearsed self_emp self_emp2

      Source |       SS       df       MS              Number of obs =     547
-------------+------------------------------           F(  3,   543) =  750.25
       Model |  600.742478     3  200.247493           Prob > F      =  0.0000
    Residual |  144.930118   543  .266906294           R-squared     =  0.8056
-------------+------------------------------           Adj R-squared =  0.8046
       Total |  745.672595   546  1.36570072           Root MSE      =  .51663

------------------------------------------------------------------------------
    lnreggdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsed |   .3447389   .0093812    36.75   0.000      .326311    .3631669
    self_emp |   .0235793   .0038133     6.18   0.000     .0160886    .0310699
   self_emp2 |  -.0005308    .000059    -9.00   0.000    -.0006467   -.0004149
       _cons |   6.495673   .0993675    65.37   0.000     6.300481    6.690865
------------------------------------------------------------------------------

. predict uhat, resid

. gen uhat2=uhat^2
```

We want to test for heteroskedasticity, so we **predict the residuals** (e_i), and then **obtain the squared residuals** (e_i^2).

Example of the Breusch-Pagan test: steps 3 & 4

We now **regress the squared residuals onto the explanatory variables from the original model**:

```
. reg uhat2 yearsed self_emp self_emp2

      Source |       SS           df       MS      Number of obs   =       547
-------------+----------------------------------   F(3, 543)       =      6.49
       Model |  3.9418781         3  1.31395937   Prob > F        =    0.0003
    Residual |  109.888897       543  .202373659   R-squared       =    0.0346
-------------+----------------------------------   Adj R-squared   =    0.0293
       Total |  113.830775       546  .208481273   Root MSE        =    .44986

------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsed |  -.0325959   .0081688    -3.99   0.000    -.0486422   -.0165496
    self_emp |  -.0018068   .0033205    -0.54   0.587    -.0083294    .0047157
   self_emp2 |   .0000143   .0000514     0.28   0.781    -.0000866    .0001152
       _cons |   .5186509   .0865251     5.99   0.000     .3486859    .6886158
------------------------------------------------------------------------------
```

The **explanatory variables are jointly significant**, as seen from the model F-test (p-value$=0.0003<0.05$). This means we **reject the null hypothesis of homoskedasticity: the errors are heteroskedastic**!

Heteroskedasticity: solution

- The solution for heteroskedasticity **does not require changing the estimates** $\widehat{\beta}_k$ (since OLS is still an unbiased estimator of β_k).

- However, we do **need to change our standard errors** since the $\widehat{Var}(\widehat{\beta}_k)$ are incorrect.

Heteroskedasticity: solution

- We therefore calculate the **heteroskedasticity-robust standard error** (also known as White standard error)

$$\widehat{Var}(\widehat{\beta}_k) = \frac{\sum e_i^2 \widehat{r}_{ik}^2}{(RSS_k)^2}$$

  where $\widehat{r}_{ik}$ is the residual for observation i from a regression of X_k on all other explanatory variables and RSS_k is the residual sum of squares from a regression of X_k on all other explanatory variables.

- Caveat: this robust standard error is **only valid in large samples**![1]

[1] Adjustments for small samples are available (e.g. using the command *hc2* or *hc3* instead of *robust*) but are not part of the exam material for this course.

Heteroskedasticity-robust standard errors in Stata

```
. reg lnreggdp yearsed self_emp self_emp2, robust

Linear regression                               Number of obs =      547
                                                F(  3,   543) = 1081.16
                                                Prob > F      =   0.0000
                                                R-squared     =   0.8056
                                                Root MSE      =   .51663
```

lnreggdp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
yearsed	.3447389	.0115737	29.79	0.000	.3220042	.3674736
self_emp	.0235793	.0047807	4.93	0.000	.0141884	.0329702
self_emp2	-.0005308	.000067	-7.92	0.000	-.0006624	-.0003992
_cons	6.495673	.1396055	46.53	0.000	6.22144	6.769906

This is can obtained easily in Stata by typing *,robust* at the end of the *reg* command.

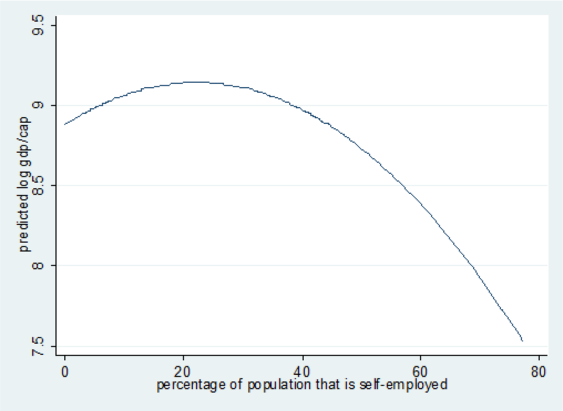
Comparing regular and robust standard errors

```
. reg lnreggdp yearsed self_emp self_emp2

      Source |       SS       df       MS              Number of obs =     547
-------------+------------------------------           F(  3,   543) =  750.25
       Model |  600.742478     3  200.247493           Prob > F      =  0.0000
    Residual |  144.930118   543  .266906294           R-squared     =  0.8056
-------------+------------------------------           Adj R-squared =  0.8046
       Total |  745.672595   546  1.36570072           Root MSE      =  .51663

------------------------------------------------------------------------------
    lnreggdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsed |   .3447389   .0093812    36.75   0.000      .326311    .3631669
    self_emp |   .0235793   .0038133     6.18   0.000     .0160886    .0310699
   self_emp2 |  -.0005308    .000059    -9.00   0.000    -.0006467   -.0004149
       _cons |   6.495673   .0993675    65.37   0.000     6.300481    6.690865
------------------------------------------------------------------------------

. reg lnreggdp yearsed self_emp self_emp2, robust

Linear regression                                      Number of obs =     547
                                                       F(  3,   543) = 1081.16
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.8056
                                                       Root MSE      =  .51663

------------------------------------------------------------------------------
             |               Robust
    lnreggdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsed |   .3447389   .0115737    29.79   0.000     .3220042    .3674736
    self_emp |   .0235793   .0047807     4.93   0.000     .0141884    .0329702
   self_emp2 |  -.0005308    .000067    -7.92   0.000    -.0006624   -.0003992
       _cons |   6.495673   .1396055    46.53   0.000      6.22144    6.769906
------------------------------------------------------------------------------
```

Comparing regular and robust standard errors

- **Robust standard errors are typically higher** than the regular ones- although they may also be lower.
- Higher standard errors means the t-statistics become smaller (in absolute value), and **estimates become less significant**.
- **In our example, the standard errors increase somewhat**, but all coefficients are still individually significant.

Sidenote: the relationship between entrepreneurship and development



A note on pure vs impure heteroskedasticity

- Some versions of Studenmund discuss that heteroskedasticity can result from a misspecified model (e.g. an omitted variable)- this is called **impure heteroskedasticity**.
- However, there are better ways to test the specification than to rely on heteroskedasticity as a symptom for misspecification, especially since we can also have heteroskedasticity in a correctly specified model- this is called **pure heteroskedasticity**.

A note on pure vs impure heteroskedasticity

- The approach we take in this course (and in the project paper), is **first to make the specification as good as possible** (weeks 1-4 of this course) **and then check for heteroskedasticity**.
- Therefore, you **do not need to discuss impure heteroskedasticity** in your paper or study it for the exam. Since we have checked the specification beforehand, we assume all found heteroskedasticity is pure.

Another example: pricing in an illegal market (from last week's lecture)

```
. reg lnprice attractive school age rich alcohol bar street

      Source |       SS       df       MS              Number of obs =    3016
-------------+------------------------------           F(  7,  3008) =  206.91
       Model |  501.703241     7  71.6718916           Prob > F      =  0.0000
    Residual |   1041.9644  3008   .34639774           R-squared     =  0.3250
-------------+------------------------------           Adj R-squared =  0.3234
       Total |  1543.66764  3015  .511995901           Root MSE      =  .58856

------------------------------------------------------------------------------
     lnprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  attractive |   .2394121   .0315921     7.58   0.000     .1774678    .3013563
      school |   .1637754   .0238151     6.88   0.000       .11708    .2104709
         age |  -.0210136   .0014531   -14.46   0.000    -.0238627   -.0181645
        rich |   .2924201   .0304404     9.61   0.000      .232734    .3521061
     alcohol |   .2403329   .0358481     6.70   0.000     .1700436    .3106222
         bar |   .2160627   .0785642     2.75   0.006     .0620178    .3701076
      street |  -.2621039   .0793876    -3.30   0.001    -.4177633   -.1064444
       _cons |   5.752484   .0912836    63.02   0.000       5.5735    5.931469
------------------------------------------------------------------------------

. predict uhat, resid

. gen uhat2=uhat^2
```

Another example: pricing in an illegal market

```
. reg uhat2 attractive school age rich alcohol bar street

      Source |       SS       df       MS              Number of obs =    3016
-------------+------------------------------           F(  7,  3008) =    8.79
       Model |  23.491717     7  3.35595957           Prob > F      =  0.0000
    Residual | 1148.15068  3008  .381699028           R-squared     =  0.0201
-------------+------------------------------           Adj R-squared =  0.0178
       Total | 1171.64239  3015  .388604442           Root MSE      =  .61782

------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  attractive |   .1954101   .0331628     5.89   0.000      .130386    .2604341
      school |   .0457579   .0249991     1.83   0.067    -.0032592     .094775
         age |  -.0038014   .0015253    -2.49   0.013    -.0067922   -.0008107
        rich |   .0298769   .0319538     0.94   0.350    -.0327767    .0925304
     alcohol |  -.0653703   .0376304    -1.74   0.082    -.1391543    .0084137
         bar |  -.1877774   .0824703    -2.28   0.023    -.3494813   -.0260735
      street |  -.1896974   .0833346    -2.28   0.023     -.353096   -.0262988
       _cons |   .6226457   .0958221     6.50   0.000     .4347622    .8105292
------------------------------------------------------------------------------
```

Another example: pricing in an illegal market

```
. reg lnprice attractive school age rich alcohol bar street, robust

Linear regression                               Number of obs =     3016
                                                F(  7,   3008) =   229.26
                                                Prob > F      =   0.0000
                                                R-squared     =   0.3250
                                                Root MSE      =   .58856
```

lnprice	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
attractive	.2394121	.0374183	6.40	0.000	.166044	.3127802
school	.1637754	.0243654	6.72	0.000	.1160009	.21155
age	-.0210136	.0013033	-16.12	0.000	-.0235691	-.0184581
rich	.2924201	.0296154	9.87	0.000	.2343515	.3504886
alcohol	.2403329	.0376984	6.38	0.000	.1664156	.3142502
bar	.2160627	.0961353	2.25	0.025	.0275651	.4045602
street	-.2621039	.0967843	-2.71	0.007	-.4518739	-.0723338
_cons	5.752484	.105823	54.36	0.000	5.544991	5.959977

Another example: pricing in an illegal market

After correcting for heteroskedastic errors:

- All estimated coefficients remain the same (this is always the case!).
- Standard errors for *attractive, school, alcohol, bar, street* increased.
- Standard errors for *age, rich* decreased.

Summary: heteroskedasticity

- **Disease** = heteroskedastic errors
- **Consequence** = coefficient estimates $\widehat{\beta}$ remain unbiased (since OLS assumptions 1-4 have not been violated), but the variance estimates $\widehat{Var}(\widehat{\beta})$ (and hence also the std errors $\sqrt{\widehat{Var}(\widehat{\beta})}$) are biased (since OLS assumption 6 has been violated). This means we cannot perform hypothesis tests (t- or F-tests).
- **Diagnosis** = Breusch-Pagan test, which involves regressing the squared residuals on all explanatory variables (there is heteroskedasticity if the p-value for the model F-test is smaller than the chosen significance level).
- **Solution** = estimate the equation with heteroskedasticity-robust standard errors (Stata command *reg y x1 x2, robust*)

Project paper

- Consider to what extent there is multicollinearity in your final (preferred) model.
- Test for heteroskedasticity in your final (preferred) specification.
- If you find any heteroskedasticity, correct for it and evaluate whether any conclusions about statistical significance are changed.
- Finish up any other parts of the cross-sectional analysis (i.e. material from weeks 1-4): next week, we move on to timeseries!