Biostatistics 200

Linear and Logistic Models
Transformations

Transformations

- One way to deal with this is to transform either x or y or both
- A common transformation is the log transformation
- Log transformations bring large values closer to the rest of the data
- There are methods to correct the standard errors for heteroscedasticity other than transformations

Log transformations

Value	Ln	Log_{10}
0	$-\infty$	$-\infty$
0.001	-6.91	-3.00
0.05	-3.00	-1.30
1	0.00	0.00
5	1.61	0.70
10	2.30	1.00
50	3.91	1.70
103	4.63	2.01

Be careful of log(0) or ln(0)
Be sure you know which log base your computer program is using

- In Stata log() will give you ln()
- If you want log10 use log10()

Transformations

- Transform FEV to ln(FEV)

```
. gen fev_ln=log(fev)
. summ fev fev_ln

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         fev |       654     2.63678    .8670591       .791      5.793
      fev_ln |       654     .915437    .3332652  -.2344573    1.75665
```

- Examine the histograms
- Run the regression of ln(FEV) on age and examine the residuals

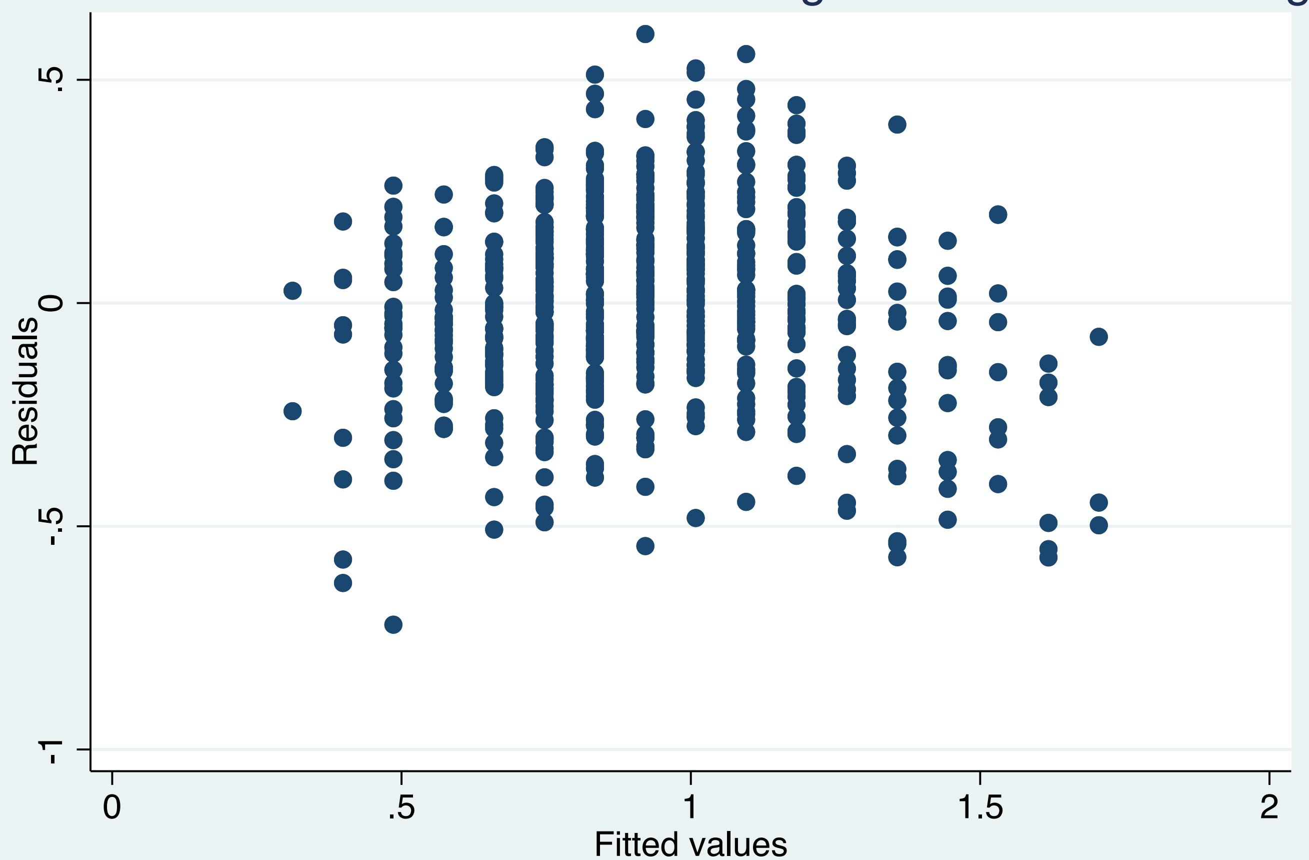
```
regress fev_ln age
rvfplot, title(Fitted values versus residuals for regression of
   lnFEV on age)
```

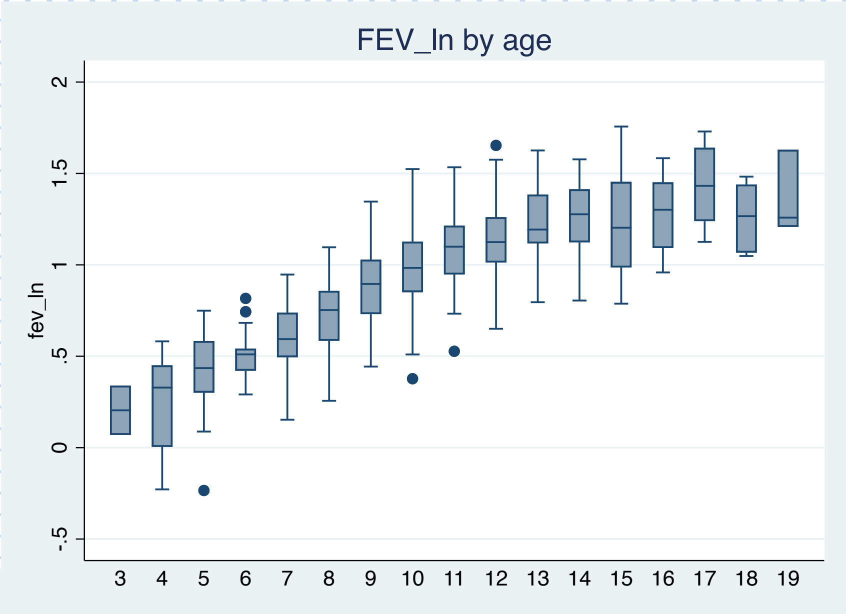
Regressions: fev & fev_ln

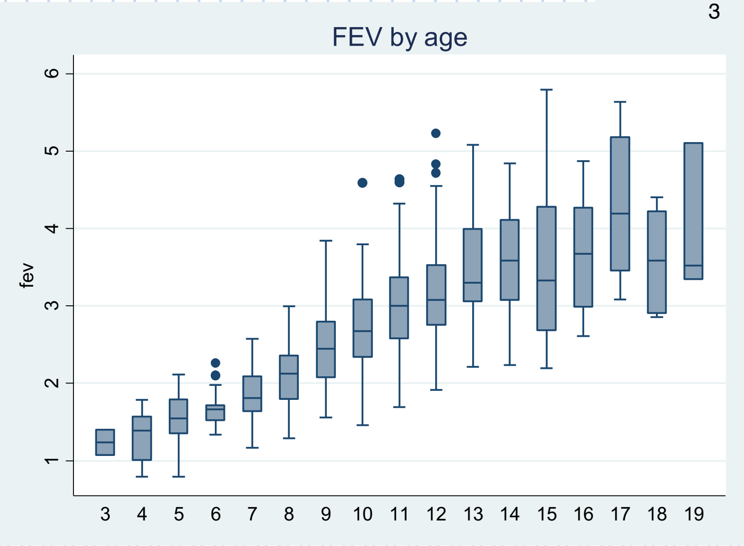
```
. regress fev_ln age

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =    961.01
       Model |  43.2100544         1  43.2100544   Prob > F        =    0.0000
    Residual |  29.3158601       652  .044962976   R-squared       =    0.5958
-------------+----------------------------------   Adj R-squared   =    0.5952
       Total |  72.5259145       653  .111065719   Root MSE        =    .21204

------------------------------------------------------------------------------
      fev_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0870833   .0028091    31.00   0.000     .0815673    .0925993
       _cons |    .050596    .029104     1.74   0.083    -.0065529    .1077449
------------------------------------------------------------------------------
```

```
. regress fev age

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =    872.18
       Model |  280.919154         1  280.919154   Prob > F        =    0.0000
    Residual |  210.000679       652  .322086931   R-squared       =    0.5722
-------------+----------------------------------   Adj R-squared   =    0.5716
       Total |  490.919833       653  .751791475   Root MSE        =    .56753

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .222041   .0075185    29.53   0.000     .2072777    .2368043
       _cons |   .4316481   .0778954     5.54   0.000      .278692    .5846042
------------------------------------------------------------------------------
```

Fitted values versus residuals for regression of lnFEV on age




FEV_ln by age
fev_ln
-.5
0
.5
1
1.5
2
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19


FEV by age
fev
1
2
3
4
5
6
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Interpretation of regression coefficients for transformed y value

- The regression equation is:

$$\ln(FEV) = \hat{\propto} + \hat{\beta} * age$$

$$= 0.051 + 0.087 \text{ age}$$

- So a one year change in age corresponds to a .087 change in ln(FEV)
- The change is on a multiplicative scale, so if you exponentiate, you get a percent change in y
- $e^{0.087}$ = 1.091 or a 9.1% change in FEV

For example…

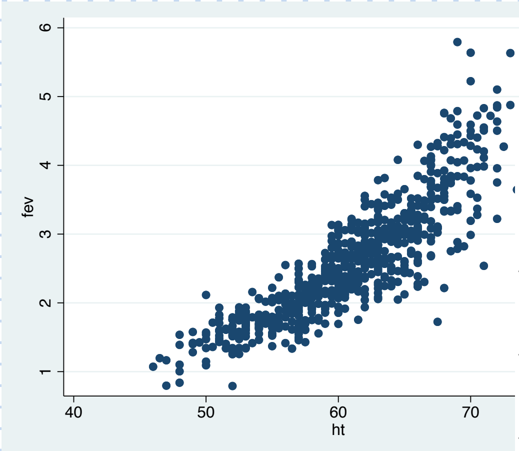
- $\text{Ln}(FEV) = 0.051 + 0.087\ \text{age}$

- $\text{Ln}(FEV_{age21}) = 0.051 + 0.087*21$
- $\text{Ln}(FEV_{age20}) = 0.051 + 0.087*20$
- $\text{Ln}(FEV_{age21})-\ln(FEV_{age20}) = 0.087$

  Remember $\ln(a)-\ln(b) = \ln(a/b)$
- $\text{Ln}(FEV_{age21}/FEV_{age20})= 0.087$
- $FEV_{age21}/FEV_{age20} = e^{0.087} = 1.09$

Now using height as the independent variable

1. Make a scatter plot of FEV by height and FEV_ln by height
2. Run a regression of FEV on height and FEV_ln on height
3. What does this tell you?

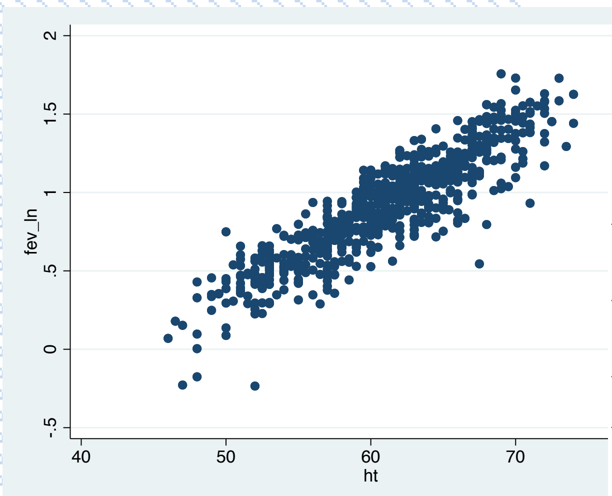


. regress fev ht

Source	SS	df	MS
Model	369.985854	1	369.985854
Residual	120.933979	652	.185481563
Total	490.919833	653	.751791475

Number of obs	=	654
F(1, 652)	=	1994.73
Prob > F	=	0.0000
R-squared	=	0.7537
Adj R-squared	=	0.7533
Root MSE	=	.43068

fev	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ht	.1319756	.002955	44.66	0.000	.1261732	.137778
_cons	-5.432679	.1814599	-29.94	0.000	-5.788995	-5.076363



. regress fev_ln ht

Source	SS	df	MS
Model	57.7021467	1	57.7021467
Residual	14.8237679	652	.02273584
Total	72.5259145	653	.111065719

Number of obs	=	654
F(1, 652)	=	2537.94
Prob > F	=	0.0000
R-squared	=	0.7956
Adj R-squared	=	0.7953
Root MSE	=	.15078

fev_ln	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ht	.0521191	.0010346	50.38	0.000	.0500876	.0541506
_cons	-2.271312	.063531	-35.75	0.000	-2.396062	-2.146562

Categorical independent variables

- We previously noted that the independent variable (the X variable) does not need to be normally distributed
- In fact, this variable can be categorical
- Coding: Dichotomous variables in regression models
  - 1 to represent the level of interest
  - 0 to represent the comparison or reference group.

  These 0-1 variables are called indicator or dummy variables.
- The regression model is the same
- The interpretation is slightly different
  - β is the change in y that corresponds to being in the group of interest vs. the reference category

Categorical independent variables

- Example sex: female x_{sex}=0, for male x_{sex} =1
- Regression of FEV and sex
- $\hat{fev} = \hat{\alpha} + \hat{\beta}\, x_{sex}$
- For male: $\hat{fev}_{male} = \hat{\alpha} + \hat{\beta}*1 = \hat{\alpha} + \hat{\beta}$
- For female: $\hat{fev}_{female} = \hat{\alpha} + \hat{\beta}*0 = \hat{\alpha}$

  So $\hat{fev}_{male} - \hat{fev}_{female} = \hat{\alpha} + \hat{\beta} - \hat{\alpha} = \hat{\beta}$

- Remember, $\hat{\alpha}$ is the mean value of y when x=0

  So here it is the mean FEV for sex=female

```
. regress fev sex

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =     29.61
       Model |  21.3239848         1  21.3239848   Prob > F        =    0.0000
    Residual |  469.595849       652  .720239032   R-squared       =    0.0434
-------------+----------------------------------   Adj R-squared   =    0.0420
       Total |  490.919833       653  .751791475   Root MSE        =    .84867

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .3612766   .0663963     5.44   0.000     .2309002     .491653
       _cons |    2.45117    .047591    51.50   0.000      2.35772     2.54462
------------------------------------------------------------------------------
```

1. What is the estimate for beta? How is it interpreted?
2. What is the estimate for alpha? How is it interpreted?
3. What hypothesis is tested where it says P>|t|?
4. How much of the variance in FEV is explained by sex?

Categorical independent variable

- Remember that the regression equation is

  $\mu_{y|x} = \alpha + \beta\, x$
- The only variables x can take are 0 and $\mu_{y|0} = \alpha$ $\mu_{y|1} = \alpha + \beta$
- The estimated mean FEV for females is $\hat{\alpha}$ and the estimated mean FEV for males is $\widehat{\alpha} + \widehat{\beta}$
- What other test have we learned that tests the same thing?

```
. ttest fev, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     318     2.45117    .0362111     .645736    2.379925    2.522414
       1 |     336    2.812446    .0547507    1.003598    2.704748    2.920145
---------+--------------------------------------------------------------------
combined |     654     2.63678    .0339047    .8670591    2.570204    2.703355
---------+--------------------------------------------------------------------
    diff |           -.3612766    .0663963                -.491653   -.2309002
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.4412
Ho: diff = 0                                     degrees of freedom =      652

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

```
. regress fev sex

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =     29.61
       Model |  21.3239848         1  21.3239848   Prob > F        =    0.0000
    Residual |  469.595849       652  .720239032   R-squared       =    0.0434
-------------+----------------------------------   Adj R-squared   =    0.0420
       Total |  490.919833       653  .751791475   Root MSE        =    .84867

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .3612766   .0663963     5.44   0.000     .2309002     .491653
       _cons |    2.45117    .047591    51.50   0.000      2.35772     2.54462
------------------------------------------------------------------------------
```

Categorical independent variables

- e.g. Race group = White, Asian/PI, Other
- If Race=White is set as reference category, dummy variables look like:
- **Then the regression equation is:**

  $y = \alpha + \beta_1 x_{Asian/PI} + \beta_2 x_{Other} + \varepsilon$

	$x_{Asian/PI}$	x_{Other}
White	0	0
Asian/PI	1	0
Other	0	1

- For race group=White

  $\hat{y} = \hat{\alpha} + \widehat{\beta_1}0 + \widehat{\beta_2}0 = \hat{\alpha}$
- For race group=Asian/PI

  $\hat{y} = \hat{\alpha} + \widehat{\beta_1}1 + \widehat{\beta_2}0 = \hat{\alpha} + \widehat{\beta_1}$
- For race group=other

  $\hat{y} = \hat{\alpha} + \widehat{\beta_1}0 + \widehat{\beta_2}1 = \hat{\alpha} + \widehat{\beta_2}$

Stata indicator variables

- You actually don't have to make the dummy variables yourself
- All you have to do is tell Stata that a variable is categorical using i. before a variable name
- The regression of FEV regressed on age group

  regress fev i.agecat

```
. regress fev i.agecat

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(4, 649)       =    191.07
       Model |  265.481813         4  66.3704533   Prob > F        =    0.0000
    Residual |   225.43802       649  .347362127   R-squared       =    0.5408
-------------+----------------------------------   Adj R-squared   =    0.5380
       Total |  490.919833       653  .751791475   Root MSE        =    .58937

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
         6-  |   .4713427    .104309     4.52   0.000     .2665188    .6761665
         9-  |   1.244846   .1010818    12.32   0.000     1.046359    1.443332
        12-  |   1.912191      .1081    17.69   0.000     1.699923    2.124459
        15-  |   2.237758   .1264743    17.69   0.000      1.98941    2.486106
             |
       _cons |   1.472385   .0943754    15.60   0.000     1.287067    1.657703
------------------------------------------------------------------------------

.
```