Statistical Reasoning

Week 11

Sciences Po - Louis de Charsonville

Spring 2018

Outline

Research Paper

Research Paper

Timeline

Final draft	**28 April**

Paper

Descriptive and inferential bivariate statistics

Describe the Relationship

- If both variables are continuous : look at the correlation using a scatterplot, Pearson's ρ (`pwcorr`)
- If you can calculate a mean on the DV (continuous or ordinal), and the IV is a dummy or categorical : compare means using `bysort` bysort and `ttest` of means ;
- If both variables are categorical : look at the percentage distribution with cross-tabulation, and look at Cramer's V for strength ;
- Make a lot of graphs (but include only the relevant ones in your paper). *"You miss 100% of the plots you don't make"*

Test for significance

- Look at p-values of each kind of test (ttest, chi2...)

Don't confuse strengh and significance

- Cramer's V and Pearson's R are **not** statistical tests, but tell you the **strength** of the association ;
- Chi-2 and t-tests are statistical tests : they tell you whether the relationship is **significant** or not ;
- `pwcorr` command with option `sig` or `star` provides both : Pearson's ρ and significance of the correlation
- R^2 = explanatory power of the predictor variables
- The p-values associated to the coefficients in the regression model is about the statistical significance of the predictor.

Do-File Requirements

Major requirements

- **The entire code should run without errors**. Run your entire do-file before submitting.
- Code should be commented and divided in three headings separating 1^{st}, 2^{nd} drafts and final paper.
- At least one multiple linear regression.
- Name **all graphs**

Minor requirements

- Comment `browse` and `lookfor` commands.
- Include your version at the beginning of the code, e.g. `version 14`. Some commands has changed, like `ci`, and will generate errors in some versions of Stata.

Paper requirements 1/2

Required structure

1. Abstract
2. Introduction
3. Theory and Hypotheseses
4. Data and Methods
5. Results
6. Discussion
7. Conclusion
8. Appendix
9. Bibliography

Paper requirements 2/2

Essential instructions

- Review **paper template**
- Respect **paragraph limits** mentioned in the template
- Improve **formatting** (styles, citations fonts...)

3-step guide

- **Rewrite** from top to bottom
- **Select** what you report
- **Balance** evidence and analysis

Abstract

- Abstract should communicate your **central contribution**
- Abstract must be concrete
- Say **what you find**, not what you look for

Example

"Using centuries of Nile flood data, I document that during deviant Nile floods, Egypt's highest-ranking religious authority was less likely to be replaced and relative allocations to religious structures increased. These findings are consistent with historical evidence that Nile shocks increased this authority's political influence by raising the probability he could coordinate a revolt. I find that the available data provide support for this interpretation and weigh against some of the most plausible alternatives. For example, I show that while Nile shocks increased historical references to social unrest, deviant floods did not increase a proxy for popular religiosity. Together, the results suggest an increase in the political power of religious leaders during periods of economic downturn."

Chaney, Eric. 2013. "Revolt on the Nile : Economic Shocks, Religion, and Political Power." Econometrica 81 (5)

Hypotheses

- State clearly the hypothesis you are making
- Relate each assumption you make with existing literature
- State every assumption. No undocumented implicit assumption.

Discussion in details

- Discuss the results in light of your hypotheses : are they confirmed or not ?
- Weave together theory, hypotheses and results : what are the theoretical implications of your results ? Are they in line with theory ?
- This is also where you mention the limits of your analysis, the various possible interpretations, the extensions which would be necessary to have more clue to conclude about your research question ;

Conclusion and Appendix

Conclusion

- Summarize the paper very briefly

Appendix

- Put all tables and graphs with titles ;
- Put **labels** and not variable names in them ;
- Multiple linear regression results should be presented in the *outreg2* format.

Don't forget

- Both form and content are important ;
- Write complete sentences ;
- Be clear and concise (*put yourself in your reader's shoes*) ;
- State clearly your assumptions
- Be humble
- **Proofread**

Regression diagnosis

Adjusted R^2

- Best models are *parsimonious* (unnecessary variables are removed)
- R^2 will always increase when we add more parameters, regardless of whether they are relevant or not
- R_a^2 adjusted for the number of parameters is

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} \tag{1}$$

with n number of observations and p number of variables.

Example - Add noise to the regression

```
qui gen noise = uniform()
qui reg lcrime lpolice
estimates store m1
qui reg lcrime lpolice noise
estimates store m2
```

```
est table m1 m2, star stats((N r2 r2_a) b(%7.2f)
```

Variable	m1	m2
lpolice	1.18^{***}	1.18^{***}
noise		0.34
_cons	2.06^{***}	2.04^{***}
N	97	97
r2	0.47	0.47
r2_a	0.47	0.46

Legend : * p<0.05 ; ** p<0.01 ; *** p<0.001

Notice that

- The adjusted R^2 decreased
- The parameter for noise is not significant
- None of the other coefficients were affected at all because noise is not correlated to any of them

Assumptions of the regression

Simple Linear Regression

$$Y = \alpha + \beta X + \epsilon \tag{2}$$

β and α are correctly estimated under the following **assumptions** :

1. H_1 : Linear in *parameters*
2. H_2 :Random sampling : $\{Y_i, X_i\}$ are independent and indentically distributed (i.i.d.)
3. H_3 : No perfect collinearity : none of the covariates is constant and there are no exact linear relationships among the IVs.
4. H_4 : Zero Conditional mean : $E(\epsilon|X) = 0$ or in *plain English* : "values of the residuals, ϵ, does not depend on X.
5. H_5 : Heteroscedasticity : $Var(\epsilon|X) = Var(\epsilon)$. Variance of the residuals does not depend of X

Gauss-Markov theorem

Gauss-Markov Theorem

Under assumptions 1-5, the estimates from the linear model are BLUE.

- BLUE : **B**est **L**inear **U**nbiased **E**stimator
- Best = parameters have the smallest variances amid all linear unbiased estimators

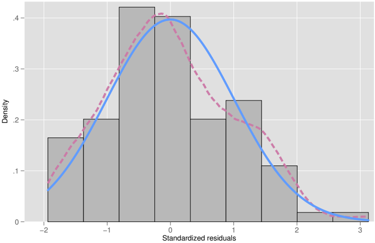
Regression diagnostics

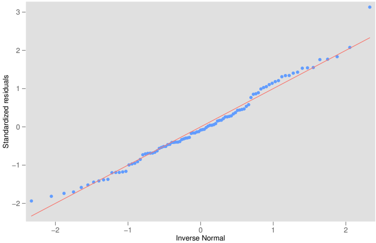
- We use regression diagnostics to check for violations of some assumptions
  - Deviations from the normality assumption
  - Outliers
  - Multicollinearity
  - Heteroskedasticity

Normality assumption 1/4

Example 1 - normal residuas

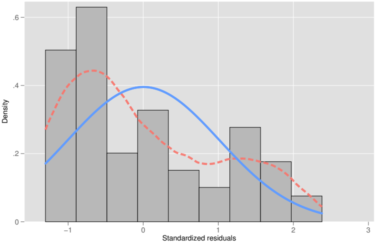
```
qui gen x = runiform()
qui gen y = x +rnormal(0,0.5)
qui reg y x
qui predict res_std, rstandard
qnorm res_std hist res_std, kdensity norm
```

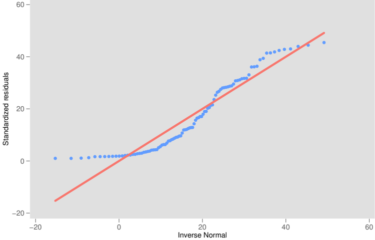




Example 2 - non-normal residuals

```
qui gen x = runiform()*10
qui gen y = x + 5*x^2 +rnormal(0,0.5)
qui reg y x
qui predict res_std, rstandard
qnorm res_std hist res_std, kdensity norm
```

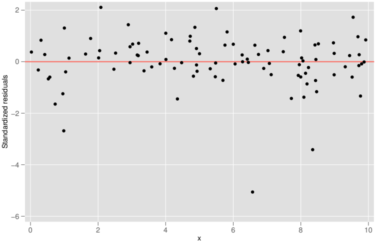




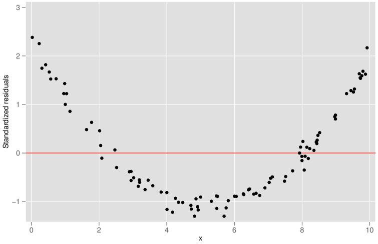
Plot residuals against X values

```
sc res_std x yline(0)
```

Normal errors



Non-normal errors



- Normality assumption is that the **residuals** are **normally-distributed**.
- Normality assumption is **not on** the unconditionnal **dependent variable,** Y.
- However, this course requires Y to be normally distributed for correlation tests and an ease of interpretation

What is an outlier ?

- a value that is larger or smaller than most of the other values of a variable
- Large errors influence results of the linear models
- *Large* is subjective

Measures of influence - Cook's distance

- Cook'distance for observation i

$$C_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - y\hat{_{j(i)}})^2}{\hat{\sigma}^2(p+1)} \tag{3}$$

- $y_{j(i)}$ is the predicted y when observation i has been removed

Model

- Data from Woolridge (cross-sectional firm data on R&D)
- *R&D*, measured as a percentage of sales (rdintens) is explained by sales
- Model :

$$rdintens = \alpha + \beta sales + \epsilon_i \tag{4}$$

Compute Cook's distance

```
reg rdintens sales
predict rdintens_cook if e(sample), cooksd
gen id = _n
gsort rdintens_cook
list id rdintens sales rdintens_cook in 1/5
```

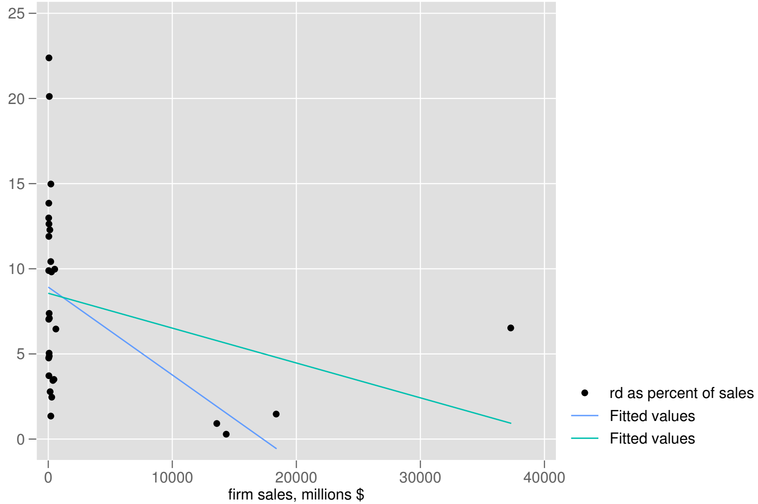
Outliers - Example

```
  list id rdintens sales rdintens_cook in 1/5
   +-----------------------------------+
   | id   rdintens     sales   rdinte~k |
   |-----------------------------------|
1. |  2   6.525412     37285   3.172364 |
2. | 19   22.38267      55.4   .1364463 |
3. | 14   20.11834      84.5   .0952832 |
4. |  5   .2858059   14345.4   .0616577 |
5. | 20   .9149596   13585.3   .0458502 |
   +-----------------------------------+
```

- Observation 2 is more influential than any other.

```
. reg rdintens sales

      Source |       SS           df       MS      Number of obs   =        29
-------------+----------------------------------   F(1, 27)        =      2.61
       Model |  77.6682236         1  77.6682236   Prob > F        =    0.1181
    Residual |  804.854368        27   29.809421   R-squared       =    0.0880
-------------+----------------------------------   Adj R-squared   =    0.0542
       Total |  882.522592        28   31.518664   Root MSE        =    5.4598

------------------------------------------------------------------------------
    rdintens |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sales |  -.0002047   .0001268    -1.61   0.118    -.0004649    .0000555
       _cons |   8.562179   1.084104     7.90   0.000     6.337781    10.78658
------------------------------------------------------------------------------

. reg rdintens sales if id != 2

      Source |       SS           df       MS      Number of obs   =        28
-------------+----------------------------------   F(1, 26)        =      6.21
       Model |  169.764751         1  169.764751   Prob > F        =    0.0194
    Residual |  710.677915        26   27.333766   R-squared       =    0.1928
-------------+----------------------------------   Adj R-squared   =    0.1618
       Total |  880.442667        27  32.6089877   Root MSE        =    5.2282

------------------------------------------------------------------------------
    rdintens |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sales |  -.0005158    .000207    -2.49   0.019    -.0009411   -.0000904
       _cons |   8.923437   1.056198     8.45   0.000     6.752391    11.09448
------------------------------------------------------------------------------
```

Outliers - Example



- Understanding the observations that might change results is important
- The more data you have, the less influential each observation is
- In some cases knowledge about the subject will help you evaluate if a value can be considered an outlier

Collinearity

- Interpreting coefficient of multiple linear models is done "holding other factors into account"
- $wage = \alpha + \beta_1 age \beta_2 educ + \epsilon$
- β_1 is the effect on average wage for an additional year of age, holding education constant
- If age and education are related, if it is not possible to hold education constant when we change the value for age.
- Ex : sample of young people all going to school, an extra year also implies another year of education

Perfect collinearity

- Perfect collinearity : one variable is a linear combination of other variables
- Estimation is impossible (H_3 is broken)

```
qui gen sales_profits = 1/3*sales + 2/3*profits
reg rd sales profits sales_profits

note: sales_profits omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =        32
-------------+----------------------------------   F(2, 29)        =    151.50
       Model |  2981673.87         2  1490836.94   Prob > F        =    0.0000
    Residual |   285366.68        29  9840.23035   R-squared       =    0.9127
-------------+----------------------------------   Adj R-squared   =    0.9066
       Total |  3267040.55        31  105388.405   Root MSE        =    99.198

------------------------------------------------------------------------------
          rd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sales |   .0173051   .0124676     1.39   0.176    -.0081939    .0428042
     profits |    .216847   .1138528     1.90   0.067    -.0160082    .4497022
sales_profits |          0  (omitted)
       _cons |   7.630929   20.13859     0.38   0.708     -33.5571    48.81896
------------------------------------------------------------------------------
```

Collinearity

- Create a highly correlated variable but no perfectly collinear

Stata

```
qui gen profits_noisy = profits + rnormal(0,5)
qui reg rd profits
est sto m1
qui reg rd profits profits_noisy
est sto m2
est table m1 m2, se p stats(N r2 r2_a F)
----------------------------------------
    Variable |     m1           m2
-------------+--------------------------
     profits |  .37204849    1.2433652
             |   .0217702    4.1541153
             |     0.0000       0.7668
profits_no~y |              -.87198564
             |               4.1572454
             |                  0.8353
       _cons |  15.836122    15.515058
             |  19.546459    19.924432
             |     0.4242       0.4425
-------------+--------------------------
           N |         32           32
          r2 |  .90685005    .90699115
        r2_a |  .90374505    .90057675
           F |  292.06137    141.39915
----------------------------------------
```

- Model fit is still good
- Coefficient for *profits* almost multiplied by three
- Standard errors of profits multiplied by 200
- F statistics went down

Variance inflation factor

- One way to diagnose collinearity is to investigate how each explanatory variable in a model is related to all other explanatory variables in the model

Variance inflation factor, **VIF**

- VIF for variable j is $VIF_j = \frac{1}{1-R^2}$
- The R^2 in VIF is obtained by regression X_j against all other explanatory variables

Intuitions

- R^2 low $\rightarrow$ VIF close to 1
- R^2 hight $\rightarrow$ VIF will be high
- A rule of thumb is that a $VIF > 10$ provides evidence of collinearity.($R^2 > 0.9$)

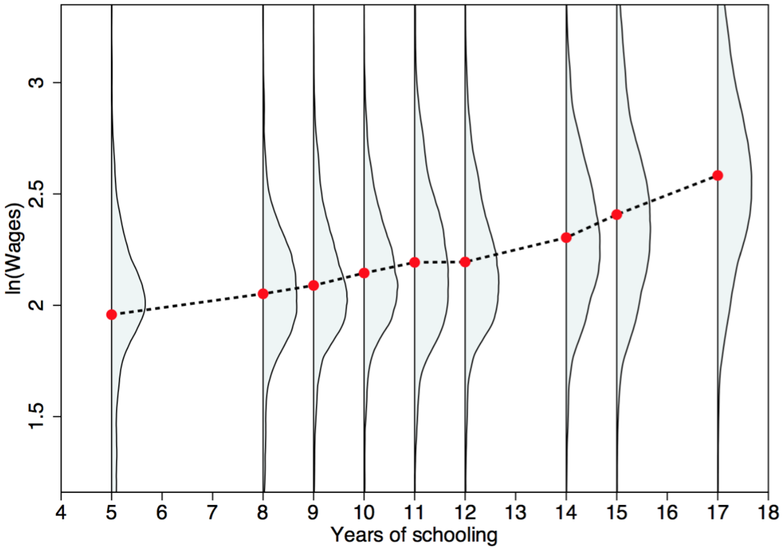
Collinearity

Signs of collinearity

- Collinearity makes estimation “unstable”
- Large changes in estimated parameters when a variable is added / deleted
- Signs of coefficients do not agree with expectations (require subject knowledge)

What to do

- Exploratory analysis to detect highly correlated predictors
- Understand what drives multicollinearity
- Easy cases : drop one variable
- Harder cases : more data / different type of model

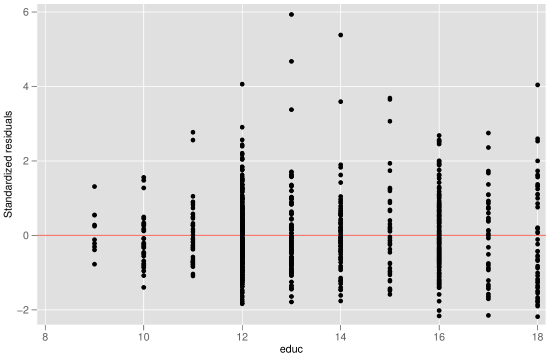


ln(Wages)
3
2.5
2
1.5
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Years of schooling

Example

$$wage = \alpha + \beta_1 educ + \beta_2 age + \epsilon \tag{5}$$

```
reg wage educ age
predict res_std, rsta
sc res_std educ, yline(0)
```



Testing for heteroskedasticity 3/4

- Use the **Breusch-Pagan** test `hettest`

```
reg wage educ age
estat hettest, rhs
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: educ age

         chi2(2)      =    41.49
         Prob > chi2  =   0.0000
```

- We reject H_0 : there is heteroskedasticity
- We can also test for age and education separately

```
estat hettest age
Breusch-Pagan / Cook-Weisberg test
for heteroskedasticity
         Ho: Constant variance
         Variables: age

         chi2(1)      =     1.59
         Prob > chi2  =   0.2073
```

```
estat hettest educ
Breusch-Pagan / Cook-Weisberg test
for heteroskedasticity
         Ho: Constant variance
         Variables: educ

         chi2(1)      =    39.70
         Prob > chi2  =   0.0000
```

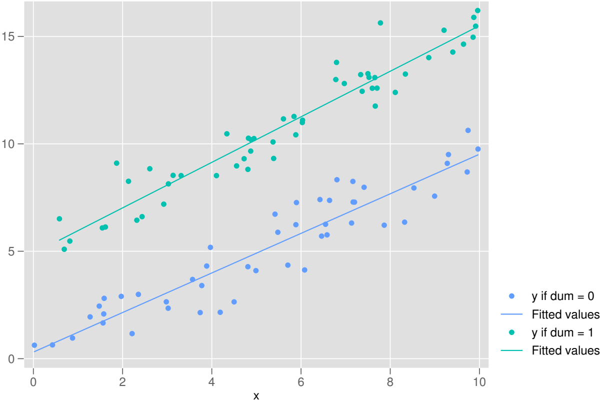
- Heteroskedasticity comes from education

Dealing with heteroskedasticity 4/4

- Use the **Huber-White** robust s.e. (sandwich estimator)
  - `reg wage educ age, vce(robust)`
  - Sandwich estimator is asymptotically unbiased : ok with large samples, not with small ones
- Use transformation of the variables, like *logs*, may help
- OLS estimates are consistent even if there is heteroskedasticity, but the standard errors are not.

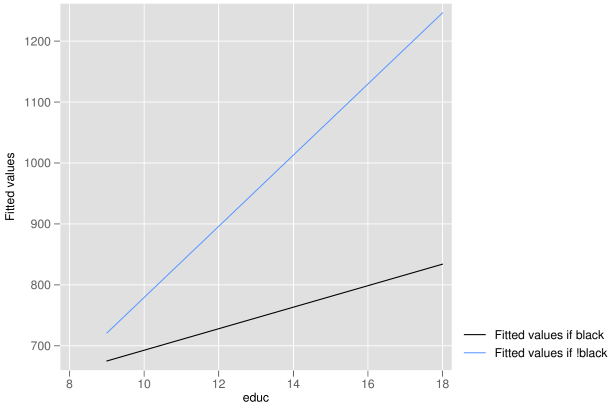
Interaction effects

- Multiple regression : each X_i has a straight line relationship with the mean of y, holding other varialbes constant
- $y = \alpha + \beta x + \delta dum + \epsilon$



- However, the effect of an explanatory variable may change considerably as the value of another explanatory variable in the model changes

```
tw (sc wage educ if black) (sc wage educ if !black)
```



Deal with interaction effects 3/6

2 main choices

- Run separate models for each categories
- Include interaction variable `x1#x2`

Example

```
reg earnings i.race i.sex race#sex

      Source |       SS           df       MS      Number of obs   =    29,557
-------------+----------------------------------   F(3, 29553)     =    409.63
       Model |  17141.3369         3  5713.77898   Prob > F        =    0.0000
    Residual |  412226.012    29,553  13.9487027   R-squared       =    0.0399
-------------+----------------------------------   Adj R-squared   =    0.0398
       Total |  429367.349    29,556  14.5272482   Root MSE        =    3.7348

------------------------------------------------------------------------------------
                  earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------------+--------------------------------------------------------
                      race |
        Black and hispanic |  -.8916536   .0705019   -12.65   0.000     -1.02984   -.7534667
                           |
                       sex |
                    Female |  -1.484246   .0532738   -27.86   0.000    -1.588665   -1.379827
                           |
                  race#sex |
Black and hispanic#Female  |   .3844973    .093391     4.12   0.000     .2014467    .5675478
                           |
                     _cons |   4.681477   .0392636   119.23   0.000     4.604518    4.758435
------------------------------------------------------------------------------------
```

$$earnings = \alpha + \beta_r race + \beta_2 sex + \beta_3 racesex + \epsilon \tag{6}$$

Interpretations

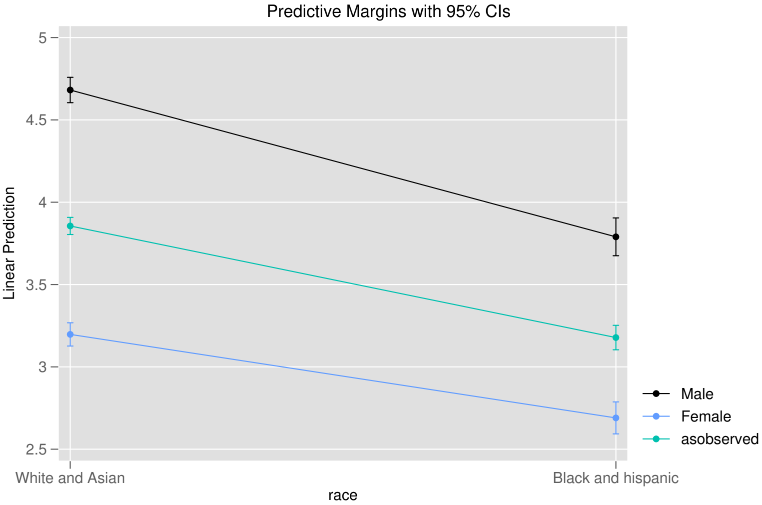
- β_1 represents the effect of $race$ on $earnings$ when $sex = 0$.
- β_2 represents the effect of sex on $earnings$ when $race = 0$.
- $\beta_3 > 0$: the effect of $race$ in earnings increase when $sex = 1$
- $\beta_3 > 0$: the effect of $race$ in earnings decrease when $sex = 1$

Keep in mind

- Include both variables alone, not only the interaction variable
- The interpretation of the coefficients of these single variables is different from a common multiple linear regression
- Use `margins` and `marginsplot`

```
margins race sex race#sex

Predictive margins                              Number of obs     =     29,557
Model VCE    : OLS

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------------------
                           |            Delta-method
                           |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------------+--------------------------------------------------------------
                      race |
           White and Asian |   3.855869   .0265465   145.25   0.000     3.803836    3.907901
        Black and hispanic |   3.178091   .0378784    83.90   0.000     3.103847    3.252334
                           |
                       sex |
                      Male |   4.387346   .0326408   134.41   0.000     4.323368    4.451323
                    Female |   3.029934   .0291434   103.97   0.000     2.972812    3.087056
                           |
                  race#sex |
      White and Asian#Male |   4.681477   .0392636   119.23   0.000     4.604518    4.758435
    White and Asian#Female |    3.19723   .0360065    88.80   0.000     3.126656    3.267805
   Black and hispanic#Male |   3.789823   .0585567    64.72   0.000     3.675049    3.904597
Black and hispanic#Female |   2.690074   .0495469    54.29   0.000      2.59296    2.787188
------------------------------------------------------------------------------------------
```

```
margins race sex race#sex
marginsplot
```



Practice

Practice

Practice : Satisfaction with Health Services in Britain and France

- Run step by step `week11.do`
- Remember to comment `run setup`
- Install packages manually if you are using Sciences Po computer