# class review

Adam Okulicz-Kozaryn
adam.okulicz.kozaryn@gmail.com

this version: Tuesday 24th April, 2018 17:28

# outline

review [if time! otherwise review at home!]

# outline

review [if time! otherwise review at home!]

## what is it?

- ◇ this set of slides reviews what we have covered earlier
- ◇ i.e. the following slides are just verbatim copies of slides you've seen earlier

# solving the problem

| | $Y_i$ | $X_i$ | $(Y_i - \bar{Y}) = y_i$ | $(X_i - \bar{X}) = x_i$ | $y_i^2$ | $x_i^2$ | $y_i x_i$ |
|---|---|---|---|---|---|---|---|
| | 2 | 1 | –2.33 | –1 | 5.53 | 1 | 2.33 |
| | 5 | 2 | 0.67 | 0 | 0.45 | 0 | 0 |
| | 6 | 3 | 1.67 | 1 | 2.79 | 1 | 1.67 |
| $\Sigma$ | 13 | 6 | 0 | 0 | 8.67 | 2 | 4 |
| *mean* | 4.33 | 2 | | | | | |

- $\hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{4}{2} = 2$
- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 4.33 - (2)(2) = 0.33$

## example: age and fear

- In this example, imagine that we have some sort of survey that measures people's fear of crime, and that our hypothesis is that fear of crime increases with age. Assume the fear measure is an index ranging from 0 to 15.
- First, we calculate the means. Second, we calculate the deviations from the means and the their squares for each observation, as well as the co-product of the X and Y deviations. Finally, we sum these up.

# example: age and fear

The Data

| obs | $X_i$ | $Y_i$ |
|---|---|---|
| 1 | 22 | 2 |
| 2 | 35 | 7 |
| 3 | 47 | 6 |
| 4 | 56 | 14 |
| 5 | 72 | 13 |
| $\sum$ | 232 | 42 |

$$\bar{X} = \frac{232}{5} = 46.4$$

$$\bar{Y} = \frac{42}{5} = 8.4$$

Deviations from the means

| Obs | $x_i$ | $x_i^2$ | $y_i$ | $y_i^2$ | $x_i y_i$ |
|---|---|---|---|---|---|
| 1 | –24.4 | 595.36 | –6.4 | 40.96 | 156.16 |
| 2 | –11.4 | 129.96 | –1.4 | 1.96 | 15.96 |
| 3 | 0.6 | 0.36 | –2.4 | 5.76 | –1.44 |
| 4 | 9.6 | 92.16 | 5.6 | 31.36 | 53.76 |
| 5 | 25.6 | 655.36 | 4.6 | 21.16 | 117.76 |
| $\sum$ | 0 | 1473.2 | 0 | 101.2 | 342.2 |

◇

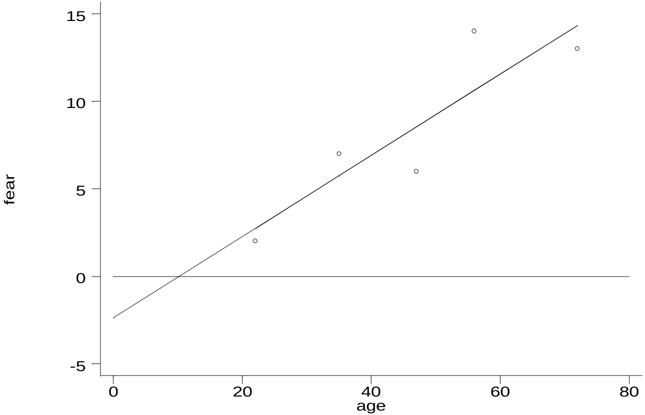
◇ $\hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{342}{1473} = .232$

◇ $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 8.4 - (.232)(46.4) = -2.365$

◇ $\hat{Y}_i = \hat{\beta}_1 + \beta_2 X_i = -2.365 + .232 X_i$

◇ how would you interpret this?

# the estimated regression line



## variance and std error of regression

- ok, we know how to calculate betas and fit the line (that min the sum of the squared resid)
- but there are lines that fit better and lines that fit worse in different samples
- we need a measure of uncertainty, i.e. how well our line fit the data...
- and the fit is measured by residuals...
- ... so our measure of uncertainty has to do with residuals !

## variance and std error of regression

- $s^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$
- $s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$
  again, the mean of the residuals is zero (hence, ē drops out)
- why divide by n-2?
- $s^2$ and $s$ are measures of the spread of the points around the estimated regression line.
- they are estimators of the variance and standard deviation of the disturbance terms:$\sigma^2$ and $\sigma$

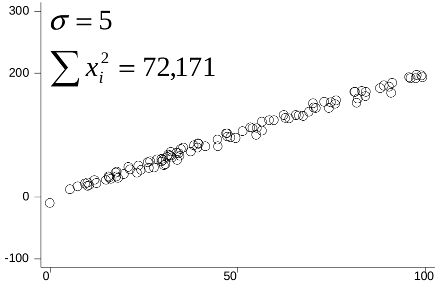
# from predicted values to std err

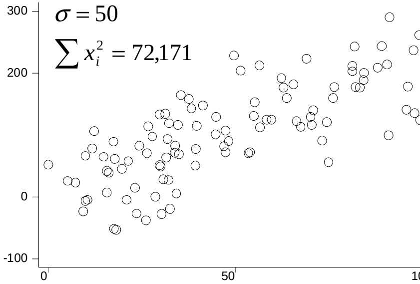
| $i$ | $\hat{Y}_i$ | $e_i$ | $e_i^2$ |
|---|---|---|---|
| 1 | 2.739 | –0.739 | 0.546 |
| 2 | 5.755 | 1.245 | 1.556 |
| 3 | 8.539 | –2.539 | 6.447 |
| 4 | 10.627 | 3.373 | 11.377 |
| 5 | 14.339 | –1.339 | 1.793 |
| $\Sigma$ | | 0 | 21.713 |

- $s = \sqrt{\frac{\sum_{i=1}^{5} e_i^2}{n-2}} = \sqrt{\frac{21.7}{3}} = 2.7$
- what is it measuring?
- $s_{\hat{\beta}_2} = \frac{s}{\sum_{i=1}^{5} x_i^2} = \frac{2.7}{\sqrt{1473}} = .07$
- how does it differ from $s$?

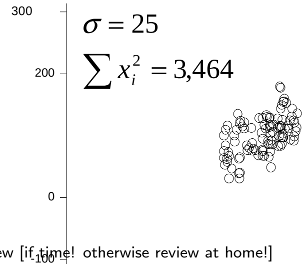
# Standard Error of the Slope Coefficient

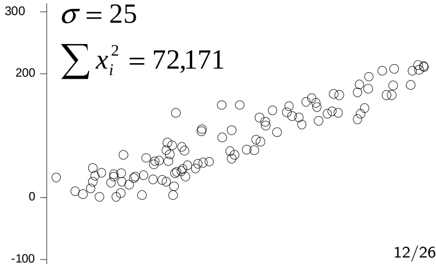
## Numerator -- variance of disturbance term




## Denominator -- variation in X



## key ols assumptions

- ◇ the true model is linear $Y_i = \beta_1 + \beta_2 X_i + u_i$
- · $cov[X_i u_i] = 0$ $X$ and $u$ are not correlated
- · $var[u_i] = \sigma^2$ constant variance
- ◇ if true, then BLUE: Best Linear Unbiased Estimators
- ◇ (there are other assumptions, too)

## predicted values and residuals

- 

| Y | X | Y_hat | e | $e^2$ |
|---|---|---|---|---|
| 1 | 17 | | | |
| 3 | 13 | | | |
| 5 | 8 | | | |
| 7 | 10 | | | |
| 9 | 2 | | | |

- $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
- for obs 1:
- $\hat{Y}_1 = 10.24 + (-0.524)(17) = 1.332$
- $e_1 = 1 - 1.33 = -0.33$

## confidence intervals

- In general, a confidence interval is the point estimator plus or minus a margin of error, which consists of a distribution parameter (z or t) times the standard error of the estimator. In this case (small sample, $\sigma$ unknown, we use the t distribution.
- $PE \pm (t_{\frac{\alpha}{2},DOF})(SE) = \hat{\beta}_2 \pm t_{0.025,3} s_{\hat{\beta}_2}$

# hypothesis test

- the null is that slope ("the unobserved true parameter")is zero (i.e. no effect)
- $H_0 : \beta_2 = 0$
- $H_A : \beta_2 \neq 0$
- $t = \frac{\hat{\beta}_2 - \beta_2}{s_{\hat{\beta}_2}}$

## exercise 1

- you regressed car's price on its weight

```
------------------------------------
       price |      Coef.   Std. Err.
-------------+----------------------
      weight |   2.044063   .3768341
```

- interpret the coefficient
- is it significant ?
- calculate 95% CI

# the 'beta' option

```
. sum wage educ exp
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        534    9.023939    5.138876          1       44.5
        educ |        534    13.01873    2.615373          2         18
         exp |        534     17.8221    12.37971          0         55

. reg wage educ exp, beta
      Source |       SS       df       MS              Number of obs =     534
-------------+------------------------------           F(  2,   531) =   67.22
       Model |  2843.72544     2  1421.86272           Prob > F      =  0.0000
    Residual |   11231.763   531   21.152096           R-squared     =  0.2020
-------------+------------------------------           Adj R-squared =  0.1990
       Total |  14075.4884   533  26.4080458           Root MSE      =  4.5991

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |    .925947   .0813995    11.38   0.000                 .4712502
         exp |   .1051282   .0171967     6.11   0.000                 .2532571
       _cons |  -4.904318   1.218865    -4.02   0.000                        .
------------------------------------------------------------------------------
```

$$\hat{\beta}_2^* = \hat{\beta}_2 \frac{s_X}{s_Y} = 0.926\left(\frac{2.615}{5.139}\right) = 0.471 \qquad \hat{\beta}_3^* = \ldots$$

# lovb

- true model:

$$Y_i = \beta_1 + \beta_2 INCL + \beta_3 EXCL + u_i$$

- we estimate:

$$Y_i = \alpha_1 + \alpha_2 INCL + v_i$$

$$\mathrm{E}[\hat{\alpha}_2] = \alpha_2 = \beta_2 + \beta_3 \left( (\rho_{EI}) \left( \frac{\sigma_E}{\sigma_I} \right) \right)$$

What you estimate using the 2 variable regression

The unbiased coefficient

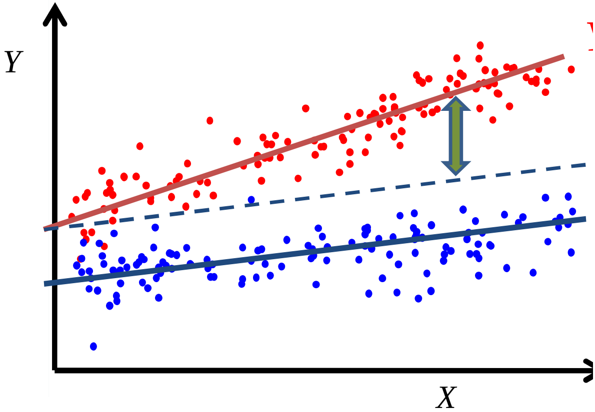
The coefficient on the left out variable

rho is the bivariate correlation of the included and excluded variables

sign of bias: $\beta_3 * \rho_{EI}$

# continuous/dummy interactions

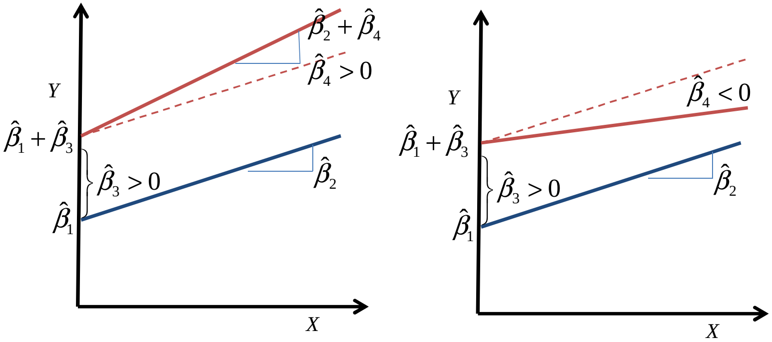
- $Y_i = \beta_1 + \beta_2 X_i + \beta_3 female_i + \beta_4 female_i * X_i + u_i$




-

# schematic

◇ $Y_i = \beta_1 + \beta_2 X_i + \beta_3 female_i + \beta_4 female_i * X_i + u_i$




◇

## interaction of dummies

- ◇ if there is an interaction effect between two variables, the effect of one variable depends on the level of the other
- ◇ eg the effect of marriage on wage depends on gender.
- ◇ interactions go both ways:
  - · the effect of gender depends on marital status, too

# interaction of dummies

- $Y_i = \beta_1 + \beta_2 female + \beta_3 married + \beta_4 female * married + u_i$

| | Male | Female | Gender Difference |
|---|---|---|---|
| Unmarried | $\hat{\beta}_1$ | $\hat{\beta}_1 + \hat{\beta}_2$ | $\hat{\beta}_2$ |
| Married | $\hat{\beta}_1 + \hat{\beta}_3$ | $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4$ | $\hat{\beta}_2 + \hat{\beta}_4$ |
| Effect of Marriage | $\hat{\beta}_3$ | $\hat{\beta}_3 + \hat{\beta}_4$ | $\hat{\beta}_4$ |

-

# example [let's calc tab from reg]

```
. table married female, c(mean wage) row col f(%7.2f)
----------------------------------
          |        Gender
  Married |    male  female   Total
----------+-----------------------
       no |    8.35    8.26    8.31
      yes |   10.88    7.68    9.40
          |
    Total |    9.99    7.88    9.02
----------------------------------
```

```
. gen femxmar = female*married
. reg wage female married femxmar
```

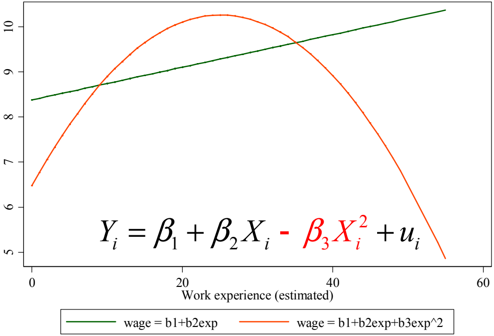
| wage | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| $\hat{\beta}_2$ female | -.0951892 | .7350367 | -0.13 | 0.897 | -1.539132 | 1.348754 |
| $\hat{\beta}_3$ married | 2.521222 | .6120814 | 4.12 | 0.000 | 1.318819 | 3.723626 |
| $\hat{\beta}_4$ femxmar | -3.09704 | .9072785 | -3.41 | 0.001 | -4.879344 | -1.314737 |
| $\hat{\beta}_1$ _cons | 8.354677 | .4936728 | 16.92 | 0.000 | 7.384882 | 9.324473 |

## interpretation: transforming variables

- Lin: One unit change in X leads to a $\beta_2$ unit change in Y.
- Log-Lin: One unit change in X leads to a $100 * \beta_2$ % change in Y. (guj ed4:p180 fig6.4; ed5:p163 ex6.4)
- Lin-Log: One percent change in X leads to a $\beta_2/100$ unit change in Y. (guj: ed4:p182 fig6.5; ed5:p165-6 ex6.5)
- Log-Log (aka log-linear or "linear in logs"): One percent change in X leads to a $\beta_2$ % change in Y (elasticity).

# quadratic model

If a *non-linear relationship* between $X$ and $Y$ is suspected, a *polynomial function of* $X$ can be used to model it.




when it flips:

$$X_i^* = -\frac{\beta_2}{2\beta_3}$$

This curve reaches a maximum wage at the point were the marginal effect of experience is zero.