

Biostat 200

Lecture 11

Today

- Linear regression: Categorical explanatory variables
- Multiple regression
- Logistic regression
- Wrap up
- Please do evaluations on qualtrics – Olivia deLeon sent you the link

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)

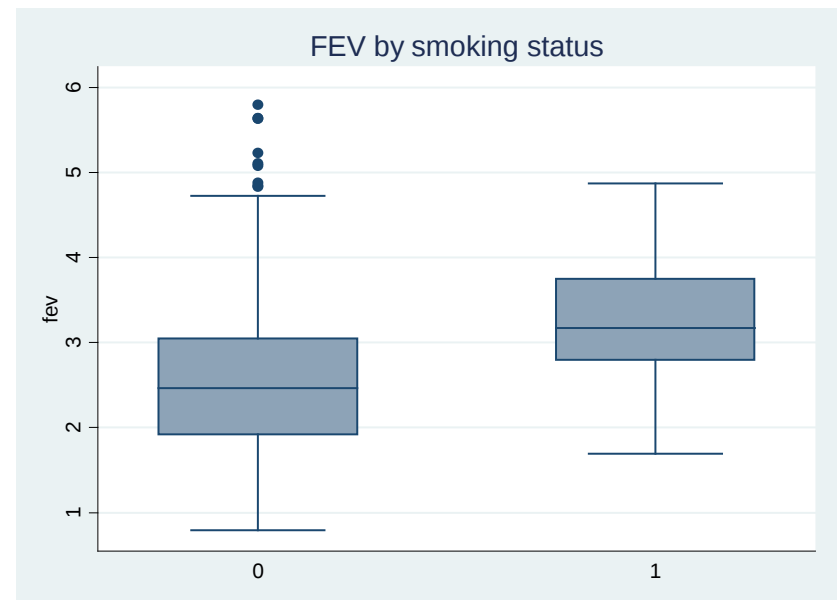
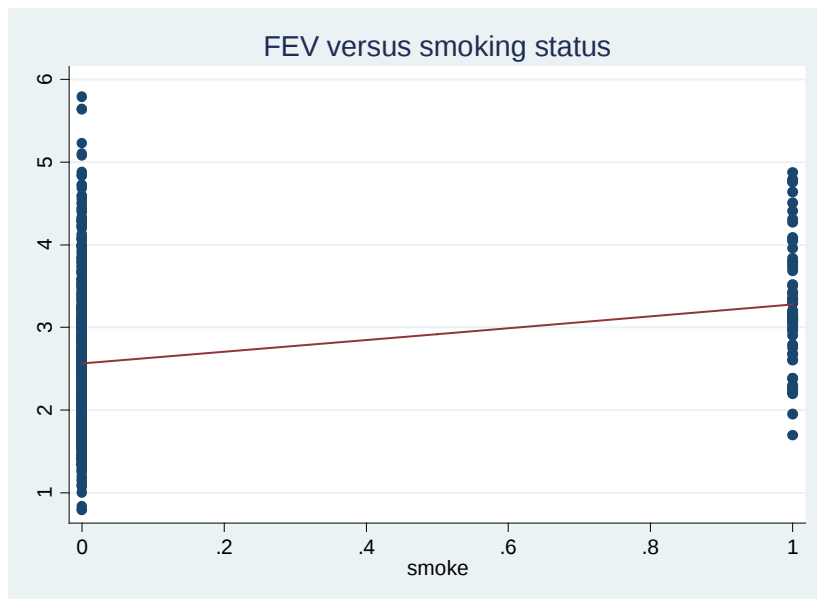
Linear regression with categorical explanatory variable

```
. regress fev smoke
```

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

$$\mu_{y/x_1, x_2, \dots, x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q$$

- α is the mean of y when all the explanatory variables are 0
- β_i is the change in the mean value of y that 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

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

- Assumptions
 - $x_1, x_2, \dots, x_q$ are measured without error
 - The distribution of y is normal with mean $\mu_{y/x_1, x_2, \dots, x_q}$ and standard deviation $\sigma_{y/x_1, x_2, \dots, x_q}$
 - The population regression model holds
 - For any set of values of the explanatory variables, $x_1, x_2, \dots, x_q$, $\sigma_{y/x_1, x_2, \dots, x_q}$ 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
Model	283.058247	2	141.529123	F(2, 651)	=	443.25
Residual	207.861587	651	.319295832	Prob > F	=	0.0000
Total	490.919833	653	.751791475	R-squared	=	0.5766
				Adj R-squared	=	0.5753
				Root MSE	=	.56506

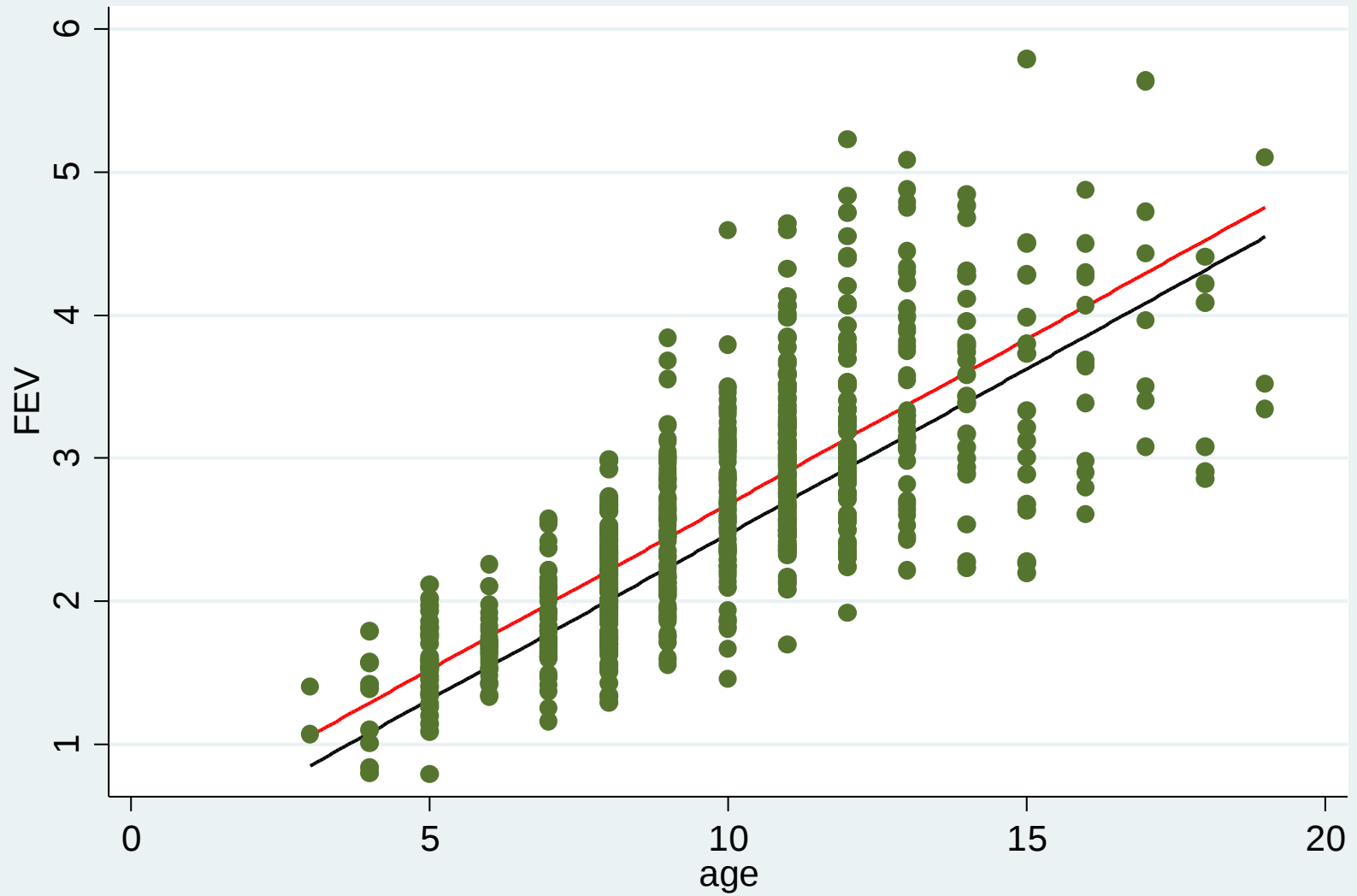
fev	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.2306046	.0081844	28.18	0.000	.2145336	.2466755
smoke	-.2089949	.0807453	-2.59	0.010	-.3675476	-.0504421
_cons	.3673731	.0814357	4.51	0.000	.2074647	.5272814

- The model is $\hat{fev} = \hat{\alpha} + \beta_1 \text{ age} + \beta_2 X_{\text{smoke}}$
- So for non-smokers, we have $\hat{fev} = \hat{\alpha} + \beta_1 \text{ age} + \beta_2 * 0$
- For smokers, $\hat{fev} = \hat{\alpha} + \beta_1 \text{ age} + \beta_2 * 1$
 - So β_2 is the mean difference in FEV for smokers versus non-smokers at each age

- 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. $\beta_2 = -.209$



FEV versus age by smoking status



Logistic regression

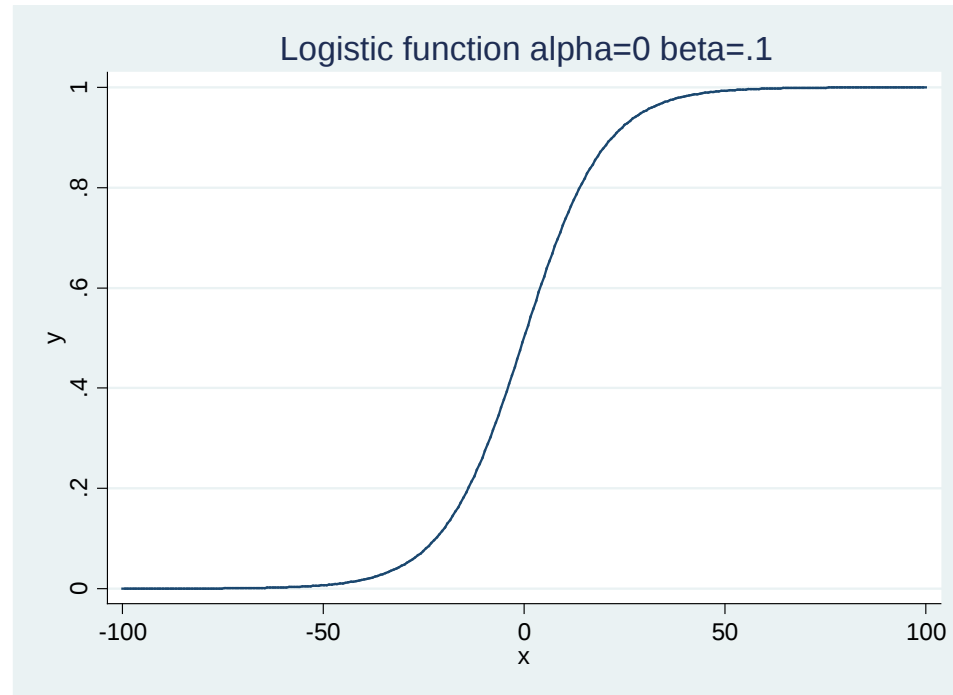
- For linear regression
 - The predicted values can be from $-\infty$ to $+\infty$
 - Y values are assumed to follow a normal distribution
- For many situations we want to model a dichotomous outcome, such as disease/no disease and cannot use linear regression
- We want to model p , the probability of disease

Logistic regression

- $p = P(Y=1)$ i.e. $P(\text{outcome}=\text{disease})$
- 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 $\alpha + \beta x = 0$
- 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

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

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

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 logit 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. $f(y) = \alpha + \beta x$ or $f(p) = \alpha + \beta x$

e.g. $\text{logit}(p) = \alpha + \beta x$

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 outcomes 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

- $p = P(Y=1)$
is the probability that $y=1$ for any x
- $p_0 = P(Y=1 | X=0)$
is the probability that $y=1$ when $x=0$
- $p_1 = P(Y=1 | X=1)$
is the probability that $y=1$ when $x=1$

Interpretation of coefficients

- One dichotomous explanatory variable
- For $x=0$

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

= odds of the outcome when $x=0$

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

$$= \alpha + \beta * x = \alpha + \beta * 0 = \alpha$$

- For $x=1$

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

= odds of the outcome when $x=1$

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

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

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 $\ln(a/b) = \ln(a) - \ln(b)$

Difference between treatment arms in PEth outcome (>50 ng/ml)

```
logistic peth_50 i.month
```

Logistic regression

Number of obs = 287
LR chi2(1) = 0.03
Prob > chi2 = 0.8534
Pseudo R2 = 0.0001

Log likelihood = -197.99359

peth_50	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
88.month	.955665	.2346518	-0.18	0.853	.5906139	1.54635
_cons	.8659794	.1290688	-0.97	0.334	.6466083	1.159775

Correlates of having had a cold in the prior month

logistic depvar indepvar

logistic peth_50 i.month

Logistic regression

Number of obs = 287
LR chi2(1) = 0.03
Prob > chi2 = 0.8534
Pseudo R2 = 0.0001

Log likelihood = -197.99359

peth_50	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
88.month	.955665	.2346518	-0.18	0.853	.5906139	1.54635
_cons	.8659794	.1290688	-0.97	0.334	.6466083	1.159775

$=e^{\beta}$

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>

Odds ratio of PEth $\geq$ 50 for comparison vs. standard group

tabodds peth_50 month, or

month	Odds Ratio	chi2	P>chi2	[95% Conf. Interval]	
6	1.000000	.	.	.	.
88	0.955665	0.03	0.8537	0.590101	1.547695
Test of homogeneity (equal odds):					
		chi2(1)	=	0.03	
		Pr>chi2	=	0.8537	
Score test for trend of odds:					
		chi2(1)	=	0.03	
		Pr>chi2	=	0.8537	

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

Continuous explanatory variable

```
. logistic peth_50 age_b
```

Logistic regression

Number of obs = 287
LR chi2(1) = 11.03
Prob > chi2 = 0.0009
Pseudo R2 = 0.0278

Log likelihood = -192.49767

peth_50	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
age_b	1.046077	.0146395	3.22	0.001	1.017774	1.075167
_cons	.2060976	.0940445	-3.46	0.001	.0842676	.5040633

- Interpretation of the coefficients: The odds ratio is for a one unit change in the predictor
- For this example 1.05 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 peth_50 age_b, coef
```

```
Logistic regression                                Number of obs   =          287  
                                                    LR chi2(1)      =          11.03  
                                                    Prob > chi2     =          0.0009  
Log likelihood = -192.49767                        Pseudo R2      =          0.0278
```

```
-----  
      peth_50 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]  
-----+-----  
      age_b   |   .0450471   .0139947     3.22   0.001    .0176181    .0724762  
      _cons   |  -1.579406   .4563106    -3.46   0.001   -2.473758   -.6850534  
-----
```

OR for a 10-year change in age = $\exp(10 \times -.045) = 1.57$

```
.
```

Continuous explanatory variable

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

```
. lincom age*10
```

```
( 1) 10*[peth_50]age_b = 0
```

peth_50		Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
-----+-----						
(1)		1.569051	.2195834	3.22	0.001	1.192654 2.064239

- Another way to find the OR for a 10-year change in age

```
gen age_10=age_b/10
```

```
. logistic peth_50 age_10
```

Logistic regression

```
Number of obs   =      287
LR chi2(1)      =      11.03
Prob > chi2     =      0.0009
Pseudo R2       =      0.0278
```

Log likelihood = -192.49767

peth_50	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age_10	1.569051	.2195834	3.22	0.001	1.192654	2.064239
_cons	.2060976	.0940445	-3.46	0.001	.0842676	.5040633



Checking logit-linearity

```
. tab agecat
```

agecat	Freq.	Percent	Cum.
-----+-----			
15	80	24.69	24.69
25	136	41.98	66.67
35	80	24.69	91.36
45	28	8.64	100.00
-----+-----			
Total	324	100.00	

```
. logistic peth_50 i.agecat
```

Logistic regression	Number of obs	=	287
	LR chi2(3)	=	15.20
	Prob > chi2	=	0.0017
Log likelihood = -190.40933	Pseudo R2	=	0.0384

peth_50	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
agecat						
25	1.934524	.6192737	2.06	0.039	1.032976	3.622914
35	3.199405	1.118076	3.33	0.001	1.612891	6.346484
45	4.232804	2.07895	2.94	0.003	1.616438	11.08402
_cons	.42	.1092154	-3.34	0.001	.2522933	.6991863
-----+-----						

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

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

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

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

- Therefore

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

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

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

so $OR = e^{\beta_2}$ adjusted for x_1

```
. logistic peth_50 age_10 i.month
```

Logistic regression

Number of obs = 287

LR chi2(2) = 11.05

Prob > chi2 = 0.0040

Log likelihood = -192.48576

Pseudo R2 = 0.0279

peth_50	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
age_10	1.57244	.2211985	3.22	0.001	1.193531	2.071641
88.month	1.039536	.2612106	0.15	0.877	.6352617	1.701085
_cons	.2017852	.0961936	-3.36	0.001	.0792703	.5136509

1.04 is the odds ratio for being in the minimally assessed vs. standard study arm when you hold age constant

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

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

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 • <code>ttest var1=hypoth val.*</code>	
Numerical; Two means, <u>paired data</u>	$H_a: \mu_1 \neq \mu_2$ (two-sided) $H_a: \mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	Paired t-test • <code>ttest var1=var2*</code>	Sign test • <code>signtest var1=var2</code> • Wilcoxon Signed-Rank <code>signrank var1=var2</code>)
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) • <code>ttest var1, by(byvar)</code> <code>unequal</code>	Wilcoxon rank-sum test • <code>ranksum var1, by(byvar)</code>
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 • <code>oneway var1 byvar</code>	Kruskal Wallis test • <code>kwallis var1, by(byvar)</code>
Dichotomous; One proportion	$H_a: p \neq p_a$ (two-sided) $H_a: p > p_a$ or $p < p_a$ (one-sided)	Proportion test • <code>prtest var1=hypoth value*</code> • <code>bitest var1=hypoth value</code>	
Dichotomous; two proportions	$H_a: p_1 \neq p_2$ (two-sided) $H_a: p_1 > p_2$ (one-sided)	Proportion test (z-test) • <code>prtest var1, by(byvar)</code>	Chi-square test • <code>tab var1 var2, chi exact</code> McNemar's for paired data: • <code>mcc var1 var2</code>
Categorical by categorical (nxk)	H_a : The rows not independent of the columns		Chi-square test • <code>tab var1 var2, chi exact</code>

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. 4 – LAB will occur
- Final exam
 - Take home exam will be posted online on Thursday Dec. 4
 - The exam is worth 30% of your grade.
 - NO COLLABORATIONS WITH YOUR CLASSMATES OR OTHERS
 - Exam due by upload to course web site Thursday, December 11, 10:30 a.m.
- Tuesday Dec. 9 – office hours by appointment

Thank you and good luck!