



Biostatistics

Collaboration Center

# Basic Biostatistics in Medical Research

## Lecture 3: Introduction to Linear Regression and Logistic Regression

*Jungwha “Julia” Lee, PhD*  
*Biostatistics Collaboration Center (BCC)*  
*Department of Preventive Medicine*  
*NU Feinberg School of Medicine*

# Outline

- **Simple linear regression**
- **Multiple linear regression**
  - estimating parameter
  - testing hypotheses about parameters
  - confidence and prediction interval
- **Logistic regression**
  - odds ratio vs. relative risk
  - univariate, multiple logistic regression
  - interpretation of coefficient
  - extensions of logistic regression

# Data Analysis Overview

- All variables you encounter in data analysis can be classified according as  
[independent | dependent] and  
[categorical/qualitative | continuous/quantitative].
- A basic data analysis strategy consist of
  1. Explore data
  2. Fit model
  3. Asses model; return to 2 if necessary
  4. Inference and interpretation

# Data Analysis Overview

- The classification of the data dictates the tool you can use:

		Dependent Variable	
		Categorical	Continuous
Independent Variable	Categorical	Chi-squared test	ANOVA
	Continuous	Logistic Regression	Linear Regression

# Simple Linear Regression

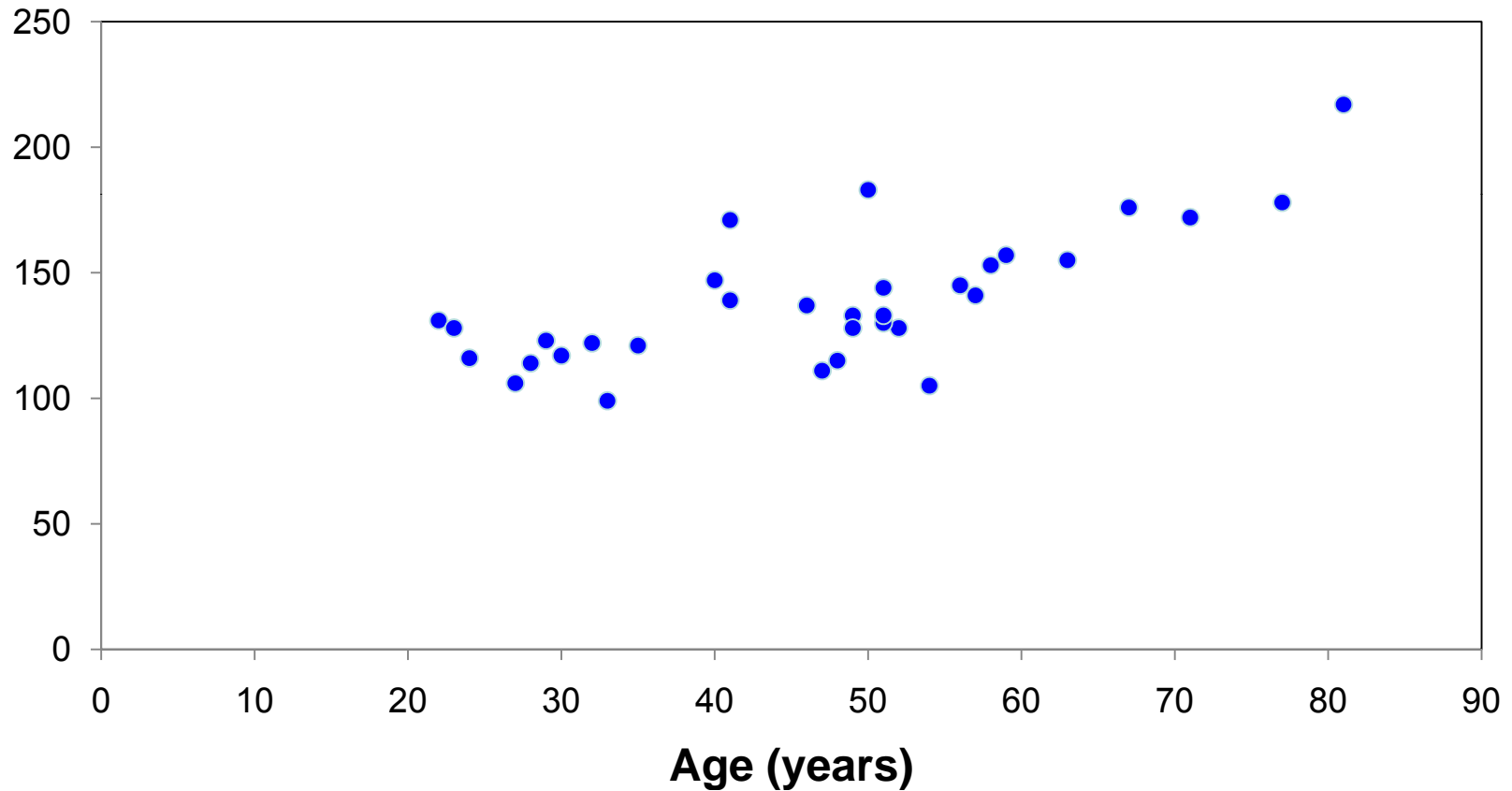
Table 1. Age and systolic blood pressure (SBP) among 33 adult women

<b>Age</b>	<b>SBP</b>	<b>Age</b>	<b>SBP</b>	<b>Age</b>	<b>SBP</b>
<b>22</b>	<b>131</b>	<b>41</b>	<b>139</b>	<b>52</b>	<b>128</b>
<b>23</b>	<b>128</b>	<b>41</b>	<b>171</b>	<b>54</b>	<b>105</b>
<b>24</b>	<b>116</b>	<b>46</b>	<b>137</b>	<b>56</b>	<b>145</b>
<b>27</b>	<b>106</b>	<b>47</b>	<b>111</b>	<b>57</b>	<b>141</b>
<b>28</b>	<b>114</b>	<b>48</b>	<b>115</b>	<b>58</b>	<b>153</b>
<b>29</b>	<b>123</b>	<b>49</b>	<b>133</b>	<b>59</b>	<b>157</b>
<b>30</b>	<b>117</b>	<b>49</b>	<b>128</b>	<b>63</b>	<b>155</b>
<b>32</b>	<b>122</b>	<b>50</b>	<b>183</b>	<b>67</b>	<b>176</b>
<b>33</b>	<b>99</b>	<b>51</b>	<b>130</b>	<b>71</b>	<b>172</b>
<b>35</b>	<b>121</b>	<b>51</b>	<b>133</b>	<b>77</b>	<b>178</b>
<b>40</b>	<b>147</b>	<b>51</b>	<b>144</b>	<b>81</b>	<b>217</b>

Source: Colton. Statistics in Medicine (1974)

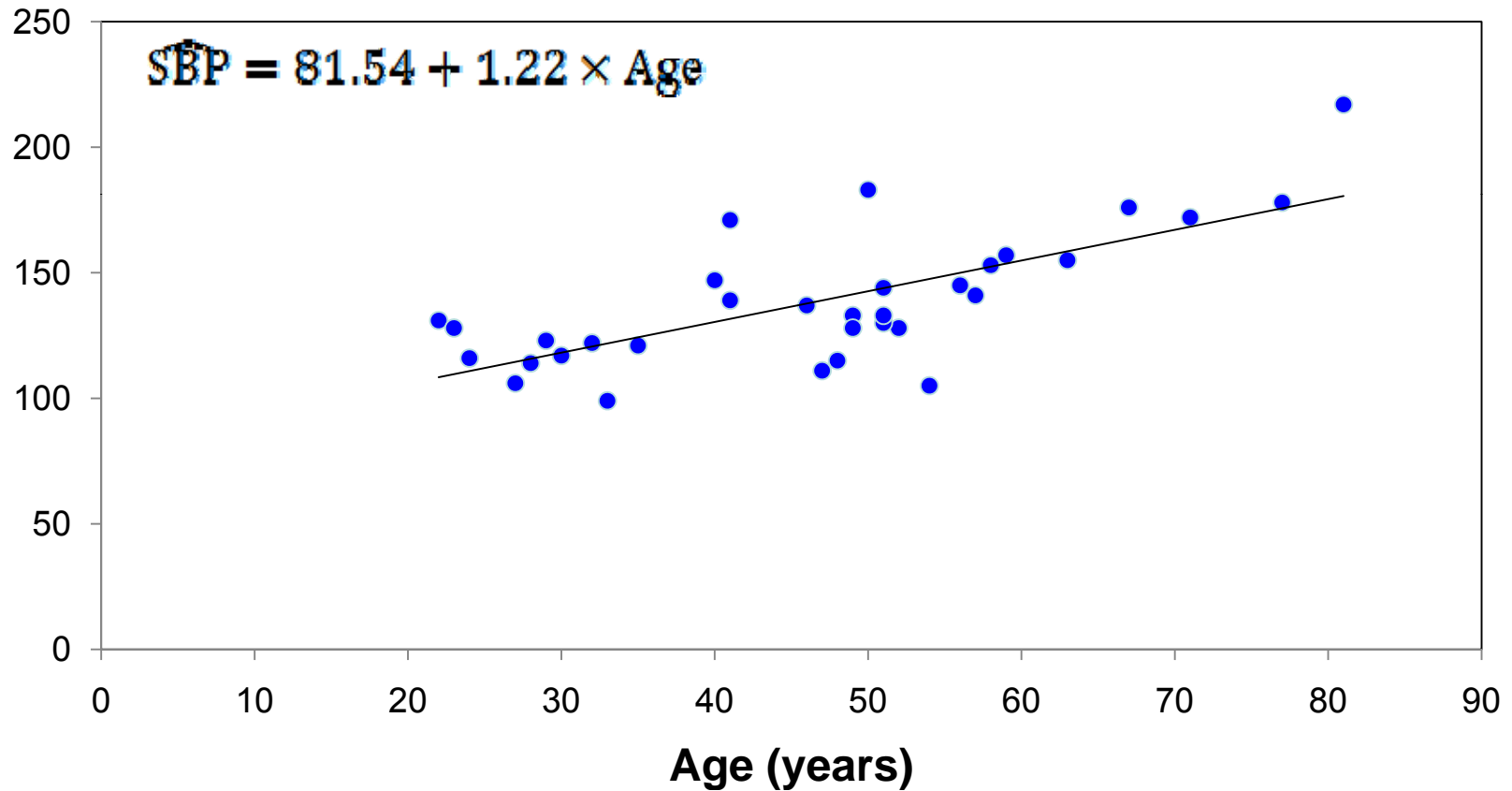
# Simple Linear Regression

SBP (mmHg)



# Simple Linear Regression

SBP (mmHg)



# Simple Linear Regression Model

- Regression Model

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, \dots, n$$

- $Y_i$  the value of the response variable in the  $i$  th trial
- $\alpha, \beta$  parameters
- $x_i$  the value of the predictor variable (a known constant)
- $\varepsilon_i$  a random error term
  - Mean  $E(\varepsilon_i)=0$
  - Variance  $\text{Var}(\varepsilon_i)=\sigma^2$
  - Covariance  $\text{Cov}(\varepsilon_i, \varepsilon_j)=0$  for all  $i, j$  ( $i \neq j$ )

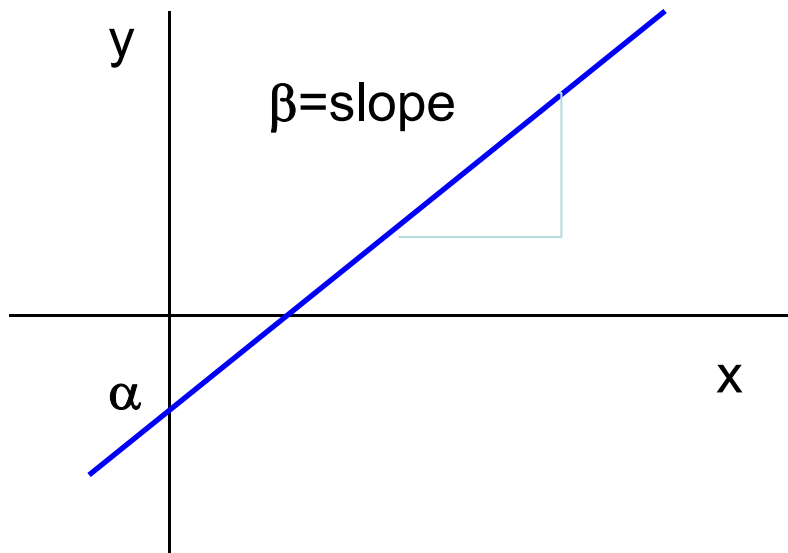
- Regression Function

$$E(Y_i) = \alpha + \beta x_i$$

# Simple Linear Regression

- Relation between 2 continuous variables (SBP and age)

$$y = \alpha + \beta x$$



- Hypothesis:
  - $H_0: \beta=0$
  - $H_1: \beta \neq 0$
- Regression coefficient  $\beta$ 
  - Measures association between y and x
  - Amount by which changes on average when x changes by one unit
  - Least squares method

# Least Squares Method

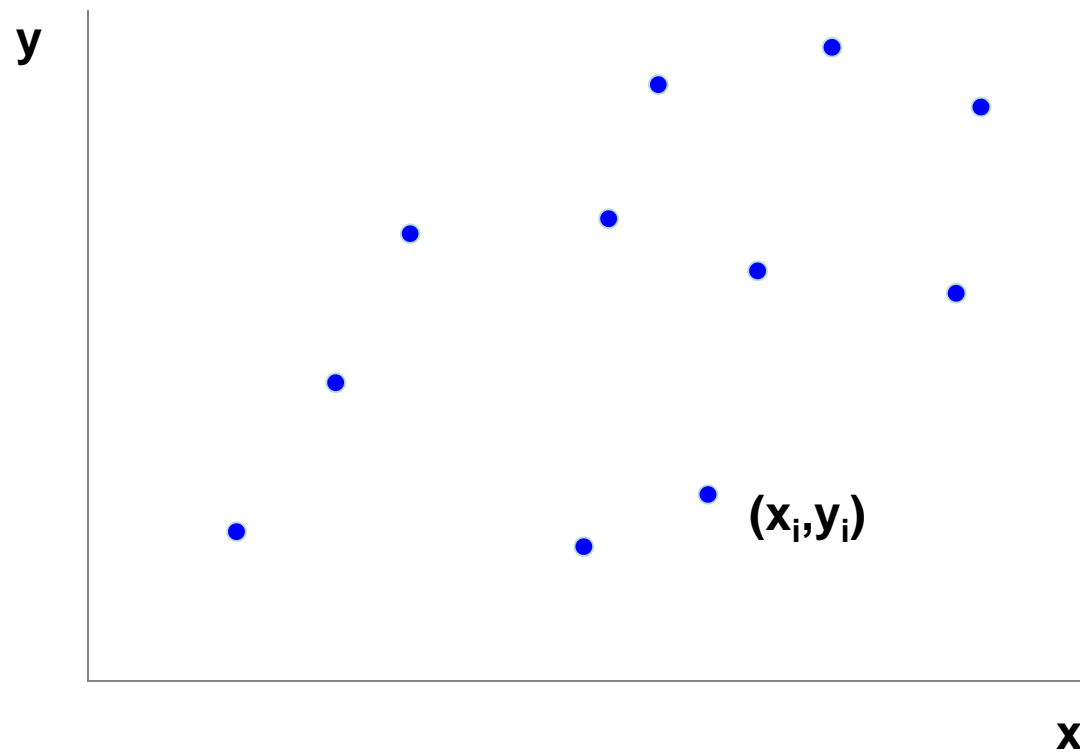
- To find “good” estimators of regression parameters  $\alpha$  and  $\beta$
- Random error

$$e_1 = y_1 - \hat{y}_1 = y_1 - (\alpha + \beta x_1)$$

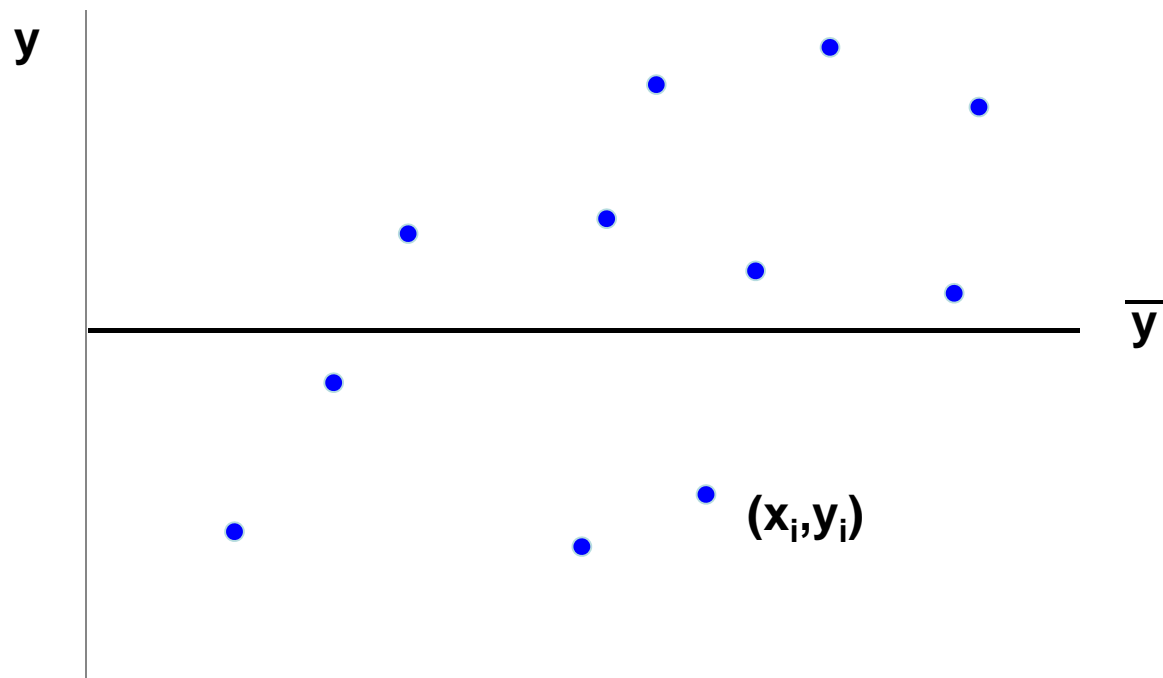
- $\hat{\alpha}$  and  $\hat{\beta}$  minimize the criterion Q for the given sample observations  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$

$$Q = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

# Partitioning Sum of Squares

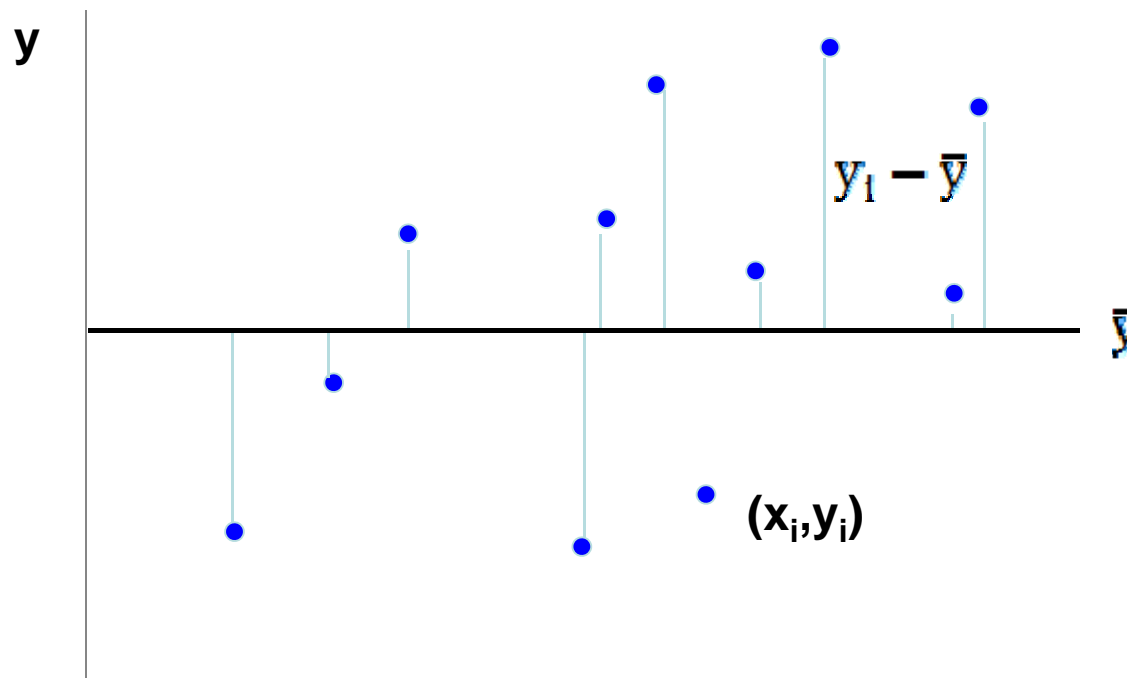


$$SS_{\text{Total}} = SS_{\text{Reg}} + SS_{\text{Error}}$$



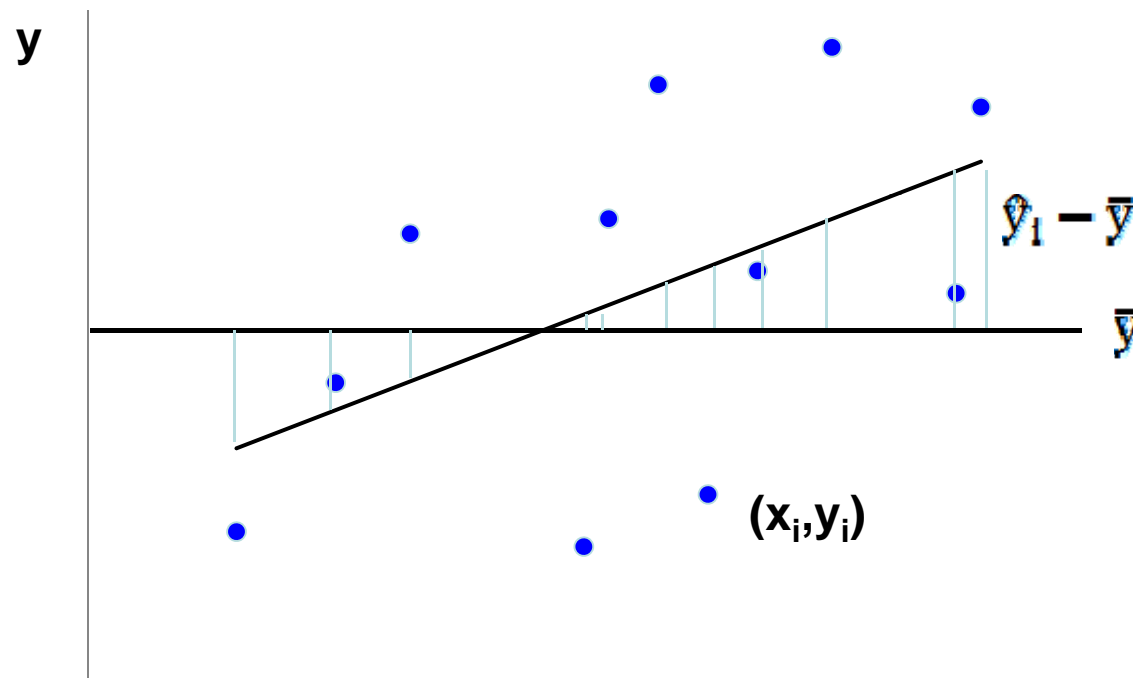
$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$SS_{\text{Total}} = SS_{\text{Reg}} + SS_{\text{Error}}$$



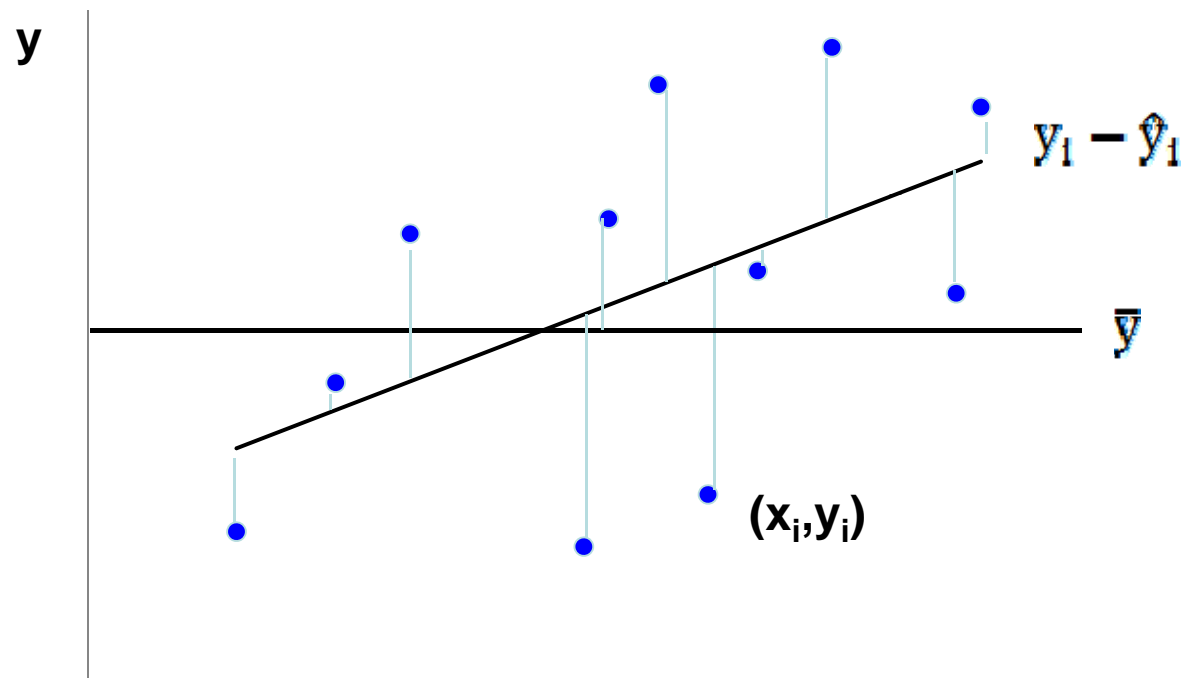
$$SS_{\text{Total}} = \sum_1 (y_i - \bar{y})^2$$

$$SS_{\text{Total}} = SS_{\text{Reg}} + SS_{\text{Error}}$$



$$SS_{\text{Reg}} = \sum_1 (\hat{y}_i - \bar{y})^2$$

$$SS_{\text{Total}} = SS_{\text{Reg}} + SS_{\text{Error}}$$



$$SS_{\text{Error}} = \sum_1 (y_i - \hat{y}_i)^2$$

# SAS Code and Output

```
proc reg; model SBP=age;run;
```

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: SBP

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = 81.5 + 1.2x$$

Number of Observations Read 33  
 Number of Observations Used 33

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	<b>1</b>	<b>SS<sub>Reg</sub> = 11450</b>	<b>MS<sub>Reg</sub> = 11450</b>	<b>32.96</b>	<b>&lt;.0001</b>
Error	31	SS <sub>Error</sub> = 10770	MS <sub>Err</sub> 347.41001		
Corrected Total	32	SS <sub>Total</sub> = 22220			

$F = \frac{MS_{Reg}}{MS_{Err}} > F_{1,31}$

Root MSE	18.63894	R-Square	0.5153
Dependent Mean	138.63636	Adj R-Sq	0.4997
Coeff Var	13.44448		

## Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	81.51675	10.46527	7.79	<.0001
<b>Age</b>	<b>1</b>	<b><math>\hat{\beta} = 1.22240</math></b>	<b>0.21293</b>	<b>5.74</b>	<b>&lt;.0001</b>

$$H_0: \beta = 0$$

$$t = \frac{\hat{\beta}}{se(\hat{\beta})} > t_{df=1}$$

# Interpretation of Output

- Hypothesis:
  - $H_0: \beta=0$  (slope)
  - $H_1: \beta \neq 0$
- Estimate of  $\beta$ 
  - $\hat{\beta} = 1.22$
- Test Statistic
  - $t=5.74$
- P-value  $<0.0001$
- Conclude:  $\beta \neq 0$ 

There is a statistically significant linear trend between age and SBP.
- Interpretation
  - one unit increase in age results in 1.22 increase in SBP on average
- 95% CI of  $\beta$ 
  - =  $\hat{\beta} \pm 1.96 \times se(\hat{\beta})$
  - =  $1.22 + 1.96 \times 0.12$
  - =  $(0.79, 1.66)$

# Multiple Linear Regression

- Relation between a continuous variable and a set of  $p$  continuous variables

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

- Partial regression coefficients  $\beta_p$ 
  - Amount by which  $y$  changes on average when  $x_p$  changes by one unit and all the other  $x_p$ 's remain constant
  - Measures association between  $x_p$  and  $y$  adjusted for all other  $x_p$
- Example
  - SBP vs. age, weight, height, etc

# Multiple Linear Regression

<b>y</b>	<b>=</b>	<b><math>\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p</math></b>
Predicted		Predictor variables
Response variable		Explanatory variables
Outcome variable		Covariates
Dependent variable		Independent variables

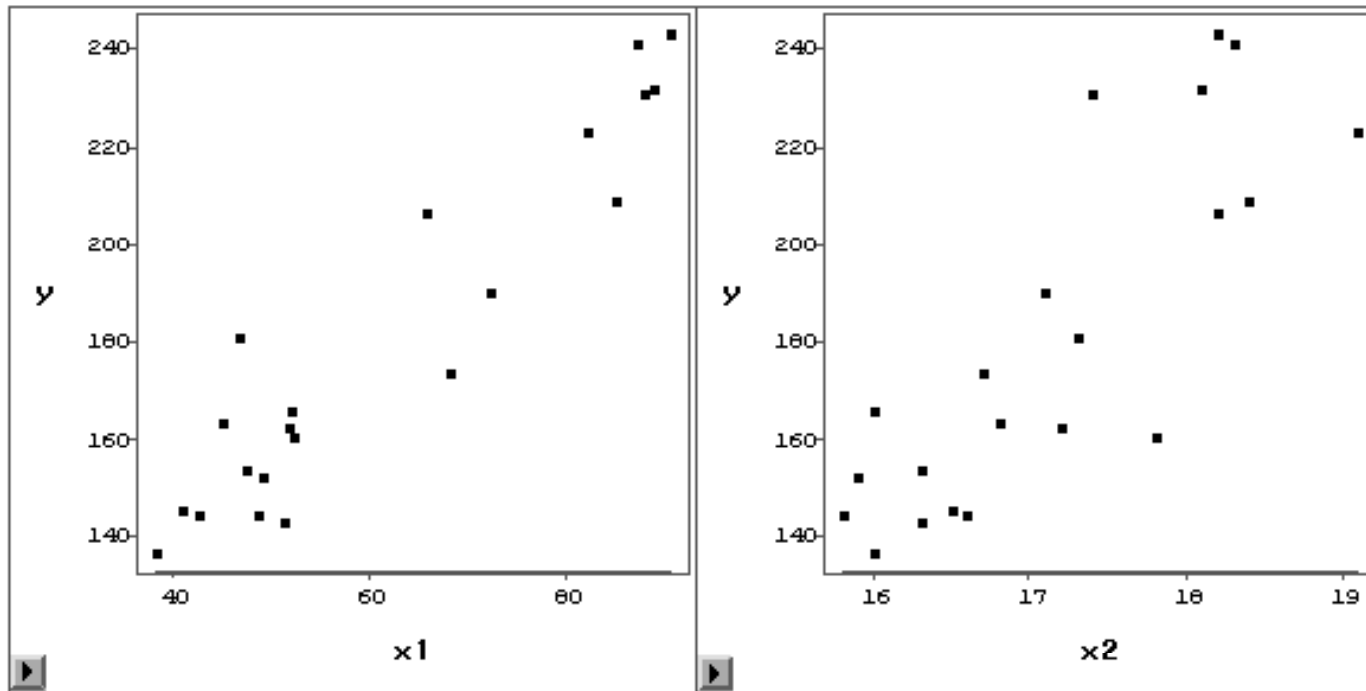
# Multiple Linear Regression

- Table 1a. Sales, Target Population, Income in 21 cities

	<b>x1</b>	<b>x2</b>	<b>y</b>
<code>data ch6fig05;</code>	38.4	16.0	137.2
<code>input x1 x2 y;</code>	87.9	18.3	241.9
<code>label x1='targetpop'</code>	72.8	17.1	191.1
<code>      x2='dispoinc';</code>	88.4	17.4	232.0
<code>cards;</code>	42.9	15.8	145.3
68.5	16.7	174.4	52.5
45.2	16.8	164.4	85.7
91.3	18.2	244.2	41.3
47.8	16.3	154.6	51.7
46.9	17.3	181.6	89.6
66.1	18.2	207.5	82.7
49.5	15.9	152.8	52.3
52.0	17.2	163.2	;
48.9	16.6	145.4	<code>run;</code>

# Multiple Linear Regression

Sales



Target Population

Income

# Correlation

```
proc corr data = ch6fig05a;run;
```

The CORR Procedure

3 Variables: x1 x2 y

## Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
x1	21	62.01905	18.62033	1302	38.40000	91.30000	targetpop
x2	21	17.14286	0.97035	360.00000	15.80000	19.10000	dispoinc
y	21	181.90476	36.19130	3820	137.20000	244.20000	

Pearson Correlation Coefficients, N = 21  
Prob > |r| under H0: Rho=0

	x1	x2	y
x1	1.00000	0.78130	<b>0.94455</b>
targetpop		<.0001	<.0001
x2	0.78130	1.00000	<b>0.83580</b>
dispoinc	<.0001		<.0001
y	0.94455	0.83580	1.00000
	<.0001	<.0001	

# SAS Code & Output

```
proc reg data = ch6fig05a; model y = x1 x2; run;
```

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: y

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -68.9 + 1.45x_1 + 9.37x_2$$

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square
Model	2	24015	12008
Error	18	2180.92741	121.16263
Corrected Total	20	26196	

Root MSE	11.00739	R-Square	0.9167 = $SS_R/SS_T$
Dependent Mean	181.90476	Adj R-Sq	0.9075
Coeff Var	6.05118		

$$H_0: \beta_1 = \beta_2 = 0$$

F Value 99.10 Pr > F <.0001

$$F = \frac{MS_{Reg}}{MS_{Err}} > F_{2,18}$$

## Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	-68.85707	60.01695	-1.15	0.2663
x1	targtpop	1	$\hat{\beta}_1 = 1.45456$	0.21178	6.87	<.0001
x2	dispoinc	1	$\hat{\beta}_2 = 9.36550$	4.06396	2.30	0.0333

$$H_0: \beta_1 = 0 \quad t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

$$H_0: \beta_2 = 0 \quad t = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)}$$

# Diagnosics

- Residuals are key to assessing the validity of the linear model assumptions.
- Outliers are unusually large observations, due to an unmodeled shift or an (unmodeled) increase in variance.
- Leverage is the potential for an observation to affect the fit of the model.
- Influence is the actual impact that an observation has on the fit of the model.

# Diagnosics (cont')

- Multicollinearity: predictor variables are correlated among themselves
  - Variance Inflation Factors (VIF)  $> 10$
  - Tolerance =  $1/\text{VIF}$

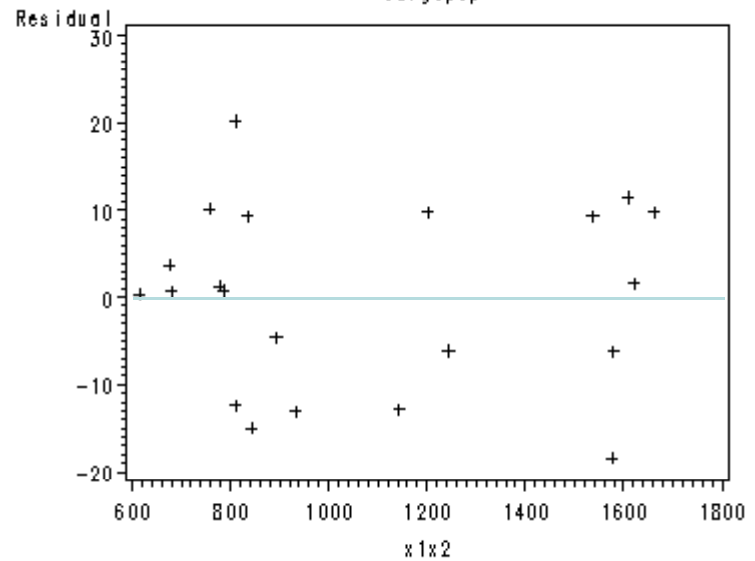
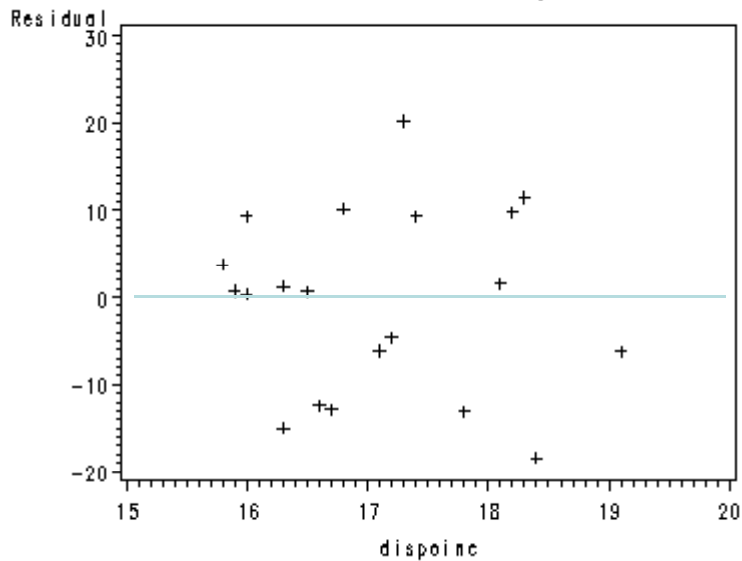
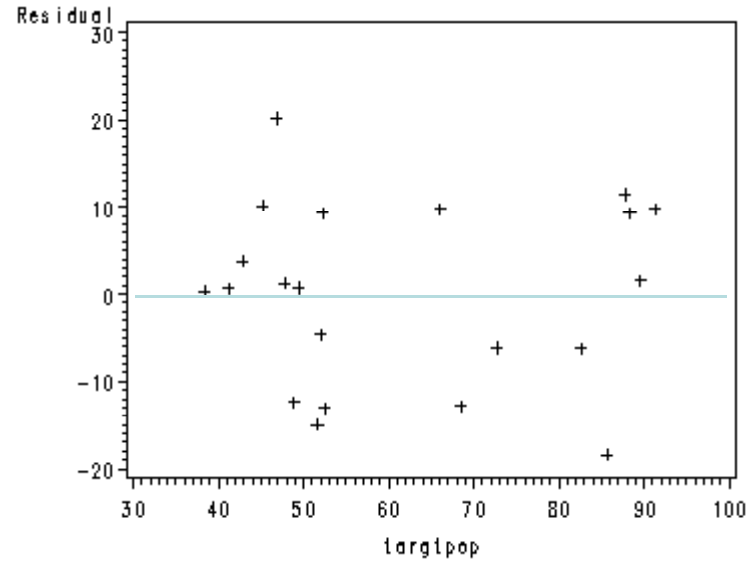
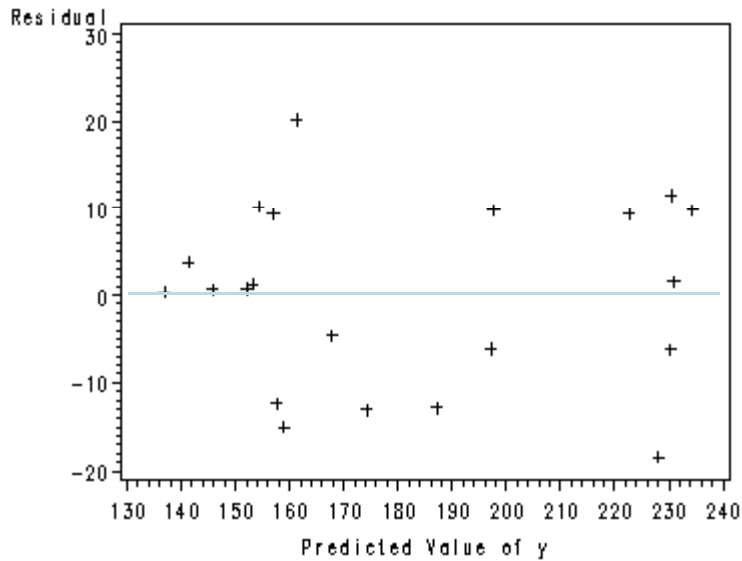
# SAS Code & Output

```
proc print data = outfig05;  
  var y x1 x2 fitted residual;  
run;
```

Obs	y	x1	x2	fitted	residual
1	174.4	68.5	16.7	187.184	-12.7841
2	164.4	45.2	16.8	154.229	10.1706
3	244.2	91.3	18.2	234.396	9.8037
4	154.6	47.8	16.3	153.329	1.2715
5	181.6	46.9	17.3	161.385	20.2151
6	207.5	66.1	18.2	197.741	9.7586
7	152.8	49.5	15.9	152.055	0.7449
8	163.2	52.0	17.2	167.867	-4.6666
9	145.4	48.9	16.6	157.738	-12.3382
10	137.2	38.4	16.0	136.846	0.3540
11	241.9	87.9	18.3	230.387	11.5126
12	191.1	72.8	17.1	197.185	-6.0849
13	232.0	88.4	17.4	222.686	9.3143
14	145.3	42.9	15.8	141.518	3.7816
15	161.1	52.5	17.8	174.213	-13.1132
16	209.7	85.7	18.4	228.124	-18.4239
17	146.4	41.3	16.5	145.747	0.6530
18	144.0	51.7	16.3	159.001	-15.0013
19	232.6	89.6	18.1	230.987	1.6130
20	224.1	82.7	19.1	230.316	-6.2161
21	166.5	52.3	16.0	157.064	9.4356

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$
$$= -68.9 + 1.45x_{i1} + 9.37x_{i2}$$
$$e_i = y_i - \hat{y}_i$$

# Residuals



# Residuals

## Output Statistics

Obs	Std Error Residual	Student Residual	-2 -1 0 1 2	Cook's D
1	10.316	-1.239	**	0.071
2	10.417	0.976	*	0.037
3	10.006	0.980	*	0.067
4	10.522	0.121		0.000
5	10.077	2.006	****	0.259
6	10.099	0.966	*	0.059
7	10.187	0.0731		0.000
8	10.491	-0.445		0.007
9	10.601	-1.164	**	0.035
10	10.252	0.0345		0.000
11	10.174	1.132	**	0.073
12	10.466	-0.581	*	0.012
13	9.603	0.970	*	0.098
14	10.186	0.371		0.008
15	9.787	-1.340	**	0.159
16	10.207	-1.805	***	0.177
17	10.355	0.0631		0.000
18	10.516	-1.427	**	0.065
19	10.082	0.160		0.002
20	9.348	-0.665	*	0.057
21	10.224	0.923	*	0.045

Sum of Residuals 0  
 Sum of Squared Residuals 2180.92741

# Confidence & Prediction Interval

Output Statistics

Obs	Dep Var y	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	174.4000	187.1841	3.8409	179.1146	195.2536	162.6910	211.6772	-12.7841
2	164.4000	154.2294	3.5558	146.7591	161.6998	129.9271	178.5317	10.1706
3	244.2000	234.3963	4.5882	224.7569	244.0358	209.3421	259.4506	9.8037
4	154.6000	153.3285	3.2331	146.5361	160.1210	129.2260	177.4311	1.2715
5	181.6000	161.3849	4.4300	152.0778	170.6921	136.4566	186.3132	20.2151
6	207.5000	197.7414	4.3786	188.5424	206.9404	172.8533	222.6295	9.7586
7	152.8000	152.0551	4.1696	143.2952	160.8150	127.3259	176.7843	0.7449
8	163.2000	167.8666	3.3310	160.8684	174.8649	143.7053	192.0280	-4.6666
9	145.4000	157.7382	2.9628	151.5136	163.9628	133.7895	181.6869	-12.3382
10	137.2000	136.8460	4.0074	128.4268	145.2653	112.2354	161.4566	0.3540
11	241.9000	230.3874	4.2012	221.5610	239.2137	205.6346	255.1402	11.5126
12	191.1000	197.1849	3.4109	190.0188	204.3510	172.9744	221.3954	-6.0849
13	232.0000	222.6857	5.3808	211.3810	233.9904	196.9448	248.4266	9.3143
14	145.3000	141.5184	4.1735	132.7502	150.2866	116.7863	166.2506	3.7816
15	161.1000	174.2132	5.0377	163.6294	184.7971	148.7807	199.6458	-13.1132
16	209.7000	228.1239	4.1214	219.4652	236.7826	203.4304	252.8174	-18.4239
17	146.4000	145.7470	3.7331	137.9041	153.5899	121.3276	170.1664	0.6530
18	144.0000	159.0013	3.2529	152.1672	165.8354	134.8870	183.1157	-15.0013
19	232.6000	230.9870	4.4176	221.7059	240.2681	206.0684	255.9056	1.6130
20	224.1000	230.3161	5.8120	218.1054	242.5267	204.1647	256.4675	-6.2161
21	166.5000	157.0644	4.0792	148.4944	165.6344	132.4018	181.7270	9.4356
22	.	191.1039	2.7668	185.2911	196.9168	167.2589	214.9490	.
23	.	174.1494	4.5986	164.4881	183.8107	149.0867	199.2121	.

Confidence Interval for mean response at point  $x_0$

Prediction Interval for new observation at point  $x_0$

# Confidence & Prediction Interval

- Confidence intervals all take the same form  
Estimator  $\pm$  StdErr(Estimator)
- Specific results for  $(1-\alpha)100\%$  intervals

Confidence Interval for slope  $\hat{\beta}$

$$\hat{\beta} \pm t_{\alpha/2, df} \hat{\sigma} \sqrt{1/S_{xx}}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

Confidence Interval for mean response at point  $x_0$

$$\hat{y}(x_0) \pm t_{\alpha/2, df} \hat{\sigma} \sqrt{1/n + (x_0 - \bar{x})^2/S_{xx}}$$

Prediction Interval for new observation at point  $x_0$

$$\hat{y}(x_0) \pm t_{\alpha/2, df} \hat{\sigma} \sqrt{1 + 1/n + (x_0 - \bar{x})^2/S_{xx}}$$

# Extensions of Linear Regression

- ANOVA=Analysis Of Variance
  - One-way ANOVA
  - Two-way ANOVA
  - .....
- ANCOVA=Analysis Of Covariance
- Random and Mixed Effects Models

# Logistic Regression

Table 2. Age and signs of coronary heart disease (CHD)

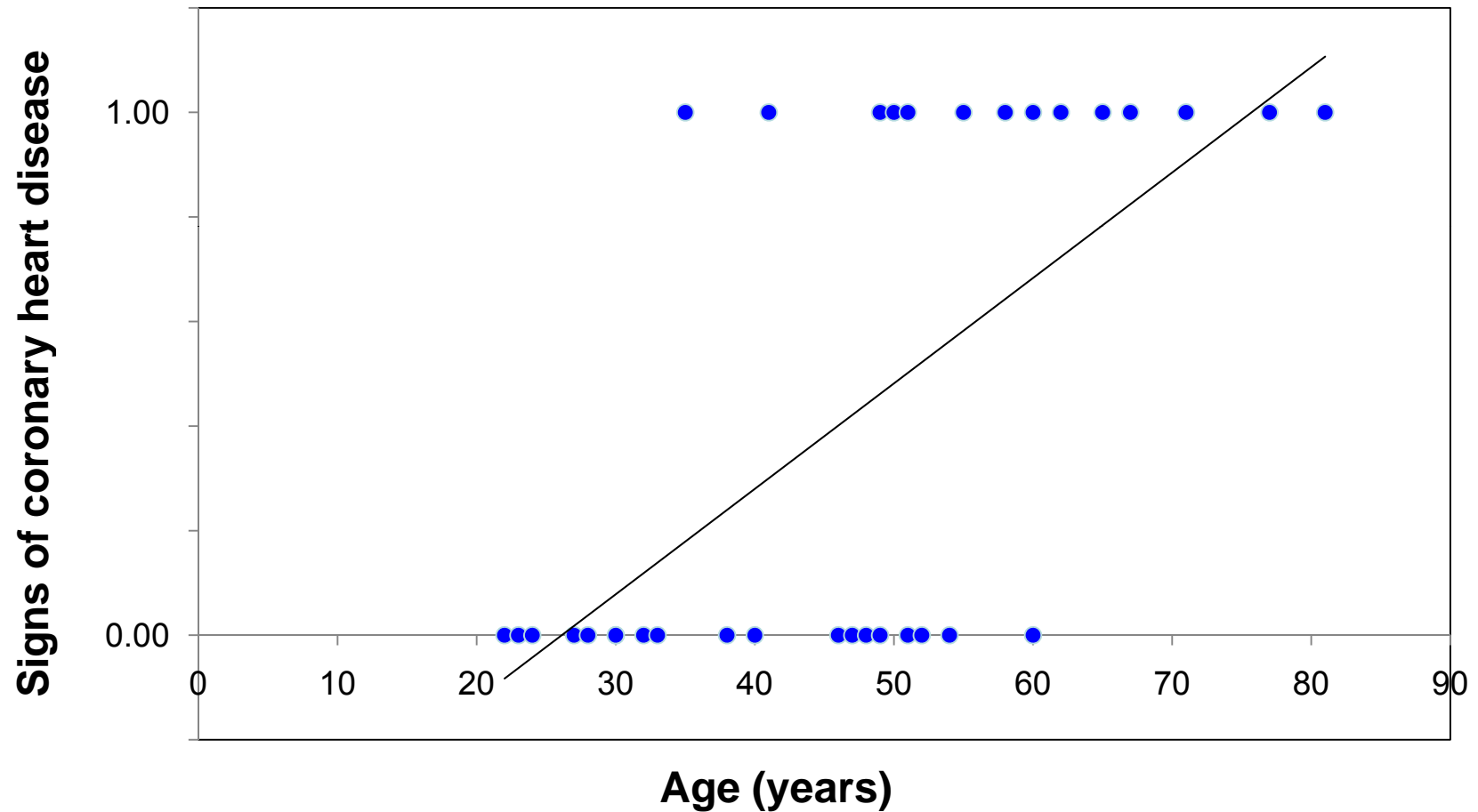
Age	CHD	Age	CHD	Age	CHD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

Source: Colton. Statistics in Medicine (1974)

# How to analyze these data?

- Compare mean age of diseased and non-diseased.
  - Diseased: 57.7 years
  - Non-diseased: 38.6 years
  - Hypothesis:  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$
  - t-test:  $p < 0.0001$
- Linear regression?

# Scatter plot: Data from Table 2



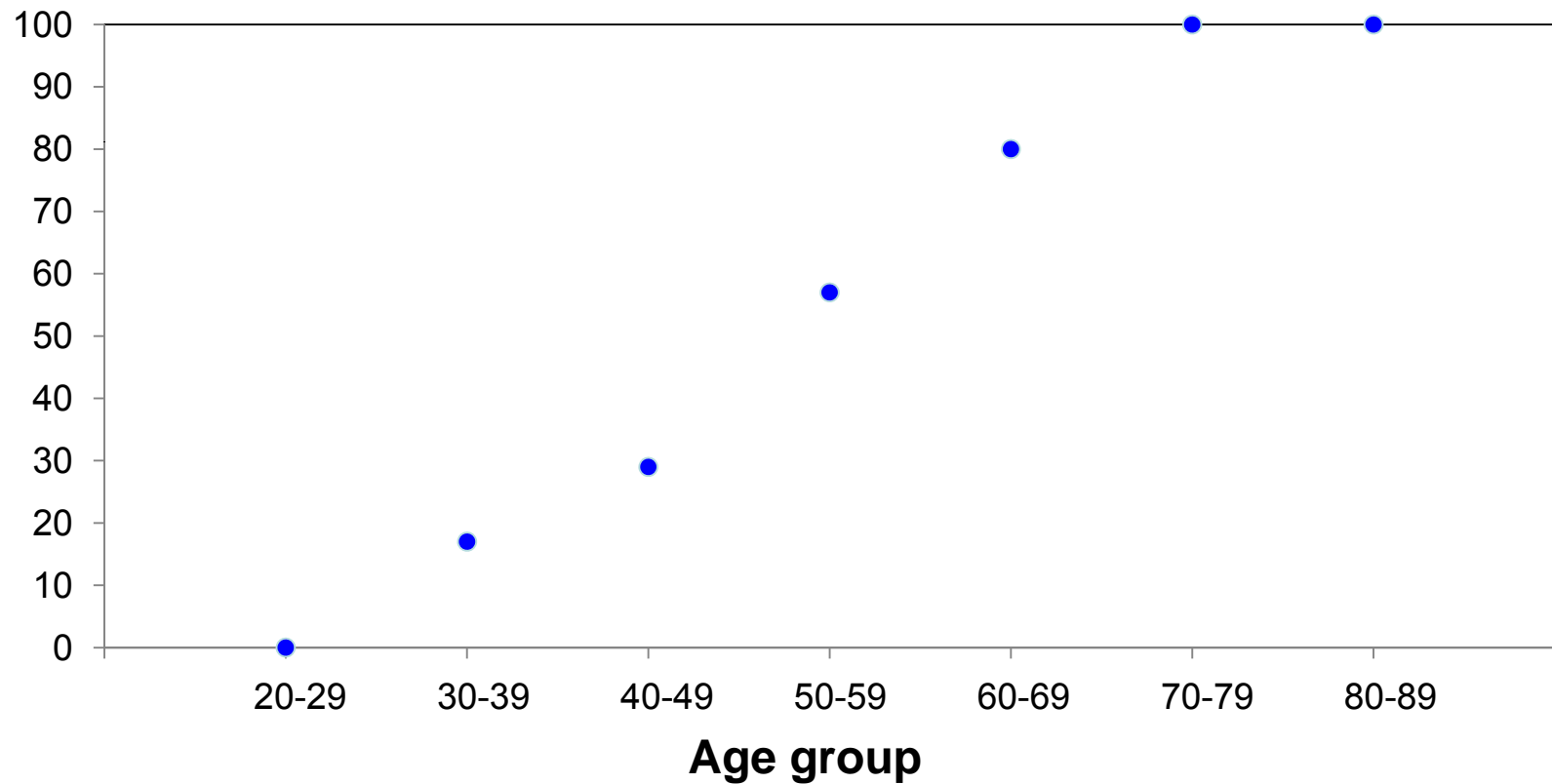
# Logistic regression

Table 3. Prevalence (%) of signs of CHD according to age group

Age group	No. in group	Diseased	
		No.	%
20-29	5	0	0
30-39	6	1	17
40-49	7	2	29
50-59	7	4	57
60-69	5	4	80
70-79	2	2	100
80-89	1	1	100

# Scatter plot: Data from Table 3

**Diseased %**

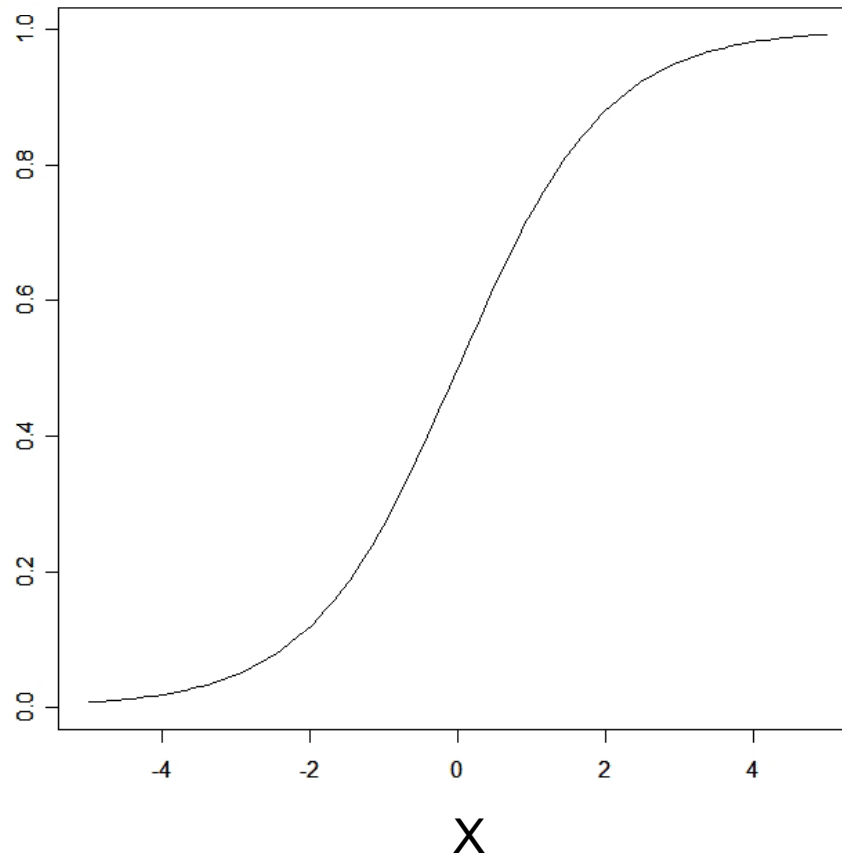


# Logistic Function

Probability of disease

$$\Pr(Y = 1 | X) = f(X)$$

$$= \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}}$$



# Objective of the Analysis

- How is a risk factor related to disease or death?
- To model the relationship between the risk of the outcome and the predictors of interest, where
  - The outcome is binary (Yes vs No):  $Y=1$  or  $0$
  - The predictors are  $X=(X_1, X_2, \dots, X_p)$
  - The risk of the outcome is  $\Pr(Y=1)$  or  $E(Y)$
- The statistical model is

$$\Pr(Y=1|X) = E(Y|X) = f(X)$$

# Construction of $f(x)$

- Linear regression:  $E(Y)=f(X)=\alpha+\beta X$
- Can we use the same model? e.g.,

$$\Pr(\text{CHD})= \alpha+\beta\cdot\text{BMI}$$

- $\Pr(\text{CHD})$  is a probability and thus will always be between 0 and 1
- Need to select  $f(X)$  so that it is always between 0 and 1

# What is logistic regression?

$$\begin{aligned}\Pr(Y = 1 | X) &= f(X) \\ &= \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}}\end{aligned}$$

# Review : Odds and Odds Ratio

## Race

		White	Non-White
Y	Disease	$a$	$b$
	No Disease	$c$	$d$

Risk  $\frac{a}{a+c}$   $\frac{b}{b+d}$

Odds  $\frac{a}{c}$   $\frac{b}{d}$

$$\text{Relative Risk (RR)} = \frac{a}{a+c} \frac{b+d}{b}$$

$$\begin{aligned} \text{Odds Ratio (OR)} &= \text{Odds for White} / \text{Odds for Non-White} \\ &= \frac{a/c}{b/d} \end{aligned}$$

# More on OR

- Ranges from 0 to infinity
- Tends to be skewed (i.e., not symmetric):
  - $OR > 1$ : Increased risk when exposed
  - $OR = 1$ : no difference in risk btw the two groups
  - $OR < 1$ : Lower risk (protective) when exposed
- Log of OR tends to be symmetric and normally distributed.
  - $\text{Log}(OR) > 0$ : increased risk
  - $\text{Log}(OR) = 0$ : no difference in risk
  - $\text{Log}(OR) < 0$ : decreased risk

# Logistic Regression

$$\begin{aligned}\text{logit } \Pr(Y = 1 | \text{Race, Age}) &= \log \frac{\Pr(Y = 1 | \text{Race, Age})}{1 - \Pr(Y = 1 | \text{Race, Age})} \\ &= \log \text{odds} = \beta_0 + \beta_1 \text{Race} + \beta_2 \text{Age}\end{aligned}$$

$$\begin{aligned}\log (\text{Odds white}) &= \beta_0 + \beta_1 + \beta_2 \text{Age} \\ \log (\text{Odds minority}) &= \beta_0 + \beta_2 \text{Age}\end{aligned}$$

$$\begin{aligned}\text{Odds white} &= \exp(\beta_0 + \beta_1 + \beta_2 \text{Age}) \\ \text{Odds minority} &= \exp(\beta_0 + \beta_2 \text{Age})\end{aligned}$$

$$\begin{aligned}\text{Odds Ratio (white vs minority)} &= \frac{\exp(\beta_0 + \beta_1 + \beta_2 \text{Age})}{\exp(\beta_0 + \beta_2 \text{Age})} \\ &= \exp(\beta_1)\end{aligned}$$

So,  $\beta_1 = \log (\text{Odds Ratio})$

# Interpretation of Coefficient

- Let  $X=1$  if exposed and  $X=0$  if not exposed
- Let  $\beta$  denote the regression coefficient for  $X$ , then  $\beta$  is the log (OR) for the outcome (e.g., heart attack), comparing the exposed to the not exposed (e.g., coffee drinking).

$$e^{\beta} = \frac{ad}{bc} = \frac{\Pr(Y = 1 | X = 1) \Pr(Y = 0 | X = 0)}{\Pr(Y = 0 | X = 1) \Pr(Y = 1 | X = 0)}$$

# Interpretation of Coefficient

- For **continuous** exposure measurements,  $\beta$  is the log of the OR for the outcome (e.g., heart attack), corresponding to one unit increase from the previous level (e.g., BMI):

$$e^{\beta} = \frac{\Pr(Y = 1 | X = x + 1) \Pr(Y = 0 | X = x)}{\Pr(Y = 0 | X = x + 1) \Pr(Y = 1 | X = x)}$$

- Which previous level? Any level!

# An Example

- Applied Logistic Regression (Hosmer & Lemeshow)
- Outcome: CHD recurrence, Yes vs No
- Predictor: Age
- The logistic regression model is

$$\Pr(\text{CHD} \mid \text{Age}) = \frac{e^{\alpha + \beta \times \text{Age}}}{1 + e^{\alpha + \beta \times \text{Age}}}$$

# An Example (cont.)

	<b>id</b>	<b>age</b>	<b>agrp</b>	<b>chd</b>
1.	1	20	1	0
2.	2	23	1	0
3.	3	24	1	0
4.	4	25	1	0
5.	5	25	1	1
6.	6	26	1	0
7.	7	26	1	0
8.	8	28	1	0
9.	9	28	1	0
10.	10	29	1	0
11.	11	30	2	0
12.	12	30	2	0
13.	13	30	2	0
14.	14	30	2	0
15.	15	30	2	0
...				

# SAS Code & Output

```
proc logistic descending; model chd = age; run;
```

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-5.3095	1.1337	21.9350	<.0001
<b>AGE</b>	<b>1</b>	<b>0.1109</b>	<b>0.0241</b>	<b>21.2541</b>	<b>&lt;.0001</b>

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
<b>AGE</b>	<b>1.117</b>	<b>1.066 1.171</b>

Association of Predicted Probabilities and Observed Responses

Percent Concordant	79.0	Somers' D	0.600
Percent Discordant	19.0	Gamma	0.612
Percent Tied Pairs	2.0	Tau-a	0.297
	2451	<b>c</b>	<b>0.800</b>

# An Example (cont.)

- The estimate of  $\beta$  :  $\hat{\beta} = 0.111$  (s.e.=0.024)
- The p-value for  $H_0 : \beta=0$  is  $<0.0001$
- The 95% conf. interval for  $\beta$  is (0.064, 0.158)
- OR corresponding to one year increase of age:

$$e^{\hat{\beta}} = e^{0.111} = 1.117$$

- with the 95% confidence interval of  $(e^{0.064}, e^{0.158}) = (1.066, 1.171)$ .

# SAS Output (cont.)

- The OR for the CHD recurrence corresponding to one year increase in age is 1.12 with a 95% CI of [1.07, 1.17].
- *c* statistics is 0.80 which measures the concordance between the predicted outcome and the observed outcome.
  - between 0.5 and 1.0

# Multiple Logistic Regression

- How to make adjustment for potential confounders?
- In linear regression:  $E(Y|X, Z) = \alpha + \beta X + \gamma Z$
- In logistic regression:

$$\Pr(Y | X, Z) = \frac{e^{\alpha + \beta X + \gamma Z}}{1 + e^{\alpha + \beta X + \gamma Z}}, \quad \text{or}$$

$$\text{logit}(p) = \log \left\{ \frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} \right\} = \alpha + \beta x + \gamma z$$

# Multiple Logistic Regression (cont.)

- The regression coefficient  $\beta$  can be interpreted as the log of the OR corresponding to one unit increase in  $X$ , while the level of  $Z$  is fixed at  $z$ .

$$e^{\beta} = \frac{\Pr(Y = 1 | X = x + 1, Z = z) \Pr(Y = 0 | X = x, Z = z)}{\Pr(Y = 0 | X = x + 1, Z = z) \Pr(Y = 1 | X = x, Z = z)}$$

# An Example: Heart Attack

- Dependent variable
  - whether the patient has had a second heart attack within 1 year  
(binary: 1= Yes vs. 0 = No).
- Two independent variables:
  - whether the patient completed a treatment consisting of anger control practice  
(binary: 1=Yes vs. 0=No).
  - score on a trait anxiety scale  
(a higher score means more anxious).

# Data

ID	Heart Attack	Treatment	Anxiety
1	1	1	70
2	1	1	80
3	0	1	50
4	1	0	60
5	0	0	40
6	0	0	65
7	0	0	75
8	1	0	80
9	1	0	70
10	1	0	60

# SAS Code

- Univariate logistic regression:

```
proc logistic descending;  
model HeartAttack = treatment;  
run;
```

- Multiple logistic regression:

```
proc logistic descending;  
model HeartAttack = treatment anxiety;  
run;
```

# SAS Output (Univariate)

The LOGISTIC Procedure

...

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.5596	0.6268	0.7974	0.372
<b>Treatment</b>	<b>1</b>	<b>-1.2528</b>	<b>0.9449</b>	<b>1.7583</b>	<b>0.185</b>

The LOGISTIC Procedure

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
<b>Treatment</b>	<b>0.2857</b>	<b>0.0448</b>	<b>1.8207</b>

# Univariate Analysis Result

- The parameter estimate is  $\hat{\beta} = -1.2528$  with a 95% confidence interval of  $(-3.1048, 0.5992)$ .
- The OR of the second heart attack comparing those received treatment to those did not is 0.2857 with a 95% CI of  $(0.0448, 1.8207)$ .

# SAS Output (Multiple)

The LOGISTIC Procedure

...

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-6.3647	3.2136	3.9204	0.0477
<b>Treatment</b>	<b>1</b>	<b>-1.0241</b>	<b>1.1701</b>	<b>0.7656</b>	<b>0.3818</b>
<b>Anxiety</b>	<b>1</b>	<b>0.1190</b>	<b>0.0550</b>	<b>4.6872</b>	<b>0.0304</b>

The LOGISTIC Procedure

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
<b>Treatment</b>	<b>0.3591</b>	<b>0.0362 3.5647</b>
<b>Anxiety</b>	<b>1.1264</b>	<b>1.0113 1.2546</b>

# Multiple Analysis Result

- The OR for the second heart attack comparing those received treatment to those did not, **while controlling for the anxiety level**, is 0.36 with a 95% CI of [0.04, 3.56] and a p-value of 0.38.
- Every unit increase in Anxiety score increases the odds of the second heart attack by 12.6% (95% CI: [1.1%, 25.5%],  $p=0.03$ ) **while treatment status remains the same.**

# Obtaining Risk, Risk Difference, and Relative Risk

- What if we want to know the risk difference or relative risk since odds ratio is too difficult to explain?
- You can still use logistic regression!

# Obtaining Risk, Risk Difference, and Relative Risk (cont.)

$$\text{Pr (Heart Attack)} = \frac{e^{(-6.36-1.02 \times \text{Treatment} + 0.12 \times \text{Anxiety})}}{1 + e^{(-6.36-1.02 \times \text{Treatment} + 0.12 \times \text{Anxiety})}}$$

# Obtaining Risk, Risk Difference, and Relative Risk (cont.)

- Anxiety level = 50:
  - $\Pr(\text{Heart attack} \mid \text{Treatment}, 50) = 0.19$
  - $\Pr(\text{Heart attack} \mid \text{No Treatment}, 50) = 0.40$
  - Risk Difference (RD) =  $0.19 - 0.40 = -0.21$
  - Relative Risk (RR) =  $0.19 / 0.40 = 48\%$
- Anxiety level = 80:
  - $\Pr(\text{Heart attack} \mid \text{Treatment}, 80) = 0.89$
  - $\Pr(\text{Heart attack} \mid \text{No Treatment}, 80) = 0.96$
  - RD =  $0.89 - 0.96 = -0.07$
  - RR =  $0.89 / 0.96 = 93\%$

# How about OR

- What is the OR for treatment vs. no treatment?
  - Anxiety level = 50
    - $\text{Pr}(\text{Heart attack}|\text{Treatment}, 50)=0.19$
    - $\text{Pr}(\text{Heart attack}|\text{No Treatment}, 50)=0.40$
    - $\text{OR} = (0.19/0.81) / (0.40/0.60) = 0.36$
  - Anxiety level = 80
    - $\text{Pr}(\text{Heart attack}|\text{Treatment}, 80)=0.89$
    - $\text{Pr}(\text{Heart attack}|\text{No Treatment}, 80)=0.96$
    - $\text{OR} = (0.89/0.11) / (0.96/0.04) = 0.36$
  - What a surprise: the same OR???

# Extensions of Logistic Regression

- Polytomous response
- Exact method
- Conditional logistic regression
- Longitudinal data

# Next Time

## ***Statistical Power*** ***Sample Size***

# Useful References

- *Intuitive Biostatistics*, Harvey Motulsky, Oxford University Press, 1995
  - Highly readable, minimally technical
- *Practical Statistics for Medical Research*, Douglas G. Altman, Chapman & Hall, 1991
  - Readable, not too technical
- *Fundamentals of Biostatistics*, Bernard Rosner, Duxbury, 2000
  - Useful reference, somewhat technical

# Contact BCC

- Please go to our web site at:  
<http://www.medschool.northwestern.edu/depts/bcc/>
- Fill out the online request form for further collaboration.