Biostat 200
Lecture 11

Today

- Linear regression: Categorical explanatory variables
- Multiple regression
- Logistic regression
- Wrap up
- Please do evaluations on moodle

From last time

- Plotting residuals in Stata
  - regress fev ht
  - rvfplot ** gives you Residuals vs. Fitted
  - rvpplot ht ** gives you Residuals vs. Predictor (indep var)

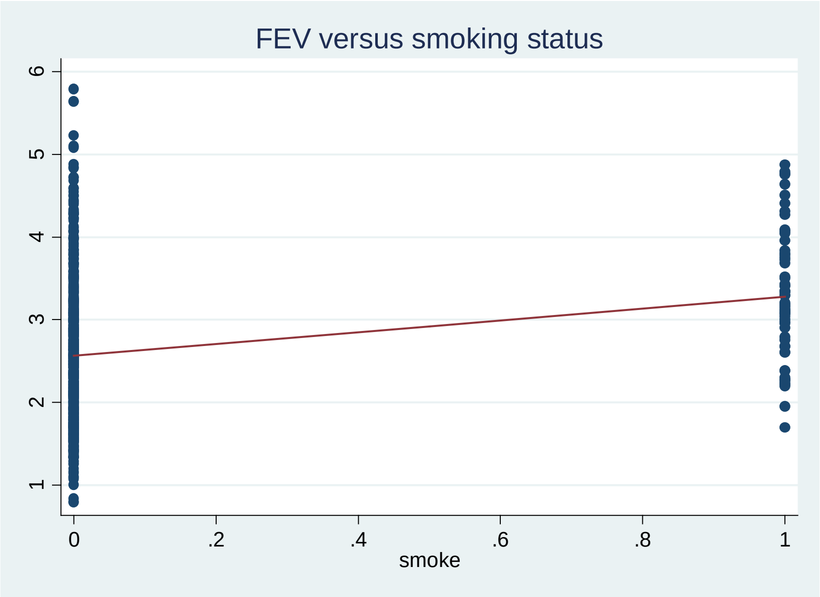
Linear regression with categorical explanatory variable

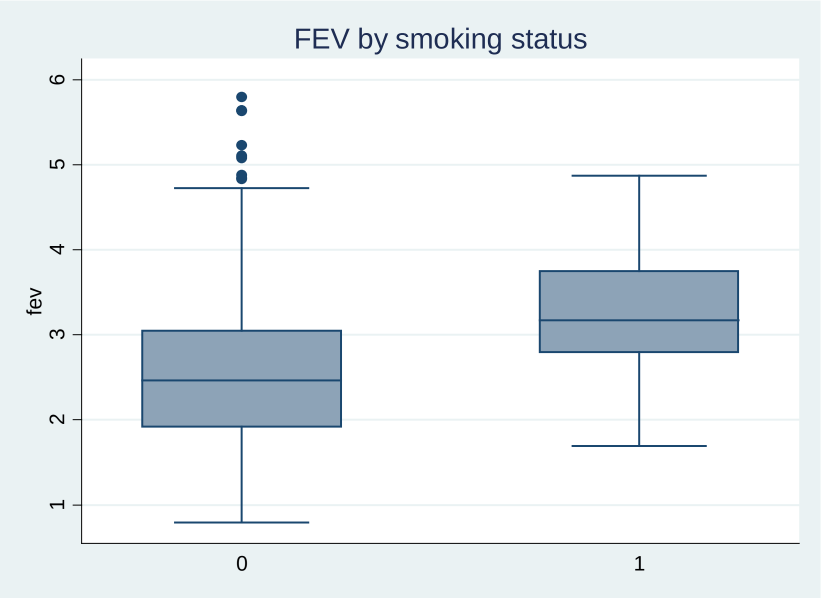
```
. regress fev smoke

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  1,   652) =   41.79
       Model |   29.569683     1   29.569683           Prob > F      =  0.0000
    Residual |   461.35015   652  .707592255           R-squared     =  0.0602
-------------+------------------------------           Adj R-squared =  0.0588
       Total |  490.919833   653  .751791475           Root MSE      =  .84119

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .7107189   .1099426     6.46   0.000     .4948346    .9266033
       _cons |   2.566143   .0346604    74.04   0.000     2.498083    2.634202
------------------------------------------------------------------------------
```

Regression with categorical predictors





- There is a lot of variability in each group here
- Smoking yes/no may be a crude measure

Multiple regression

- Additional explanatory variables might add to our understanding of a dependent variable
- We can posit the population equation

  $$\mu_{y|x1,x2,\ldots,xq} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q$$

- α is the mean of y when all the explanatory variables are 0
- β_i is the change in the mean value of y the corresponds to a 1 unit change in x_i when all the other explanatory variables are held constant

- Because there is natural variation in the response variable, the model we fit is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon$$

- Assumptions
  - $x_1, x_2, \ldots, x_q$ are measured without error
  - The distribution of y is normal with mean $\mu_{y|x1,x2,\ldots,xq}$ and standard deviation $\sigma_{y|x1,x2,\ldots,xq}$
  - The population regression model holds
  - For any set of values of the explanatory variables, $x_1, x_2, \ldots, x_q$, $\sigma_{y|x1,x2,\ldots,xq}$ is constant – homoscedasticity
  - The outcomes are independent

You can fit continuous and/or categorical variables

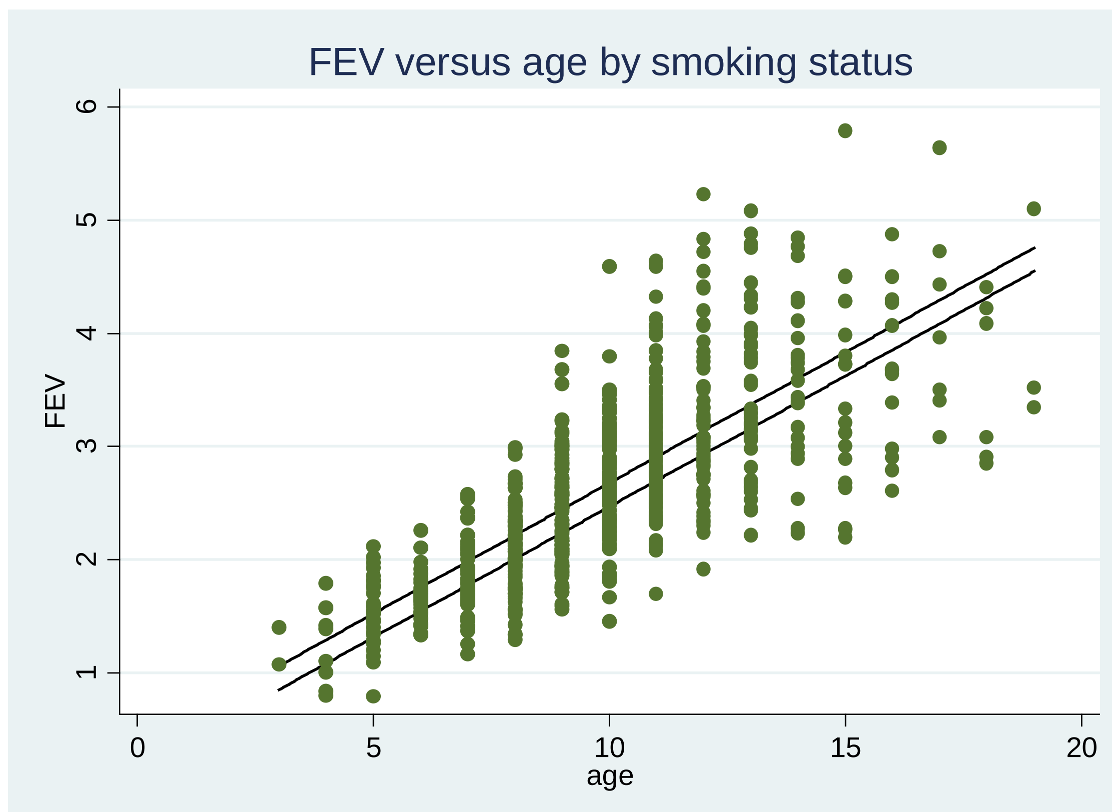
```
. regress fev age smoke

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  2,   651) =  443.25
       Model |  283.058247     2  141.529123           Prob > F      =  0.0000
    Residual |  207.861587   651  .319295832           R-squared     =  0.5766
-------------+------------------------------           Adj R-squared =  0.5753
       Total |  490.919833   653  .751791475           Root MSE      =  .56506

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2306046   .0081844    28.18   0.000     .2145336    .2466755
       smoke |  -.2089949   .0807453    -2.59   0.010    -.3675476   -.0504421
       _cons |   .3673731   .0814357     4.51   0.000     .2074647    .5272814
------------------------------------------------------------------------------
```

- The model is $\hat{fev} = \hat{\alpha} + \beta_1$ age $+ \beta_2 X_{smoke}$
- So for non-smokers, we have $\hat{fev} = \hat{\alpha} + \beta_1$ age
- For smokers, $\hat{fev} = \hat{\alpha} + \beta_1$ age $+ \beta_2$
  - So β_2 is the mean difference in FEV for smokers versus non-smokers <u>at each age</u>

- When you have one continuous variable and one dichotomous variable, you can think of fitting two lines that only differ in y intercept by the coefficient of the dichotomous variable
- E.g. β_2=-.209



Logistic regression

- For linear regression
  - The predicted values can be from $-\infty$ to $+\infty$
  - Y values are assumed to follow a normal distribution

- For many situations we want to model a dichotomous outcome, such as disease/no disease and cannot use linear regression

- We want to model p, the probability of disease

Logistic regression

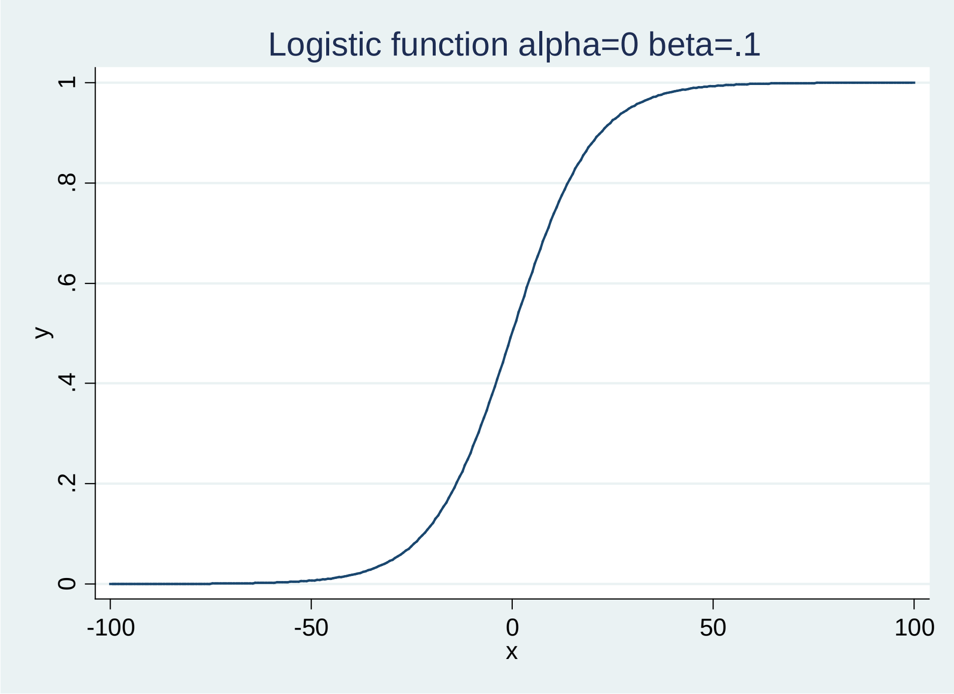
- p=P(Y=1)   i.e. P(outcome=disease)

- A model of the form p= $\alpha + \beta x$ would be able to take on negative values or values more than 1

- p=e $^{\alpha + \beta x}$ is an improvement because it cannot be negative , but it still could be greater than 1

Logistic regression

- How about the function? $$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$



- This function =.5 when *α + β x =0*
- The function models the probability slowly increasing over the values of X, until there is a steep rise, and another leveling off

Logistic regression

- The logistic function $$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

- Now check this out:

$$1-p = 1-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1+e^{\alpha+\beta x}-e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{\alpha+\beta x}}$$

- So odds of disease/success are

$$\frac{p}{1-p} = \frac{\dfrac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{\dfrac{1}{1+e^{\alpha+\beta x}}} = e^{\alpha+\beta x}$$

- And therefore ln(p/(1-p)) = $\alpha + bx$

Logistic regression

- So instead of assuming that the relationship between X and p is linear , we are assuming that the relationship between ln(p/(1-p)) and X is linear.

- ln(p/(1-p)) is called the <u>logit</u> function

Logistic regression

- This transformation of the outcome p is called the logit link

- It is part of a family of models called generalized linear models

- In generalized linear models, some function of the outcome variable is assumed to be linear

  e.g. $f(\mathrm{y}) = \alpha + bx$ or $f(\mathrm{p}) = \alpha + bx$

  e.g. $\mathrm{logit(p)} = \alpha + bx$

Logistic regression

- Assumptions
  - Y_i follows a bernoulli distribution (0,1)
  - The mean of Y at every X, P(Y=1|x) is given by the logistic function
  - The values of the outcome are independent

Logistic regression

- Estimation
  - The parameters are estimated via a procedure called *maximum likelihood*
  - There is not a closed form solution of the maximization (i.e. we cannot write down a solution) – it is an iterative procedure

Notation

- p_0 = P(Y=1|X=0)

  is the probability that y=1 when x=0

- p_1 = P(Y=1|X=1)

  is the probability that y=1 when x=1

- p = P(Y=1)

  is the probability that y=1 for any x

Interpretation of coefficients

- One dichotomous explanatory variable
- For x=0

  $\ln(P(Y=1|X=0)/(1- P(Y=1|X=0)))$

  $= \ln(p_0/(1-p_0))$

  $= \alpha + \beta*0 = \alpha$
- For x=1

  $\ln(P(Y=1|X=1)/(1- P(Y=1|X=1)))$

  $= \ln(p_1/(1-p_1))$

  $= \alpha + \beta*1 = \alpha + \beta$

Interpretation of coefficients

$$\ln\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}\right) = \ln(OR)$$

- $\ln[(p_1/(1-p_1)/(p_0/(1-p_0)]$

  $= (\ln(p_1/(1-p_1)) - \ln(p_0/(1-p_0)) =$

  $= (\alpha + \beta) - \alpha = \beta$

→ $OR = e^{\beta}$

*Remember that ln(a/b) = ln(a)-ln(b)

Correlates of having had a cold in the prior month

```
. logistic coldany i.rested_mostly

Logistic regression                                 Number of obs   =        452
                                                    LR chi2(1)      =       2.43
                                                    Prob > chi2     =     0.1188
Log likelihood = -272.65565                         Pseudo R2       =     0.0044

------------------------------------------------------------------------------
     coldany | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.rested_m~y |   .6999519    .162216    -1.54   0.124     .4444262    1.102394
------------------------------------------------------------------------------
```

Correlates of having had a cold in the prior month

logistic depvar indepvar

```
. logistic coldany i.rested_mostly

Logistic regression                               Number of obs   =        452
                                                  LR chi2(1)      =       2.43
                                                  Prob > chi2     =     0.1188
Log likelihood = -272.65565                       Pseudo R2       =     0.0044

------------------------------------------------------------------------------
     coldany | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.rested_m~y |   .6999519    .162216    -1.54   0.124     .4444262    1.102394
------------------------------------------------------------------------------
```

$=e^{\beta}$

This is used in comparing nested models of the same data set

Test of whether the model fit differs from the model with only α and no covariates

The Pseudo R^2 is an attempt to be analogous with linear regression but it is not

Stata notes

- logistic depvar indepvars
  - gives you odds ratios
- logistic depvar indepvars, coef
  - gives you coefficients (alphas and betas)
- logit depvar indepvars
  - gives you coefficients
- logit depvar indepvars, or
  - gives you odds ratios

R-square statistic in logistic regression

“In OLS regression, the R-square statistic indicates the proportion of the variability in the dependent variable that is accounted for by the model (i.e., all of the independent variables in the model). Unfortunately, creating a statistic to provide the same information for a logistic regression model has proved to be very difficult. Many people have tried, but no approach has been widely accepted by researchers or statisticians. The output from the **logit** and **logistic** commands give a statistic called "pseudo-R-square", and the emphasis is on the term "pseudo". This statistic should be used only to give the most general idea as to the proportion of variance that is being accounted for. ... There is little agreement regarding an R-square statistic in logistic regression, and that different approaches lead to very different conclusions. If you use an R-square statistic at all, use it with great care. “

http://www.ats.ucla.edu/stat/stata/webbooks/logistic/chapter1/statalog1.htm

Correlates of having had a cold in the prior month

```
. tabodds coldany rested_mostly, or

---------------------------------------------------------------------------
rested_mos~y |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
          No |    1.000000          .            .              .         .
         Yes |    0.699952       2.38       0.1231       0.443668   1.104276
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =     2.38
                                  Pr>chi2  =   0.1231

Score test for trend of odds:     chi2(1)  =     2.38
                                  Pr>chi2  =   0.1231
```

For <u>one independent variable that is categorical</u>, the odds ratio estimate is the same for logistic regression as for tabular methods.

Continuous explanatory variable

```
.
. . logistic coldany age

Logistic regression                               Number of obs   =        451
                                                  LR chi2(1)      =       9.49
                                                  Prob > chi2     =     0.0021
Log likelihood = -268.77615                       Pseudo R2       =     0.0174

------------------------------------------------------------------------------
     coldany | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9739414   .0088552    -2.90   0.004     .9567392    .9914529
------------------------------------------------------------------------------
```

- Interpretation of the coefficients: The odds ratio is for a one unit change in the predictor
- For this example the 0.974 is the odds ratio for a year difference in age

Continuous explanatory variable

- When the explanatory variable is a continuous variable, there is no equivalent tabular method without dividing the variable into groups (e.g. age groups)

- You need to use logistic regression when you want to model the relationship between a dichotomous outcome and a continuous explanatory variable

Continuous explanatory variable

- To find the OR for a 10-year change in age

```
.

. logistic coldany age, coef

Logistic regression                               Number of obs   =        451
                                                  LR chi2(1)      =       9.49
                                                  Prob > chi2     =     0.0021
Log likelihood = -268.77615                       Pseudo R2       =     0.0174

------------------------------------------------------------------------------
     coldany |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0264041   .0090922    -2.90   0.004    -.0442244   -.0085838
       _cons |   .0902805    .338474     0.27   0.790    -.5731163    .7536773
------------------------------------------------------------------------------
```

OR for a 10-year change in age = exp(10*-.0264) = 0.768

- To find the OR for a 10-year change in age

```
. gen age_10=age/10
(2 missing values generated)

. logistic coldany age_10

Logistic regression                               Number of obs   =        451
                                                  LR chi2(1)      =       9.49
                                                  Prob > chi2     =     0.0021
Log likelihood = -268.77615                       Pseudo R2       =     0.0174

------------------------------------------------------------------------------
     coldany | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   .7679418   .0698225    -2.90   0.004     .6425926    .9177426
------------------------------------------------------------------------------
```

Checking logit-linearity

```
. tab agecat

  RECODE of |
        age |      Freq.     Percent        Cum.
------------+-----------------------------------
      18-30 |        161       35.54       35.54
      30-45 |        202       44.59       80.13
      45-60 |         42        9.27       89.40
        >60 |         48       10.60      100.00
------------+-----------------------------------
      Total |        453      100.00

. logistic coldany i.agecat

Logistic regression                               Number of obs   =        451
                                                  LR chi2(3)      =       8.45
                                                  Prob > chi2     =     0.0375
Log likelihood = -269.29729                       Pseudo R2       =     0.0154

------------------------------------------------------------------------------
     coldany | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
          2  |   1.037197   .2348894     0.16   0.872     .6654165    1.616699
          3  |   .6011029   .2486731    -1.23   0.219     .2671845    1.352342
          4  |   .3648981   .1615593    -2.28   0.023     .1532147    .8690459
------------------------------------------------------------------------------
```

Logistic regression with >1 explanatory variable (multiple logistic regression)

- In multiple logistic regression, the probability of success depends on more than one explanatory variable

$$p = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}$$

$$\text{and } \frac{p}{1-p} = e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}$$

- Therefore

$$\ln(p/(1-p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Interpretation of the coefficients

- Example: 2 independent variables, X_1 a numerical variable, X_2 a dichotomous exposure variable

- $\text{logit}(p) = \log(p/(1-p)) = \alpha + \beta_1 x_1 + \beta_2 x_2$

- When $x_2=1$ (exposed), holding x_1 constant,

$$\text{logit}(p_1) = \alpha + \beta_1 x_1 + \beta_2$$

- When $x_2=0$ (not exposed), holding x_1 constant,

$$\text{logit}(p_0) = \alpha + \beta_1 x_1$$

Interpretation of the coefficients

So $\mathrm{logit}(p_1/p_0) = (\alpha + \beta_1 x_1 + \beta_2) - (\alpha + \beta_1 x_1) = \beta_2$

So

$$\log\left(\frac{\dfrac{p_1}{1-p_1}}{\dfrac{p_0}{1-p_0}}\right) = \beta_2$$

$$\text{so } OR = e^{\beta_2} \text{ adjusted for } \mathrm{x}_1$$

```
. logistic coldany  age_10 rested_mostly

Logistic regression                               Number of obs   =        450
                                                  LR chi2(2)      =      10.67
                                                  Prob > chi2     =     0.0048
Log likelihood = -266.96282                       Pseudo R2       =     0.0196

------------------------------------------------------------------------------
     coldany | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   .7774313   .0715981    -2.73   0.006     .6490378    .9312238
rested_mos~y |   .7685218    .181155    -1.12   0.264     .4841857    1.219833
------------------------------------------------------------------------------
```

For this example the 0.777 is the odds ratio for a ten-year difference in age when you hold rested status constant

0.768 is the odds ratio for being well rested when you hold age constant

- Want more on linear regression and logistic regression? Go to

  http://www.ats.ucla.edu/stat/stata/webbooks/

- Explore correlates of having a cold in our class data set. Hours of sleep? Having children under the age of 6? These while holding age constant?

Statistical hypothesis tests

Data and comparison type	Alternative hypotheses	Parametric test Stata command	Non-parametric test Stata command
Numerical; One mean	H_a: $\mu \neq \mu_a$ (two-sided) H_a: $\mu > \mu_a$ or $\mu < \mu_a$ (one-sided)	Z or t-test • ttest var1=hypoth val.*	
Numerical; Two means, paired data	H_a: $\mu_1 \neq \mu_2$ (two-sided) H_a: $\mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	Paired t-test • ttest var1=var2*	Sign test • signtest var1=var2 • Wilcoxon Signed-Rank signrank var1=var2)
Numerical; Two means, independent data	H_a: $\mu_1 \neq \mu_2$ (two-sided) H_a: $\mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	T-test (equal or unequal variance) • ttest var1, by(byvar) unequal	Wilcoxon rank-sum test • ranksum var1, by(byvar)
Numerical, Two or more means, independent data	H_a: $\mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$ or $\mu_2 \neq \mu_3$ etc.	ANOVA • oneway var1 byvar	Kruskal Wallis test • kwallis var1, by(byvar)
Dichotomous; One proportion	H_a: $p \neq p_a$ (two-sided) H_a: $p > p_a$ or $p < p_a$ (one-sided)	Proportion test • prtest var1=hypoth value* • bitest var1=hypoth value	
Dichotomous; two proportions	H_a: $p_1 \neq p_2$ (two-sided) H_a: $p_1 > p_2$ (one-sided)	Proportion test (z-test) • prtest var1, by(byvar)	Chi-square test • tab var1 var2, chi exact McNemar's for paired data: • mcc var1 var2
Categorical by categorical (nxk)	H_a : The rows not independent of the columns		Chi-square test • tab var1 var2, chi exact

Further methods

Dependent variable (y)	**Independent variable(s) type(s) (x)**	**Methods**	**Non-parametric methods**
Numerical/continuous	One numerical/continuous variable	Pearson correlation; simple linear regression	Spearman rank correlation
Numerical/continuous	One categorical variable	Linear regression, t-test, ANOVA	Sign test, Signed rank test, Wilcoxon rank sum, Kruskal-Wallis
Numerical/continuous	Multiple variables/types of variables	Linear regression	
Dichotomous	One categorical variable	Logistic regression (or proportion test for 2x2)	Chi-square test or Fisher exact test
Dichotomous	One numerical/continuous variable	Logistic regression	
Dichotomous	Multiple variables/types of variables	Logistic regression	

Last sessions

- Thursday, Dec. 5 – no lab
- Final exam
  - <u>Take home exam </u>will be posted online on Thursday Dec. 5
  - The exam is worth 30% of your grade.
  - <u>NO COLLABORATIONS WITH YOUR CLASSMATES OR OTHERS</u>
  - Exam due by e-mail to TAs Thursday, December 12, 10:30 a.m.
- Tuesday Dec. 10 – office hours <u>by appointment</u>

Thank you and good luck!