Session 3: Further regression

Formula

Formula for multiple regression

Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k + e$

$e \sim N(0, \sigma)$

- Interpretation of β_j
  - How much $\hat{Y}$ changes for a 1-unit in X_j holding all other values constant
  - The estimated effect on Y of a 1-unit change in X_j, "controlling for" or "taking account" of all the other Xs

sociology

Residuals

$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k$

$Y = \hat{Y} + e$

$e \sim N(0, \sigma)$

- Mean of zero
- Standard deviation of σ (RMSE)
- Normally distributed
- Should have no structured relationship to X variables

sociology

Session 3: Further regression

R^2

R^2

- R^2: coefficient of multiple determination
- TSS = sum of squared deviation from the mean = $\sum(Y_i - \bar{Y})^2$
- RSS = sum of squared deviation from the regression prediction = $\sum(Y_i - \hat{Y})^2$
- $R^2 = \frac{TSS - RSS}{TSS}$
- Range: 0 (no relationship) to 1 (perfect linear relationship)
- PRE: Proportional Reduction in Error

sociology

R^2 and correlation

- In bivariate regression, R^2 is the square of the correlation coefficient between Y and X
- In multiple regression, it is the square of the correlation between Y and $\hat{Y}$
- (In bivariate regression the correlation between X and $\hat{Y}$ is 1)

sociology

Session 3: Further regression

Indicator variables

Multicategory explanatory variables -> Indicator variables

- We often use "indicator coding" or "dummy coding"
- For 2-category variables, set one category to 0, the other to 1: interpret as the effect of being in the second category (e.g., female) compared with the first.

```
. reg income age i.sex

      Source |       SS           df       MS      Number of obs   =       959
-------------+----------------------------------   F(2, 956)       =     45.72
       Model |  33922983.9         2    16961492   Prob > F        =    0.0000
    Residual |   354670636       956  370994.389   R-squared       =    0.0873
-------------+----------------------------------   Adj R-squared   =    0.0854
       Total |   388593620       958  405630.083   Root MSE        =    609.09

------------------------------------------------------------------------------
      income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |  -3.144945   1.083398    -2.90   0.004    -5.271057   -1.018833
             |
         sex |
     female  |   -352.678   39.51326    -8.93   0.000    -430.2208   -275.1353
       _cons |   1035.878   54.58935    18.98   0.000     928.7494    1143.007
------------------------------------------------------------------------------
```

sociology

More than two categories

With more that two categories we create a set of binary variables, "indicator variables" or "dummy variables":

	d1	d2	d3	d4
a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1

For m categories, m-1 dummy variables are sufficient.

We interpret the parameter as the estimated effect of being in that category relative to the omitted or "reference" category.

Stata handles this automatically with the `i.` prefix.

sociology

Example: education

```
. reg income age i.sex i.qual

      Source |       SS           df       MS      Number of obs   =       959
-------------+----------------------------------   F(5, 953)       =     54.14
       Model |  85960604.5         5  17192120.9   Prob > F        =    0.0000
    Residual |   302633015       953  317558.253   R-squared       =    0.2212
-------------+----------------------------------   Adj R-squared   =    0.2171
       Total |   388593620       958  405630.083   Root MSE        =    563.52

------------------------------------------------------------------------------
            income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------------+----------------------------------------------------------------
               age |  -.3897295    1.04777    -0.37   0.710    -2.445933    1.666474
                   |
               sex |
           female  |  -336.9623   36.75947    -9.17   0.000    -409.1011   -264.8234
                   |
              qual |
A-level, other sub-d.. |  -459.9208   78.54165    -5.86   0.000    -614.0554   -305.7862
O-level, commercial,.. |   -701.695   77.16016    -9.09   0.000    -853.1185   -550.2716
  Sub-O-level, no qual |  -864.9695   76.41768   -11.32   0.000    -1014.936   -715.0032
                   |
             _cons |   1563.508   81.83797    19.10   0.000     1402.904    1724.111
------------------------------------------------------------------------------
```


Session 3: Further regression

Hypothesis testing

Hypothesis testing: one parameter at a time

- t-test: $abs(\hat{\beta}_j/\mathrm{se}_j) > t$
- Interpretation:
  - Null: population value of β is 0; this variable has no influence once the other variables are taken account of

Example

```
. reg income age i.sex

      Source |       SS           df       MS      Number of obs   =       959
-------------+----------------------------------   F(2, 956)       =     45.72
       Model |  33922983.9         2    16961492   Prob > F        =    0.0000
    Residual |   354670636       956  370994.389   R-squared       =    0.0873
-------------+----------------------------------   Adj R-squared   =    0.0854
       Total |   388593620       958  405630.083   Root MSE        =    609.09

------------------------------------------------------------------------------
      income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |  -3.144945   1.083398    -2.90   0.004    -5.271057   -1.018833
             |
         sex |
     female  |   -352.678   39.51326    -8.93   0.000    -430.2208   -275.1353
       _cons |   1035.878   54.58935    18.98   0.000     928.7494    1143.007
------------------------------------------------------------------------------
```


Hypothesis testing: all parameters together

- F-test:
  - $\beta_1 = \beta_2 \ldots = \beta_k = 0$
- Null hypothesis: no X variable has an effect once the others are taken care of.
- A "global" test: the null is that there is no relevant variable in the model
- Calculation based on TSS and RSS, but also number of cases and number of parameters estimated
- Uses F distribution (two df parameters: k and n-k-1, k is number of parameters, n the number of cases)

Hypothesis testing: additional parameters

- Delta F-test compares "nested" models
  - Model 1: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_g X_g$
  - Model 1: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_g X_g + \beta_h X_h ... + \beta_k X_k$
- Null hypothesis: $\beta_h = \ldots = \beta_k = 0$
- That is, given the variables already in the model, the additional variables contribute no explanatory power.
- Useful when adding multi-category variables, or related groups of variables

delta-F example: group of indicator variables

```
. qui reg income age i.sex

. est store base

. qui reg income age i.sex i.qual

. ftest base
Assumption: base nested in .

F(  3,     953) =     54.62
       prob > F =    0.0000
```

Note: ftest is an add-on command. Do `ssc install ftest` to install

Session 3: Further regression

Multicollinearity

Multicollinearity

- Multicollinearity arises where variable that individually "work" share too much of their explanatory power
- When both are in the model, they may both be insignificant
- Not simply correlation, but that they share too much of their correlation with Y
- Often arises when the 2 variables both measure the same phenomenon
- Usually a small sample problem
- Don't worry unless you see variables inexplicably becoming insignficant

Bodyfat correlations

```
. use http://www.stata-press.com/data/r14/bodyfat
(Body Fat)

. corr *
(obs=20)

             |  triceps    thigh   midarm  bodyfat
-------------+------------------------------------
     triceps |   1.0000
       thigh |   0.9238   1.0000
      midarm |   0.4578   0.0847   1.0000
     bodyfat |   0.8433   0.8781   0.1424   1.0000
```


Triceps predicting bodyfat

```
. reg bodyfat tricep

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =     44.30
       Model |  352.269824         1  352.269824   Prob > F        =    0.0000
    Residual |  143.119689        18  7.95109386   R-squared       =    0.7111
-------------+----------------------------------   Adj R-squared   =    0.6950
       Total |  495.389513        19  26.0731323   Root MSE        =    2.8198

------------------------------------------------------------------------------
     bodyfat | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     triceps |   .8571866   .1287808     6.66   0.000     .5866282    1.127745
       _cons |  -1.496107   3.319235    -0.45   0.658     -8.46956    5.477347
------------------------------------------------------------------------------
```


Thigh predicting bodyfat

```
. reg bodyfat thigh

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =     60.62
       Model |  381.965845         1  381.965845   Prob > F        =    0.0000
    Residual |  113.423669        18  6.30131492   R-squared       =    0.7710
-------------+----------------------------------   Adj R-squared   =    0.7583
       Total |  495.389513        19  26.0731323   Root MSE        =    2.5102

------------------------------------------------------------------------------
     bodyfat | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       thigh |   .8565467   .1100156     7.79   0.000     .6254124    1.087681
       _cons |  -23.63449   5.657414    -4.18   0.001    -35.52028   -11.74871
------------------------------------------------------------------------------
```


Midarm predicting bodyfat

```
. reg bodyfat midarm

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =      0.37
       Model |  10.0516092         1  10.0516092   Prob > F        =    0.5491
    Residual |  485.337904        18  26.9632169   R-squared       =    0.0203
-------------+----------------------------------   Adj R-squared   =   -0.0341
       Total |  495.389513        19  26.0731323   Root MSE        =    5.1926

------------------------------------------------------------------------------
     bodyfat | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      midarm |   .1994287   .3266297     0.61   0.549    -.4867949    .8856523
       _cons |   14.68678   9.095926     1.61   0.124    -4.423052    33.79661
------------------------------------------------------------------------------
```


Omnibus model: none significant

```
. reg bodyfat tricep thigh midarm

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(3, 16)        =     21.52
       Model |  396.984607         3  132.328202   Prob > F        =    0.0000
    Residual |  98.4049068        16  6.15030667   R-squared       =    0.8014
-------------+----------------------------------   Adj R-squared   =    0.7641
       Total |  495.389513        19  26.0731323   Root MSE        =      2.48

------------------------------------------------------------------------------
     bodyfat | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     triceps |   4.334085   3.015511     1.44   0.170    -2.058512    10.72668
       thigh |  -2.856842   2.582015    -1.11   0.285    -8.330468    2.616785
      midarm |  -2.186056   1.595499    -1.37   0.190    -5.568362     1.19625
       _cons |   117.0844   99.78238     1.17   0.258    -94.44474    328.6136
------------------------------------------------------------------------------
```


VIF for the model

```
. estat vif

    Variable |       VIF       1/VIF  
-------------+----------------------
     triceps |    708.84    0.001411
       thigh |    564.34    0.001772
      midarm |    104.61    0.009560
-------------+----------------------
    Mean VIF |    459.26
```


Drop triceps

```
. reg bodyfat thigh midarm

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(2, 17)        =     29.40
       Model |  384.279748         2  192.139874   Prob > F        =    0.0000
    Residual |  111.109765        17  6.53586854   R-squared       =    0.7757
-------------+----------------------------------   Adj R-squared   =    0.7493
       Total |  495.389513        19  26.0731323   Root MSE        =    2.5565

------------------------------------------------------------------------------
     bodyfat | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       thigh |   .8508818   .1124482     7.57   0.000     .6136367    1.088127
      midarm |   .0960295   .1613927     0.60   0.560    -.2444792    .4365383
       _cons |  -25.99696    6.99732    -3.72   0.002    -40.76001   -11.2339
------------------------------------------------------------------------------
```


New VIF

```
. estat vif

    Variable |       VIF       1/VIF  
-------------+----------------------
      midarm |      1.01    0.992831
       thigh |      1.01    0.992831
-------------+----------------------
    Mean VIF |      1.01
```


Session 3: Further regression

Residuals

Residuals

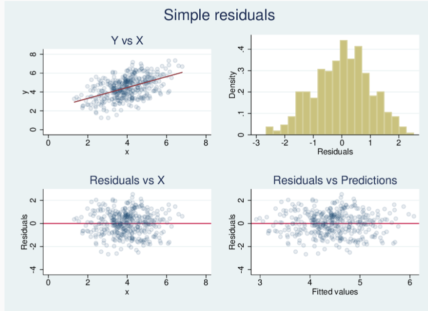
$Y = b_0 + b_1 X_1 + ... + b_k X_k + e$

$e \sim N(0, \sigma)$

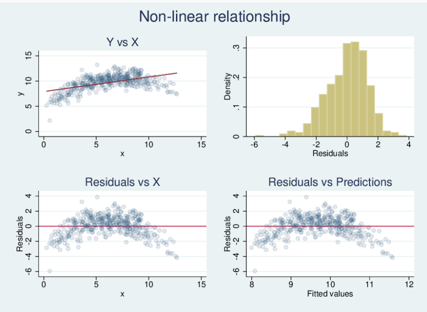
Characteristics

- Residuals will
  - have mean 0
  - be as small as possible
  - have no linear relationship to X variables
- Residuals should
  - be approximately normally distributed (symmetric is often enough)
  - not have a non-linear relationship to any X variable
  - have a constant spread, that is not related to X or Y values
- If correlated with variables not in the model, perhaps those variables should be included

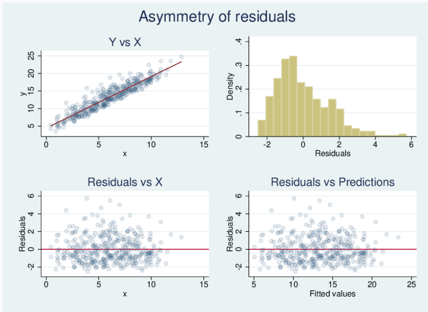
Examining residuals: ideal



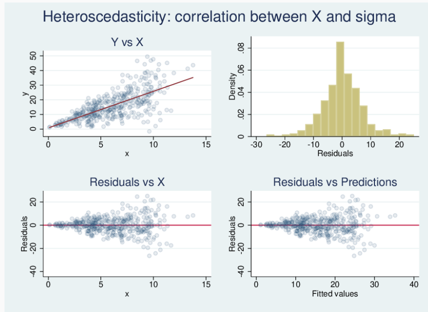
Examining residuals: Non-linear



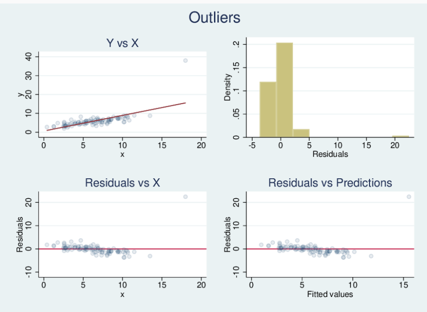
Examining residuals: asymmetric



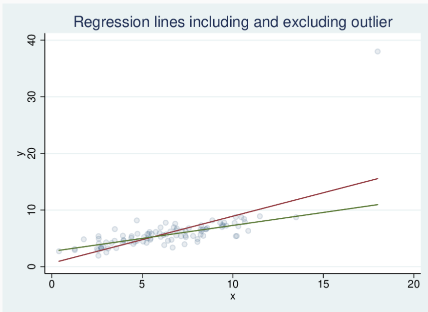
Examining residuals: heteroscedasticity



Examining residuals: Spotting outliers



Examining residuals: Influence of outliers



Session 3: Further regression

Influence

Outliers may have undue influence

- `dfbeta`
- Cook's distance

Outlier interactive app

http://teaching.sociology.ul.ie:3838/influence/

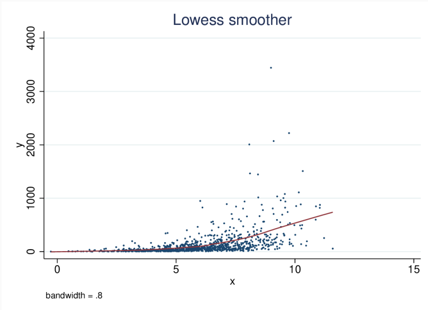
Session 3: Further regression

Log regression

Multiplicative relationship

- Where the underlying relationship is multiplicative, linear regression doesn't work well
- Implies an additive increase where a multiplicative one is better
- If we take the log of the dependent variable:
  - better estimates
  - often cures heteroscedasticity

Simulation: Y increases 65% for X +1



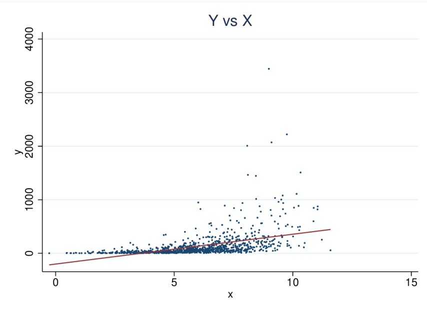
Linear regression

```
. reg y x

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =    274.71
       Model |  12181477.5         1  12181477.5   Prob > F        =    0.0000
    Residual |  44253675.2       998  44342.3599   R-squared       =    0.2158
-------------+----------------------------------   Adj R-squared   =    0.2151
       Total |  56435152.7       999  56491.6443   Root MSE        =    210.58

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |   55.69088   3.360033    16.57   0.000     49.09734    62.28442
       _cons |  -200.7041   20.95566    -9.58   0.000    -241.8263   -159.5819
------------------------------------------------------------------------------
```


Predictions



Log(Y)

```
. gen ly = log(y)

. reg ly x

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =   1032.66
       Model |   956.12538         1   956.12538   Prob > F        =    0.0000
    Residual |  924.030142       998  .925881905   R-squared       =    0.5085
-------------+----------------------------------   Adj R-squared   =    0.5080
       Total |  1880.15552       999  1.88203756   Root MSE        =    .96223

------------------------------------------------------------------------------
          ly | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |   .4933914   .0153537    32.14   0.000     .4632622    .5235205
       _cons |   1.062305   .0957568    11.09   0.000     .8743972    1.250213
------------------------------------------------------------------------------
```


Interpretation

- For a 1 unit change in X, $log(\hat{Y})$ rises by 0.4933914
- Thus for a 1 unit change in X, Y rises by $e^{0.4933914} = 1.638$
- $e^{0.4933914}$ is the antilog of 0.4933914

Predictions

Predicted values

- Where the dependent variable is logged the prediction of the Y value is not simply the anti-log of the predicted log(Y)
- When we take the anti-log we must take account of the fact that residuals above the line expand by more than residuals below the line
- Thus a small correction

$$\widehat{log(Y)} = a + bX$$
$$\hat{Y} = e^{\widehat{log(Y)}} * e^{\mathrm{RMSE}^2/2}$$

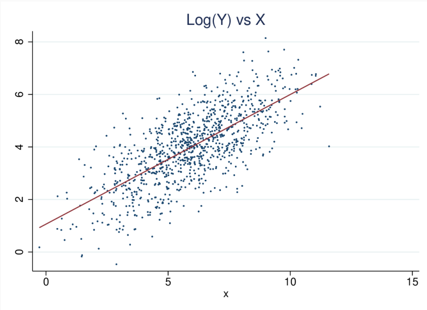
- where RMSE is the standard deviation of the regression

Calculations

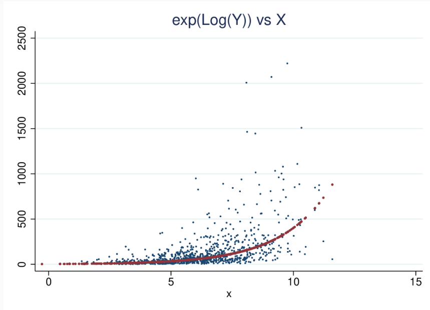
```
gen ly = log(y)
reg ly x

predict lyhat
gen elyh = exp(lyhat)
gen elyh2 = elyh * exp(rmse^2/2)
```

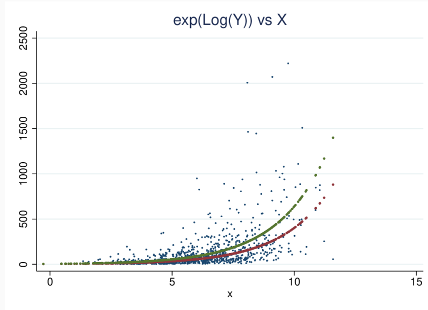

Predictions: predict log(Y) on log scale



Predictions: only $e^{\widehat{log(Y)}}$



Predictions: with correction



Predicting COVID-19

- We can apply log regression to the COVID-19 data
- A straight line on a log scale means a constant proportional increase.
- We can estimate this increase, regressing log(cases) on date.
- The slope, b, is the amount by which $\widehat{\log \text{cases}}$ rises per day
- e^b is then the multiplier by which `cases` rises per day

```
reg lcases date
```

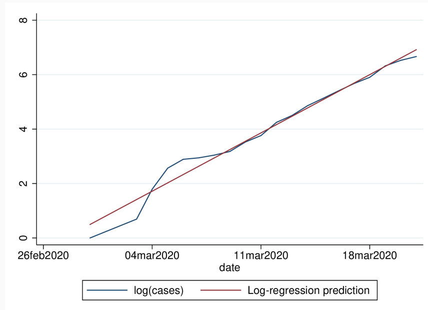

Stata output

```
. reg lc date

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =    746.82
       Model |  66.1088015         1  66.1088015   Prob > F        =    0.0000
    Residual |  1.59336573        18  .088520318   R-squared       =    0.9765
-------------+----------------------------------   Adj R-squared   =    0.9752
       Total |  67.7021673        19  3.56327196   Root MSE        =    .29752

------------------------------------------------------------------------------
          lc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        date |   .3058309   .0111911    27.33   0.000     .2823193    .3293426
       _cons |  -6719.833   246.0411   -27.31   0.000    -7236.746    -6202.92
------------------------------------------------------------------------------
```


Logs with log regression



Steady increase

The log of `cases` rises by 0.3058 per day

This means `cases` rises by a factor of $e^{0.3058} = 1.358$

The increase is 1.358 - 1 = 0.358, or almost 36% per day

Implies a doubling about every 2.6 days

But exponential increase is temporary

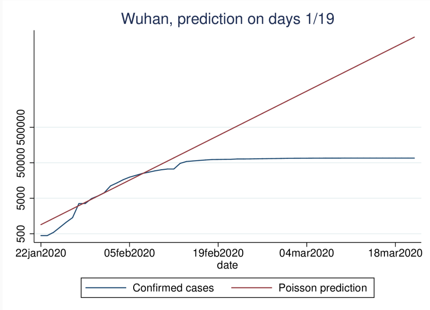
Exponential increase cannot go on indefinitely

Even if nothing is done, the rate of increase will decline as fewer people are left unexposed

And interventions (isolation, tracing) will reduce the rate

See China, for example

Wuhan, with prediction based on 1st 19 days



Summary

If there is a constant rate of increase, logs give us straight lines

Graph the log, or use a log scale on the Y-axis

Log regression allows us to estimate the rate

Exponential increase isn't forever, but modelling the exponential helps us see where the rate starts to drop

Code available here: `http://teaching.sociology.ul.ie/so5032/irecovid.do`