Statistical Thinking in Biomedical Research

Session #3

Statistical Modeling

Lily Wang, PhD

Department of Biostatistics

(modified from notes by J.Patrie, R.Abbott, U of Virginia and WD Dupont, Vanderbilt U)

Objectives

This presentation assumes no mathematical background, and focuses basic introduction to linear regression models, what they can do for you and how to interpret results.

For further reading, please refer to

- MPH and MSCI programs at Vanderbilt
- *Medical Statistics*, by BR Kirkwood *et al*.
- *Statistical Modeling for Biomedical Researchers*, by WD Dupont

Regression

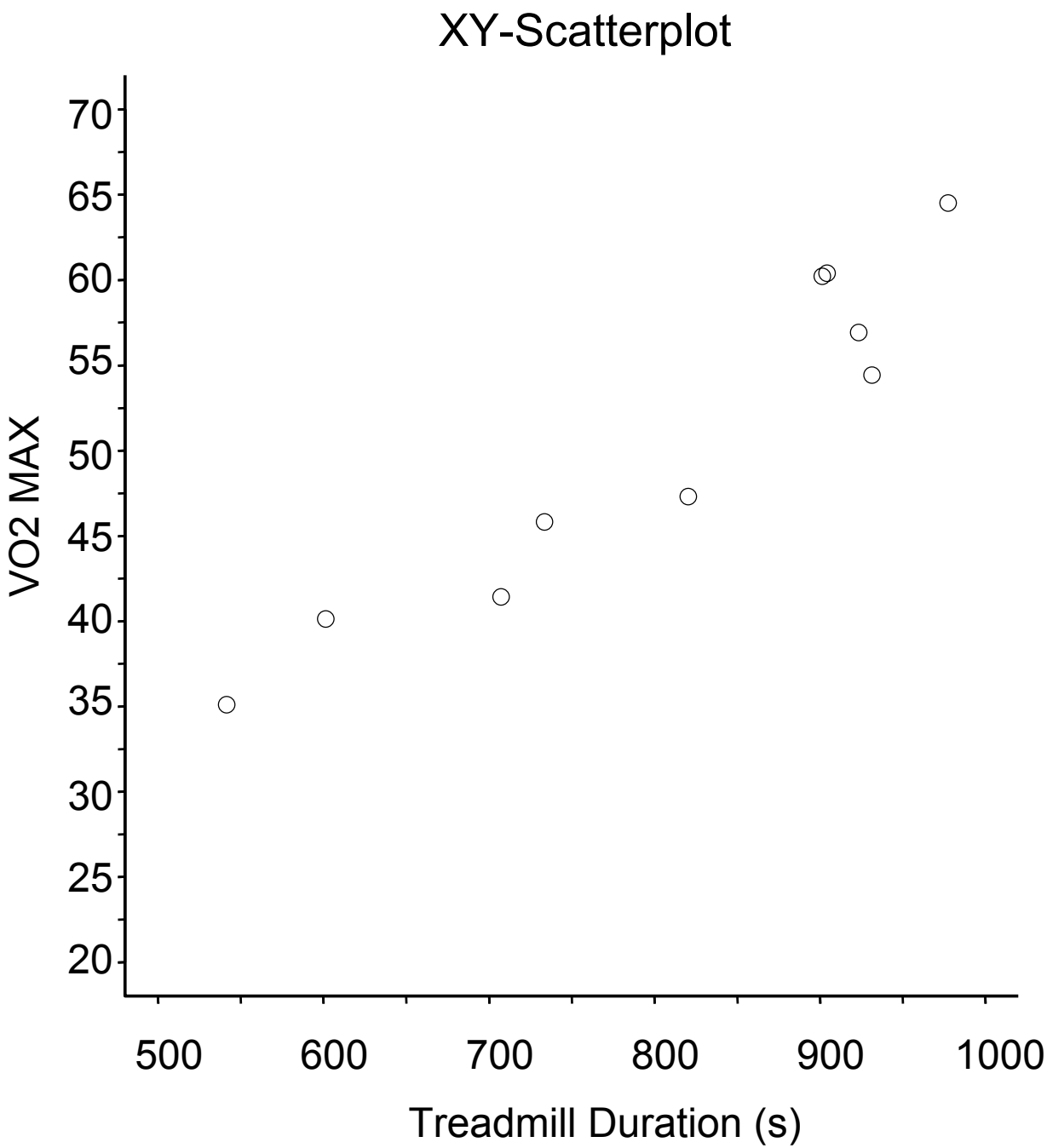
Regression provides the mathematical bases for drawing statistical inference about the existence of a functional relationship between two or more variables.

More explicitly, the methods of regression quantitatively characterize the functional relationship between an outcome variable and one or more independent variables that are believed to influence the value of the outcome.

Example: Treadmill Exercise Data.

Exercise Data from 10 Healthy Active Males.

Treadmill Duration (s)	VO_2 MAX
706	41.5
732	45.9
930	54.5
900	60.3
903	60.5
976	64.6
819	47.4
922	57.0
600	40.2
540	35.2


XY-Scatterplot
VO2 MAX
70
65
60
55
50
45
40
35
30
25
20
500
600
700
800
900
1000
Treadmill Duration (s)

Regression is typically utilized for one of the following reasons

- To assess whether or not a response variable, or perhaps a function of the response variable, is associated with one or more independent variables.

- To control for secondary factors that may influence the response variable but which are not consider as the primary independent variables of interest.

- To predict the value of the response variable at specific values of the independent variables.

The Simple General Linear Model Setting.

- The data consist of N paired measurements. Each pair consisting of a measurement of a response variable Y, and a measurement of a independent variable X.

- The elements of the response variable Y are assumed to be random, to have a continuous scale measure, and to have been measured without error.

- The elements of the independent variable X are assumed to be non-random and to have been measured without error. The elements of X can have either a continuous scale of measure or represent dichotomous (e.g gender; male,female), ordinal (e.g income; low, medium, high), or nominal (e.g. ethnicity; African American, Hispanic, White) categories.

.

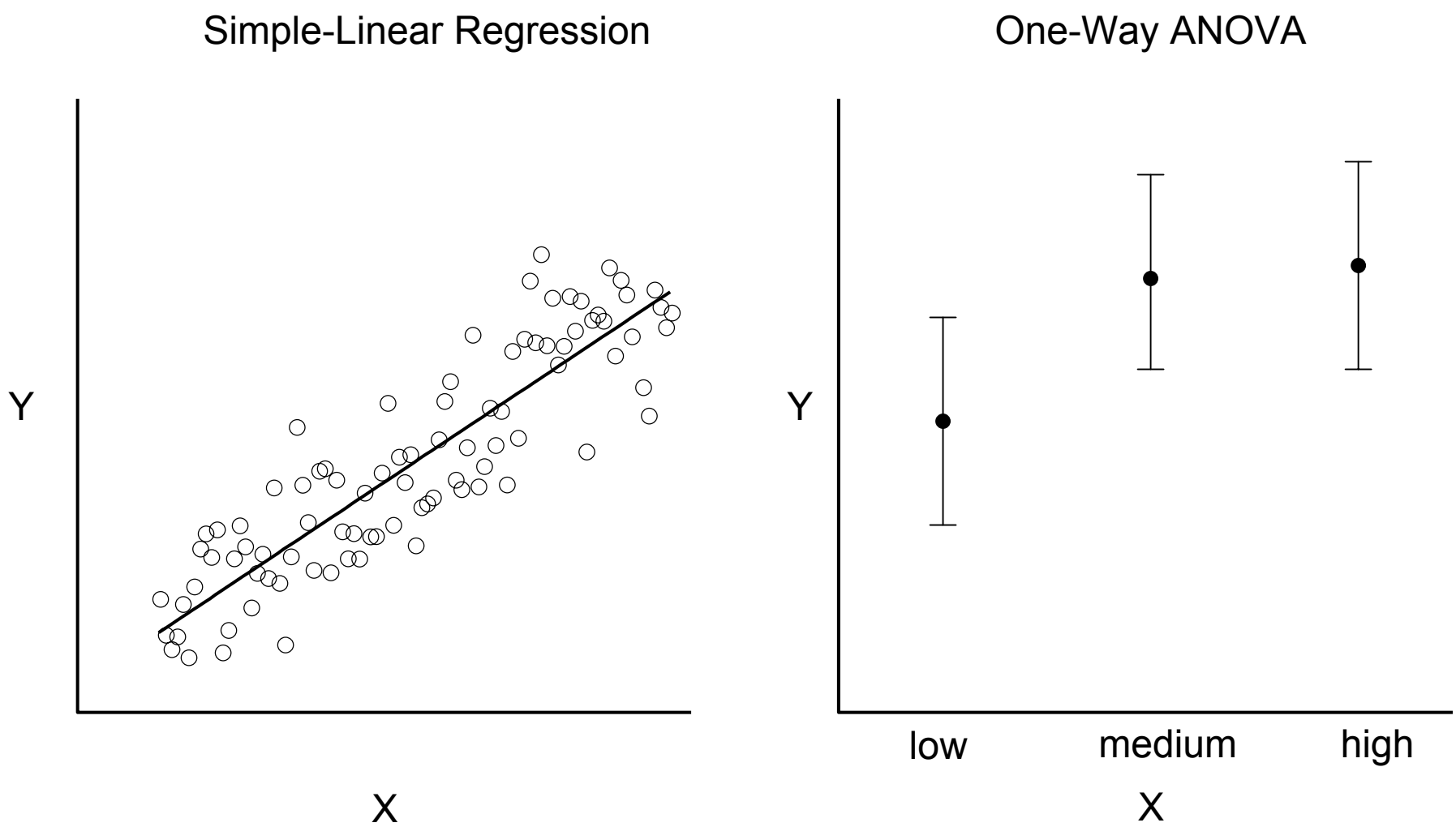
X-Continuous

- When the independent variable X has a continuous scale of measure the functional relationship between X and Y is summarized by a simple linear equation in which the expected value of y_i $(i = 1,2\cdots,n)$ is estimated as a linear function of x_i.
- When the independent variable X is continuous the analysis is referred to as “simple-linear regression”.

X-Categorical

- When the independent variable X is categorical the functional relationship between X and Y is typically summarized by a comparison of the mean of Y across the categories of X.
- When X is categorical the analysis is referred to as “One-Way ANOVA”.

Simple-Linear Regression and One-Way ANOVA Scenarios



The Simple-Linear Regression Equation.

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$(i = 1, 2, \cdots, n)$$

where

y_i = the i*th* value of the response.
x_i = the i*th* value of the independent variable.
α = the intercept parameter (y intercept).
β = the slope parameter ($\Delta y/\Delta x$).
ε_i =the random error associated with the i*th* response value.

The ε_i(s) are assumed to be independent identically distributed normal random variables with mean 0 and variance σ^2.

The Least Squares Simple-Linear Regression Model.

$$E(y_i|x_i) = \alpha + \beta x_i$$

$$(i = 1, 2, \cdots n)$$

where

$E(y_i|x_i)$ = the expected value of y_i at x_i. This is the population expected value or long-run average of y conditioned on the value of x. Example: population average blood pressure for a 30-year old.

x_i = the i*th* value of the independent variable.

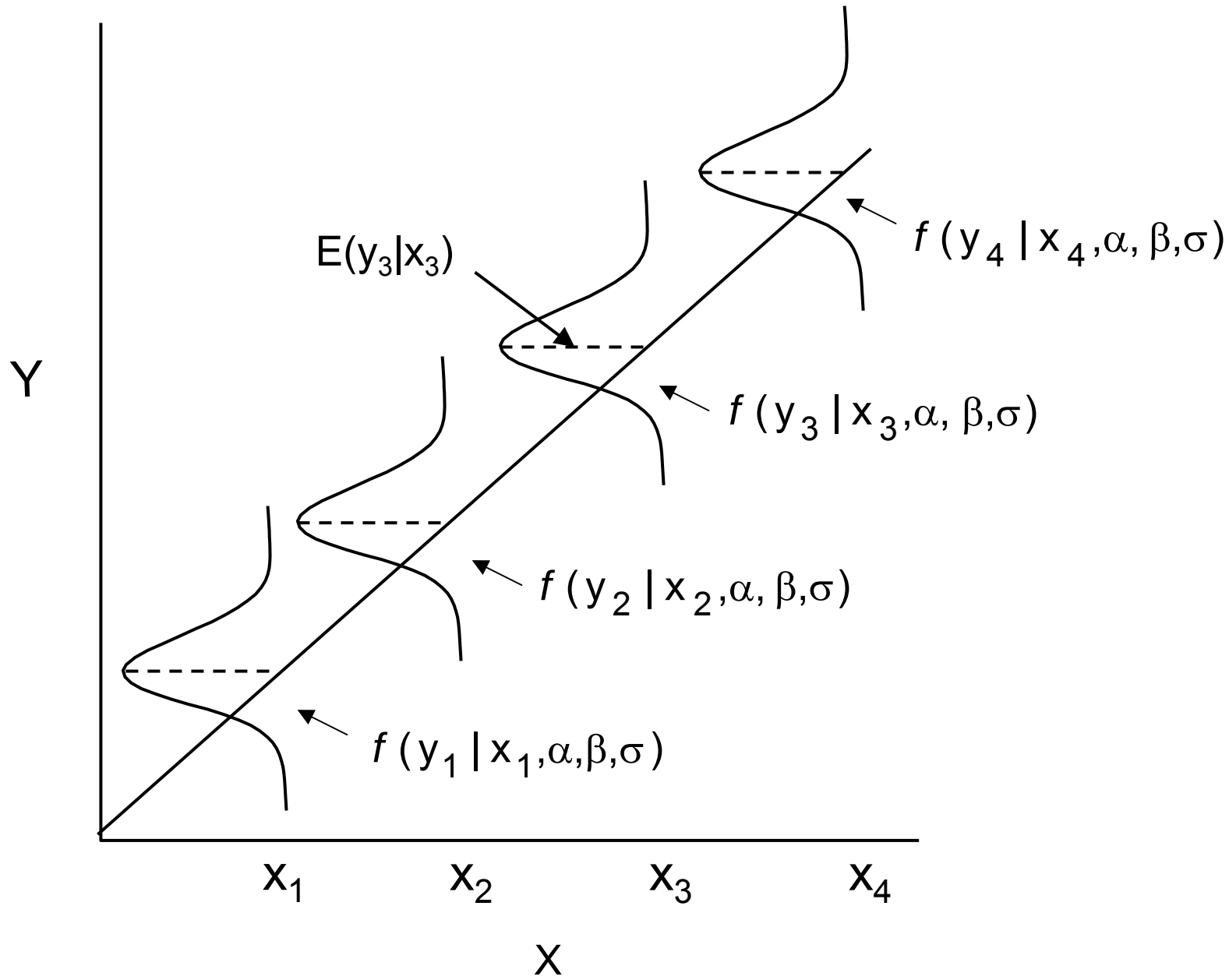
α = the intercept parameter (y intercept).

β = the slope parameter (Δ y/Δx).

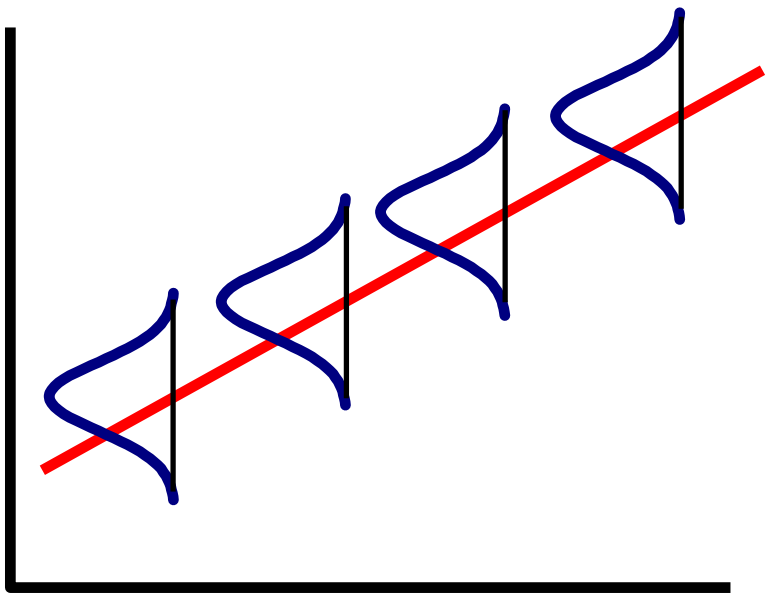
The Least Squares Linear Regression Model Assumptions.

- $E(y_i|x_i)$ is a linear in X.
- For each x_i, the conditional distribution of y_i $f(y_i|x_i,\alpha,\beta,\sigma)$ is normal.
- For each x_i, the conditional distribution of y_i $f(y_i|x_i,\alpha,\beta,\sigma)$ has variance σ^2.
- The y_i (*s*) are independent.

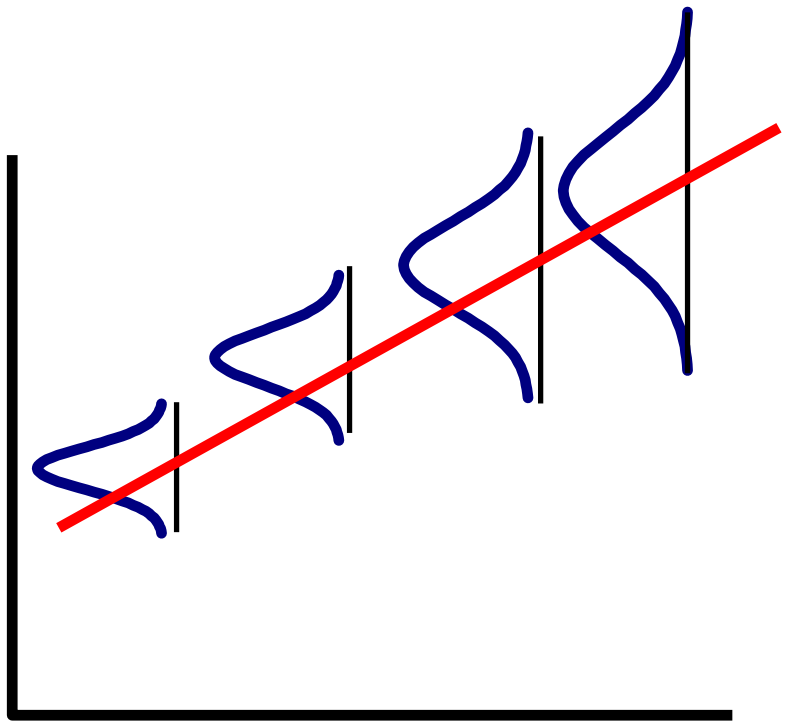
Model Assumptions



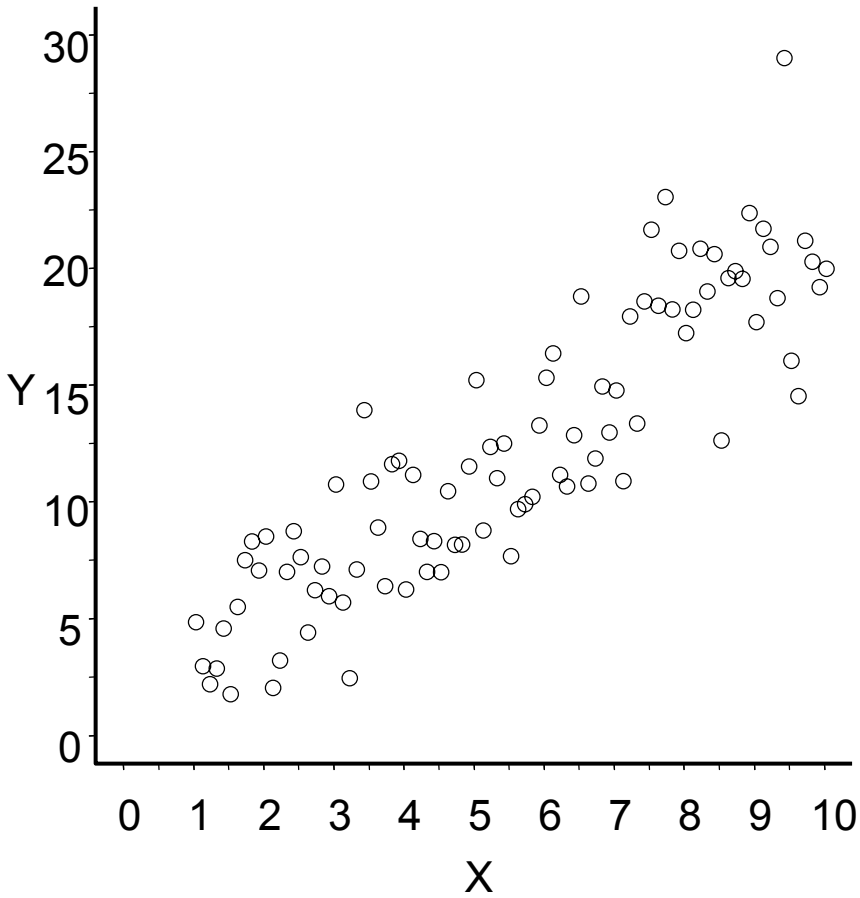
Homoscedastic



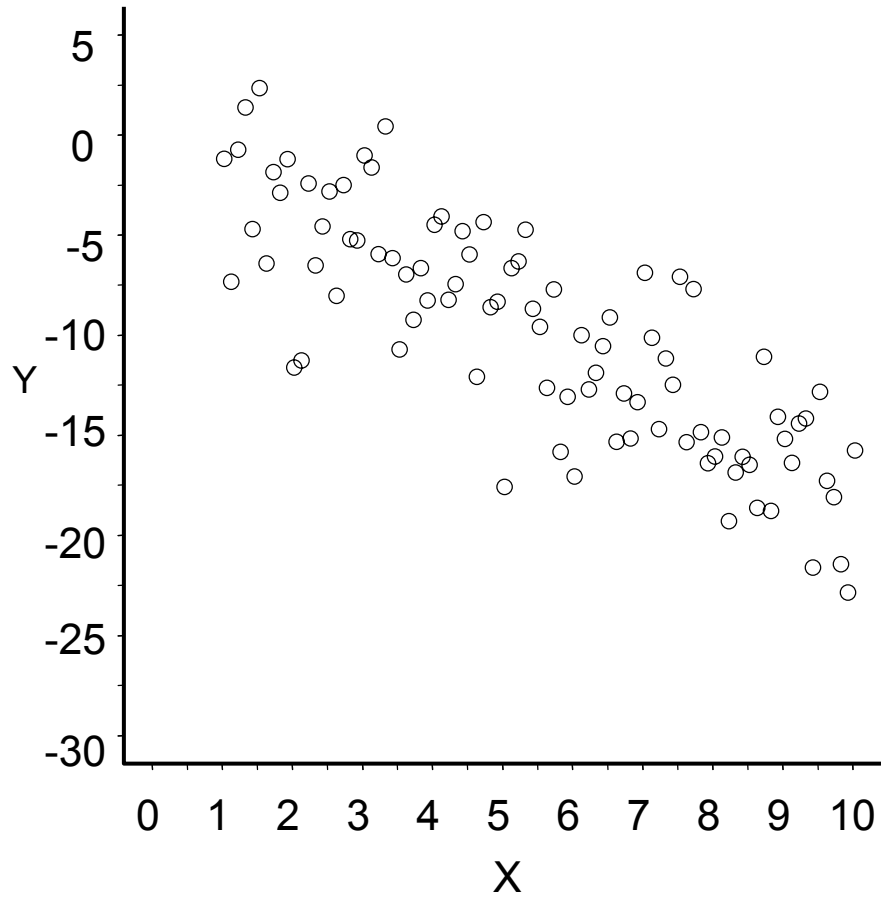
Heteroscedastic



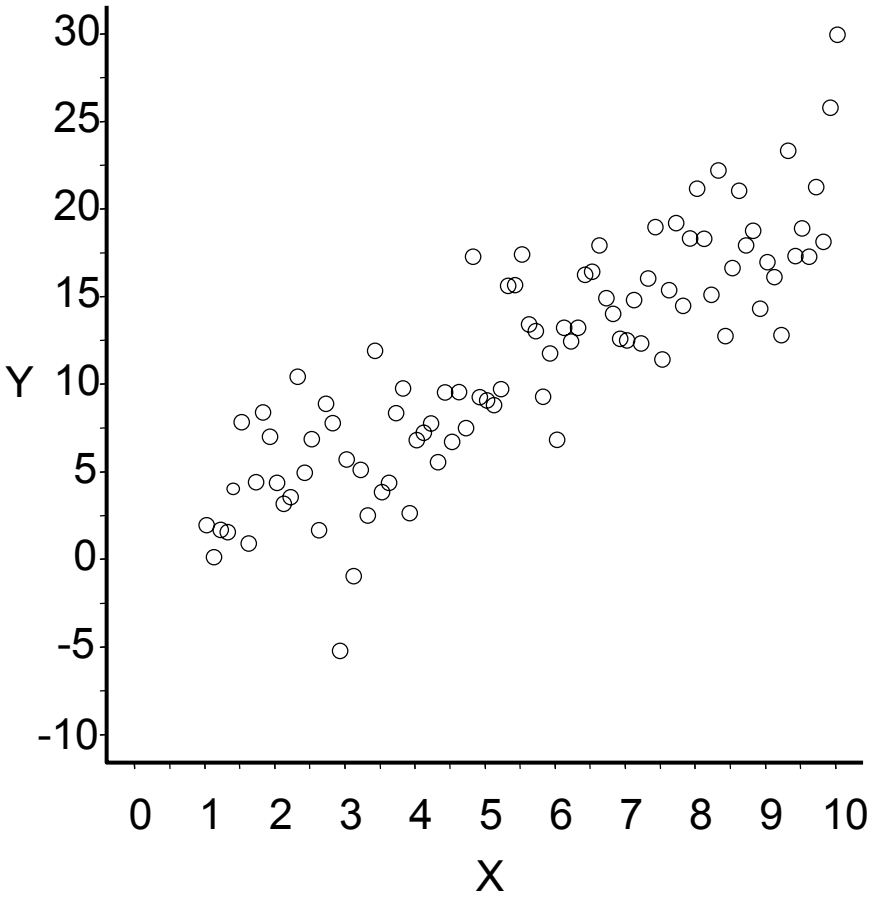
Examples of the Assumed Underlying Data Generating Process



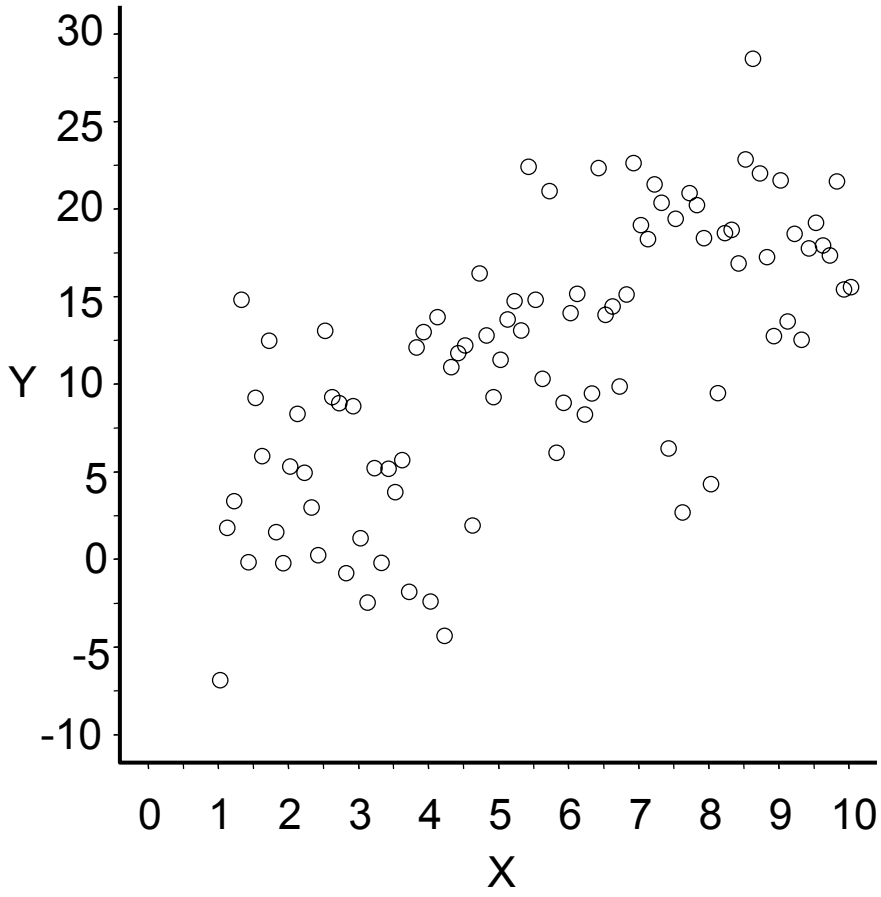
$y = 1 + 2x + \varepsilon$ where $\varepsilon \sim N(0,3)$



$y = 1 - 2x + \varepsilon$ where $\varepsilon \sim N(0,3)$



$y = 1 + 2x + \varepsilon$ where $\varepsilon \sim N(0,3)$



$y = 1 + 2x + \varepsilon$ where $\varepsilon \sim N(0,6)$

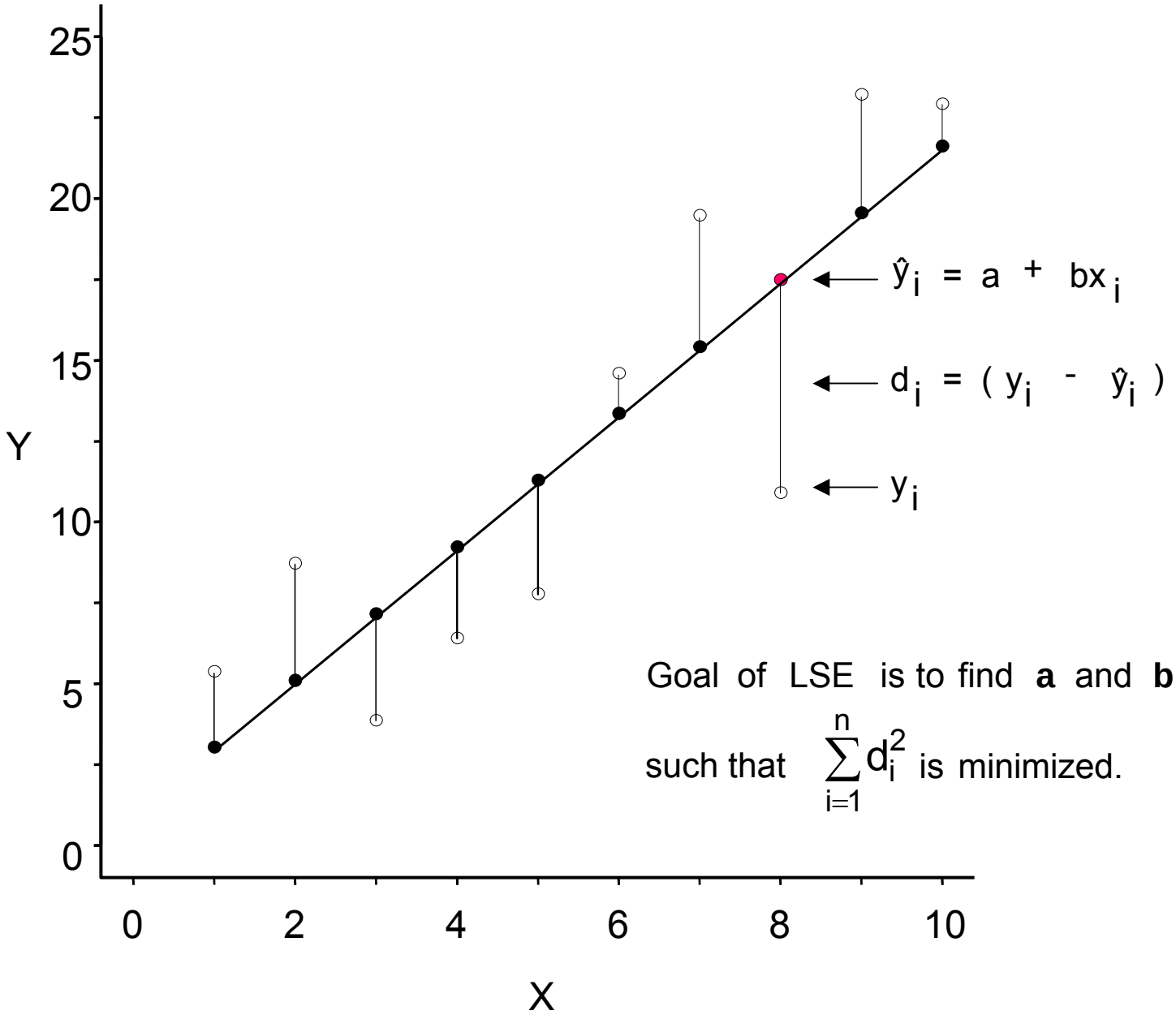
Least Squares Parameter Estimation.

The Goal of Least Squares Estimation

- The goal of least squares parameter estimation is to estimate the value of α and β from the observed sample of data in such a manner that the discrepancy between the observed value of the response and the predicted value of the response is minimized.

- The least squares estimation process selects from among the set of all possible regression lines the line which minimizes the sum of the squared difference between the predicted value of the response and the observed value of the response.

Least-Squares Estimation



Goal of LSE is to find **a** and **b** such that $\sum_{i=1}^{n} d_i^2$ is minimized.

The Properties of the Least Squares Estimators.

When the least squares model assumptions are valid the estimators for α and β have the following mathematical properties:

- The estimators are unbiased.
- The estimators have minimum variance among all unbiased estimators of α and β.

Estimating α and β by the Method of Least Squares (LS).

$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$ the mean value of the independent variable X.

$\bar{y} = \sum_{i}^{n} y_i$ the mean value of the response variable Y.

$L_{xx} = \sum_{i}^{n} (x_i - \bar{x})^2$ the corrected sum of squares of X.

$L_{xy} = \sum_{i}^{n} (x_i - \bar{x})(y_i - \bar{y})$ the corrected sum of cross products.

$b = \frac{L_{xy}}{L_{xx}}$ the LS estimate for the slope parameter β.

$a = \bar{y} - b\bar{x}$ the LS estimate for the intercept parameter α.

Estimating of the Precision of the Least Squares Estimators.

- Based on the observed sample of data, we can estimate the level of precision of the least squares estimators for α and β. This is accomplished by computing the standard error (SE) of the estimator.

- The SE provides an estimate of the magnitude of the variation that we would expect to see in the parameter estimate from one sample of data to the next.

- The SE is a function of the magnitude of the discrepancy between y_i and the predicted value of y_i, the sample size (n), and the range of X.

The SE(s) of the Least Squares Estimators a and b.

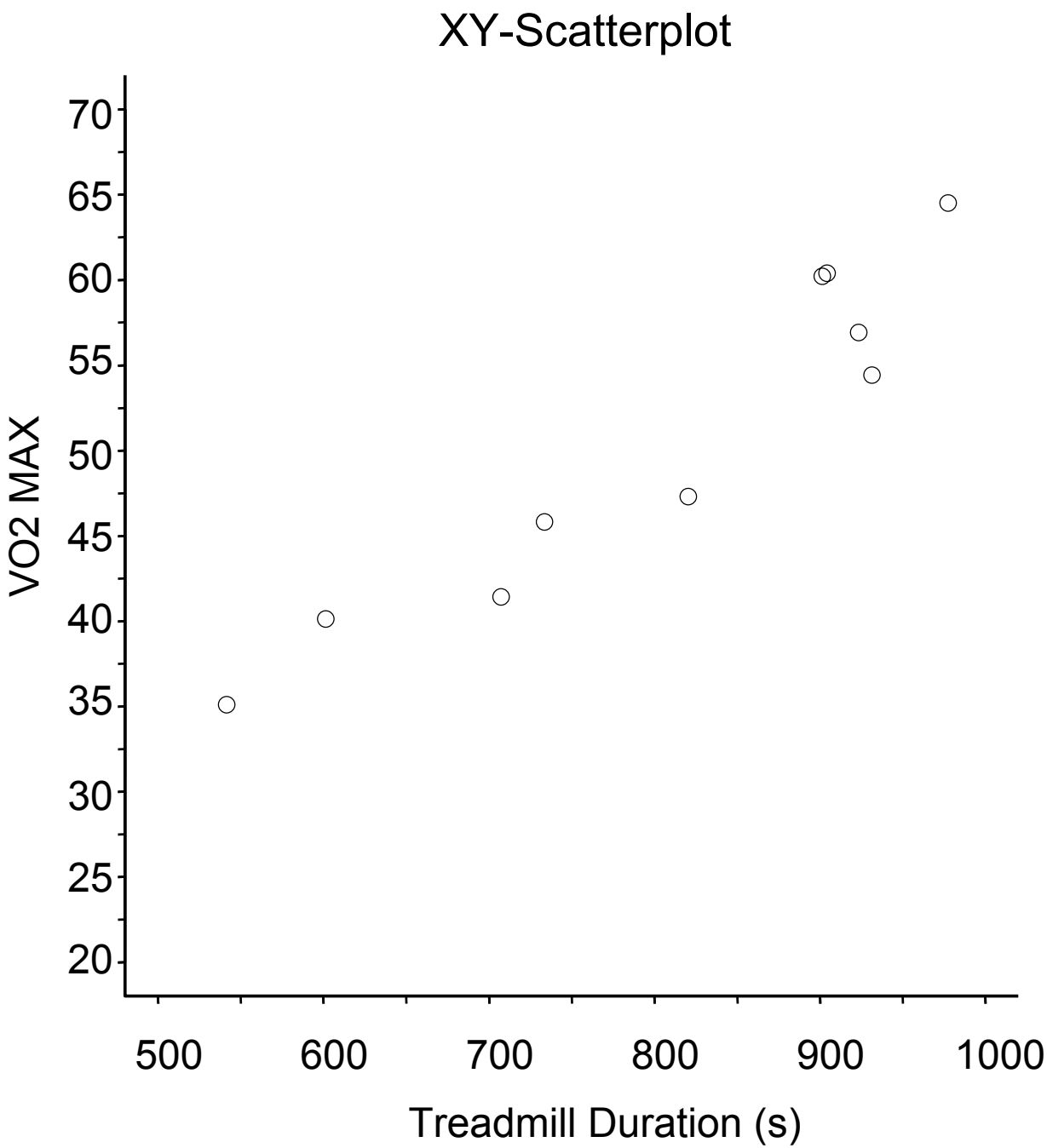
$$SE(a) = \sqrt{s^2_{y.x}\left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right)} \quad \text{where} \quad s^2_{y.x} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p)$$

$$SE(b) = \sqrt{\frac{s^2_{y.x}}{L_{xx}}}$$

Example: Treadmill Exercise Data.

Exercise Data from 10 Healthy Active Males.

Treadmill Duration (s)	VO_2 MAX
706	41.5
732	45.9
930	54.5
900	60.3
903	60.5
976	64.6
819	47.4
922	57.0
600	40.2
540	35.2


XY-Scatterplot
VO2 MAX
70
65
60
55
50
45
40
35
30
25
20
500
600
700
800
900
1000
Treadmill Duration (s)

The Least Squares Estimates for α and β.

$$\bar{x} = \frac{1}{10}(706 + 732 + \cdots + 540) = 802.8$$

$$\bar{y} = \frac{1}{10}(41.5 + 45.9 + \cdots + 35.2) = 50.7$$

$$L_{xx} = (706 - 802.8)^2 + (732 - 802.8)^2 + \cdots + (540 - 802.8)^2 = 204711.6$$

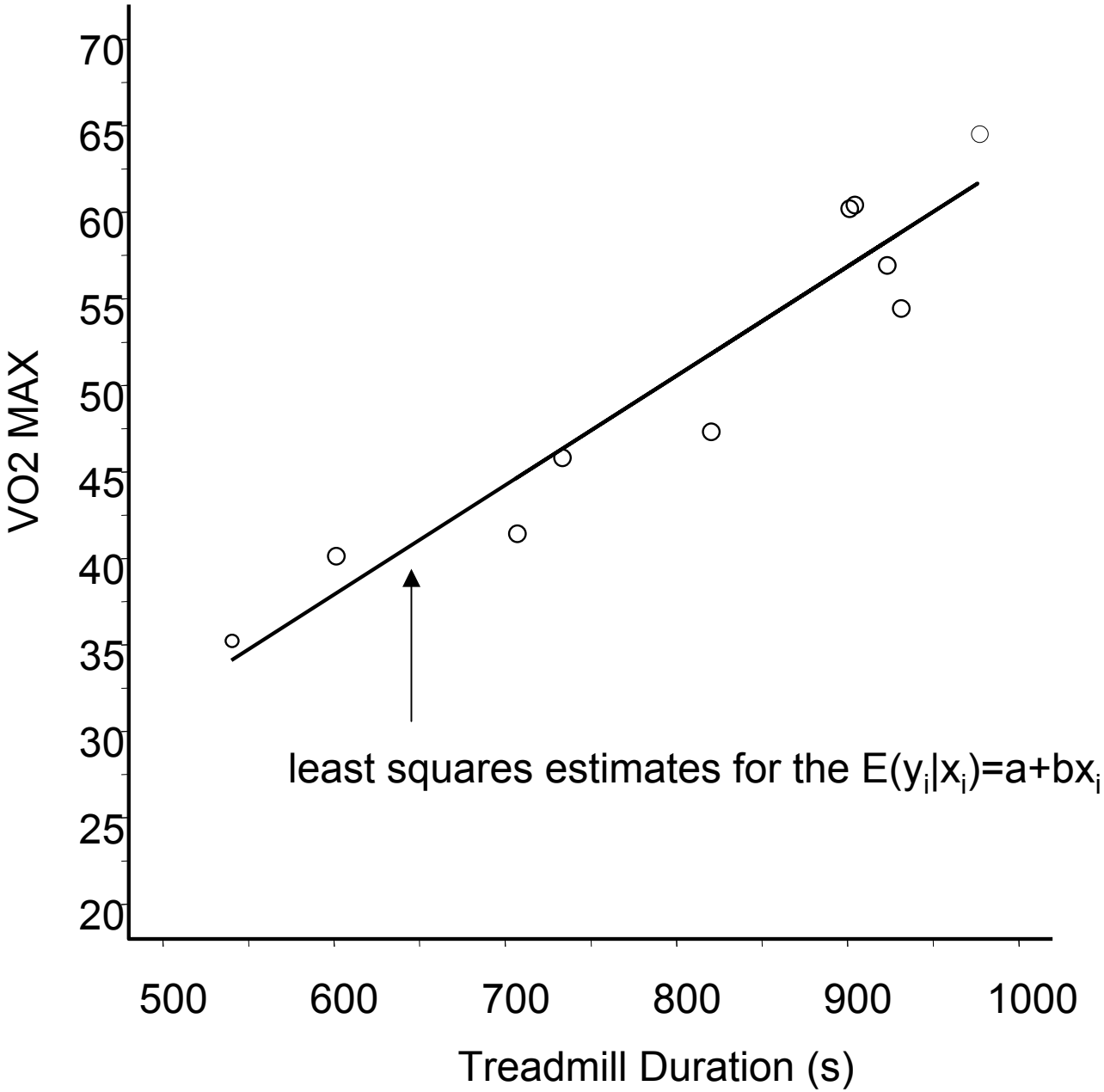
$$L_{xy} = (706 - 802.8)(41.5 - 50.7) + \cdots + (540 - 802.8)(35.2 - 50.7) = 12936.6$$

$$b = \frac{L_{xy}}{L_{xx}} = \frac{12936.6}{204711.6} = 0.063$$

$$a = \bar{y} - b\bar{x} = 50.7 - 0.0632 \text{ x } 802.8 = -0.026$$

Regression Equation

$$E(y_i \mid x_i) = -0.026 + 0.063 \frac{VO_2 \text{ Max}}{s}(\text{treadmill duration}_i \text{ (s)})$$


70
65
60
55
50
45
40
35
30
25
20
VO2 MAX
500
600
700
800
900
1000
Treadmill Duration (s)
least squares estimates for the $E(y_i|x_i)=a+bx_i$

The Least Squares Estimate of SE for a and b.

$$s^2_{y.x} = \frac{1}{8}\left[(41.5-44.6)^2 + (45.9-46.2)^2 + \cdots + (35.2-34.1)^2\right] = 10.9$$

$$SE(a) = \sqrt{10.9\left(\frac{1}{10} + \frac{(802.8)^2}{204711.6}\right)} = 5.9$$

$$SE(b) = \sqrt{\frac{10.9}{204711.6}} = 0.007$$

Hypothesis Tests for the Simple Linear Regression Model

- For the simple-linear regression model we assume under the null hypothesis that there is no linear association between the value of the outcome (y_i) and the value of the independent variable (x_i), or equivalently that the value of β is equal to zero.

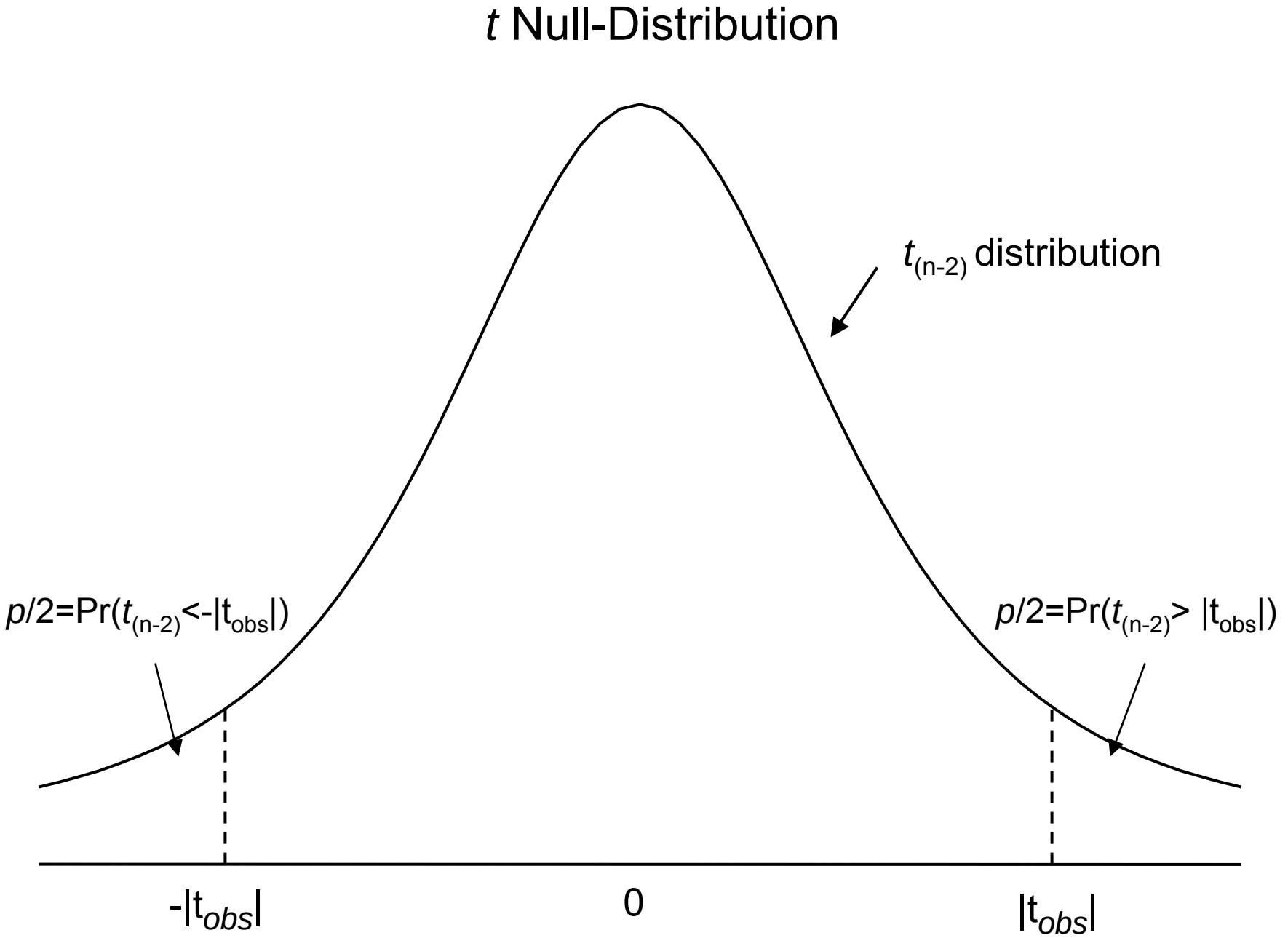
- In layman's terms the hypothesis test is asking the question what is the chance, given the sample of data that we observed, of observing a sample of data that is more contradictory of the null hypothesis of no association.

t-Test for the Hypothesis of No Association between Y and X.

Hypothesis: Ho: $\beta = 0$ versus H_a: $\beta \neq 0$

$$t_{obs} = \frac{b}{\sqrt{SE(b)}}$$

Under Ho: t_{obs} follows a t_{n-2} distribution. For a two-sided test with significance level α we reject Ho: if $| t_{obs} | > t_{(n-2,1-\alpha/2)}$.


t Null-Distribution
t(n-2) distribution
p/2=Pr(t(n-2)<-|tobs|)
p/2=Pr(t(n-2)> |tobs|)
-|tobs|
0
|tobs|

Example: Regression of VO_2 Max onto Duration of Exercise.

Hypothesis: Ho: $\beta = 0$ versus H_a: $\beta \neq 0$

$$t_{obs} = \frac{b}{\sqrt{SE(b)}} = \frac{L_{xy}/L_{xx}}{\sqrt{(s^2_{y.x}/Lxx)}} = \frac{0.063}{0.0072} = 8.7$$

For a two-sided test with significance level 0.05

$t_{(8,0.975)} = 2.31$

$| t_{obs} | > t_{(8,0.975)}$, which $\Rightarrow$ we should reject Ho: $\beta = 0$.

Typical Computer Generated Regression Summary.

Table 1. Regression ANOVA Table.

Source	df	SS	MS	F	P
Regression	1	817.52	817.52	75.10	<0.001
Error	8	87.09	10.89		
Total	9	904.61			

Table 2. Parameter Estimates.

Parameter	Parameter Estimate	SE	t_{obs}	$P(T>t_{obs})$
Intercept	-0.022	5.946		
Duration	0.063	0.007	8.7	<0.001

- If b and a are respectively, the estimated slope and intercept of the least squares regression line, and SE(b) and SE(a) are the least squares estimates of standard error, then the two-sided 100%(1-alpha) confidence intervals for β and α are given by $b \pm t_{(n-2,1-\alpha/2)}SE(b)$ and $a \pm t_{(n-2,1-\alpha/2)}SE(a)$.

- The 100%(1-alpha) confidence interval for β and α can be interpreted in the following manner. If we were to resample an infinite number of times from the same reference population and for each of the respective samples we were to fit a regression model to the data and then construct 100% (1-alpha) confidence intervals for β and α by the preceding criterion, we would expect 100%(1-alpha) of these confidence intervals to contain the true value of the respective parameter.

Example: Regression of VO_2 Max onto Duration of Exercise

95% CI for α.

$$95\%\ CI(\alpha) = a \pm t_{(8,0.975)}SE(a).$$
$$= -0.026 +/- 2.31 \times 5.9$$
$$= (-13.6, 13.6)$$

95% CI for β.

$$95\%\ CI(\beta) = b \pm t_{(8,0.975)}SE(b).$$
$$= 0.063 +/- 2.31 \times 0.007$$
$$= (0.047, 0.079)$$

Residual Diagnostics for Checking Goodness of Model Fit.

- Once you have fit your initial regression model the things to check include the following:

  a) The assumption of constant variance.
  b) The assumption of normality.
  c) The correctness of functional form.
  d) Check outliers.

- All of the preceding diagnostics checks can be conducted with graphics. The residuals ($e_i = y_i - \hat{y}_i$) from the model or a function of the residuals play an important role in all of the model diagnostic procedures.

a) Checking the assumption of constant variance.

- Plot the studentized residuals from your model versus their fitted values. Examine if the variability between the residuals remains relatively constants across the range of the fitted values.

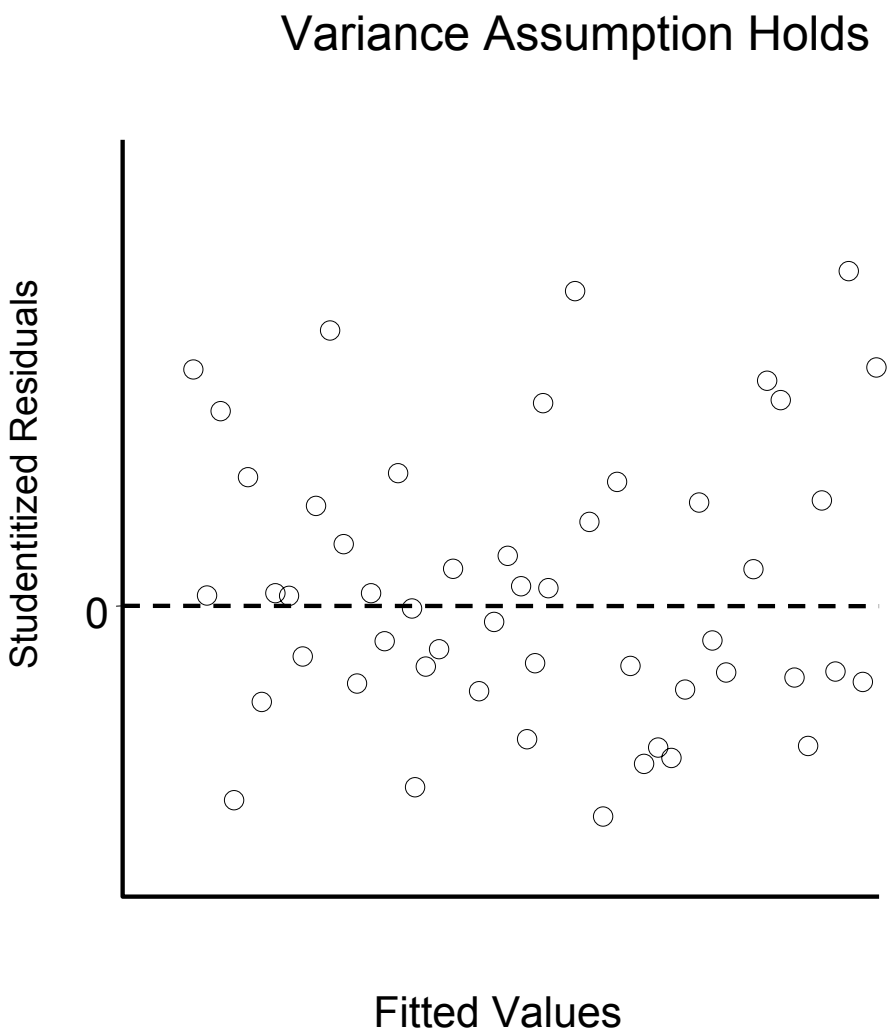
b) Checking the assumption of normality.

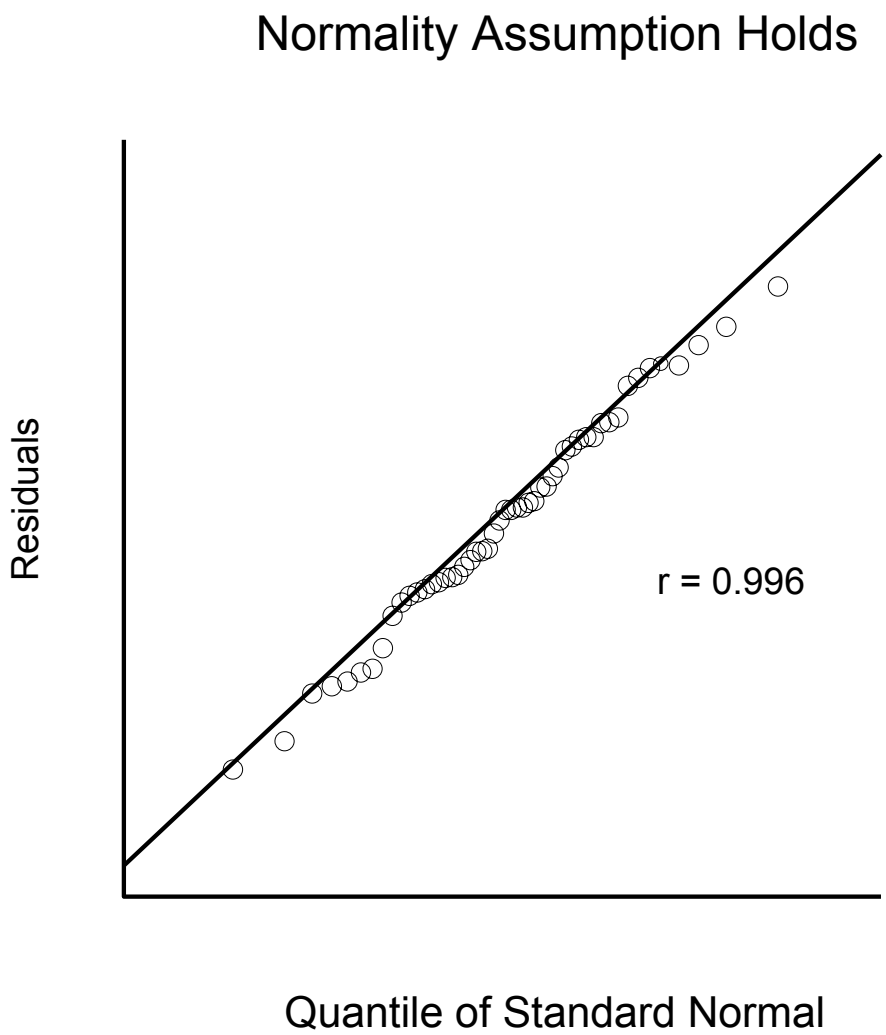
- Plot the residuals from your model versus the expected value of the residual under normality (Normal Probability Plot). If the residuals are normally distributed the residuals will fall along a 45° line.

c) Checking for the correctness of functional form.

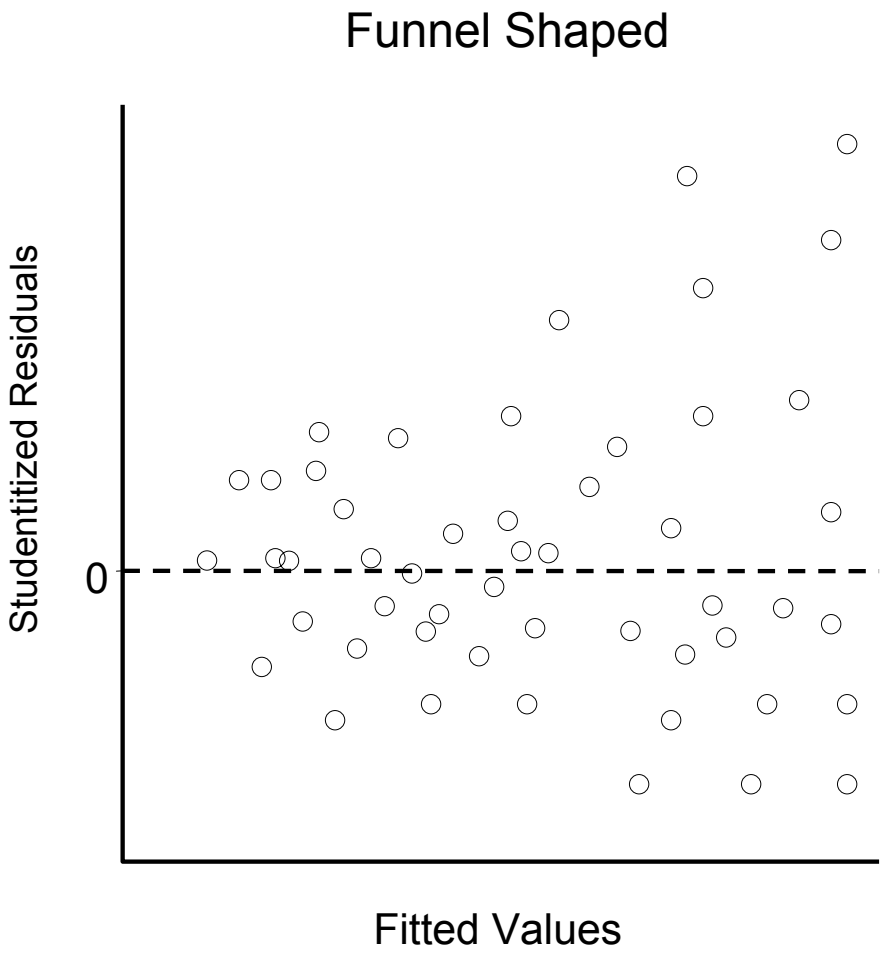
- Plot the residuals from your model versus their fitted values. Examine the residual plot for evidence of a non-linear trend in the value of the residual across the range of the fitted values.

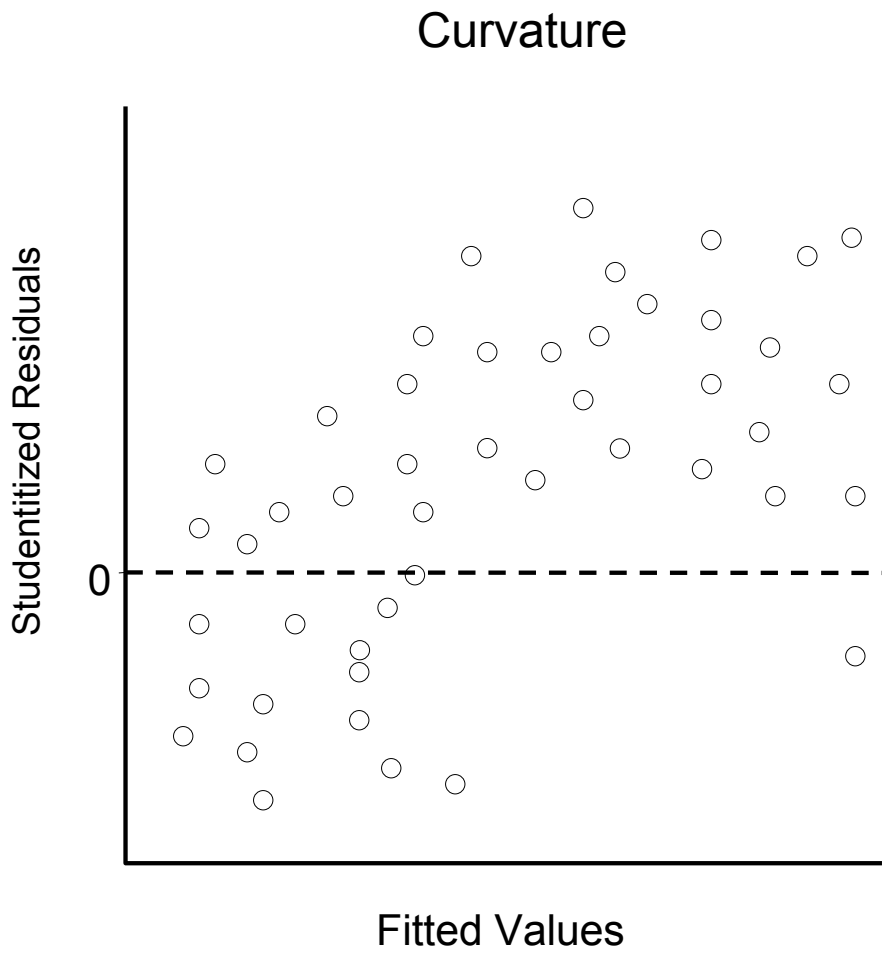
Residual Diagnostic





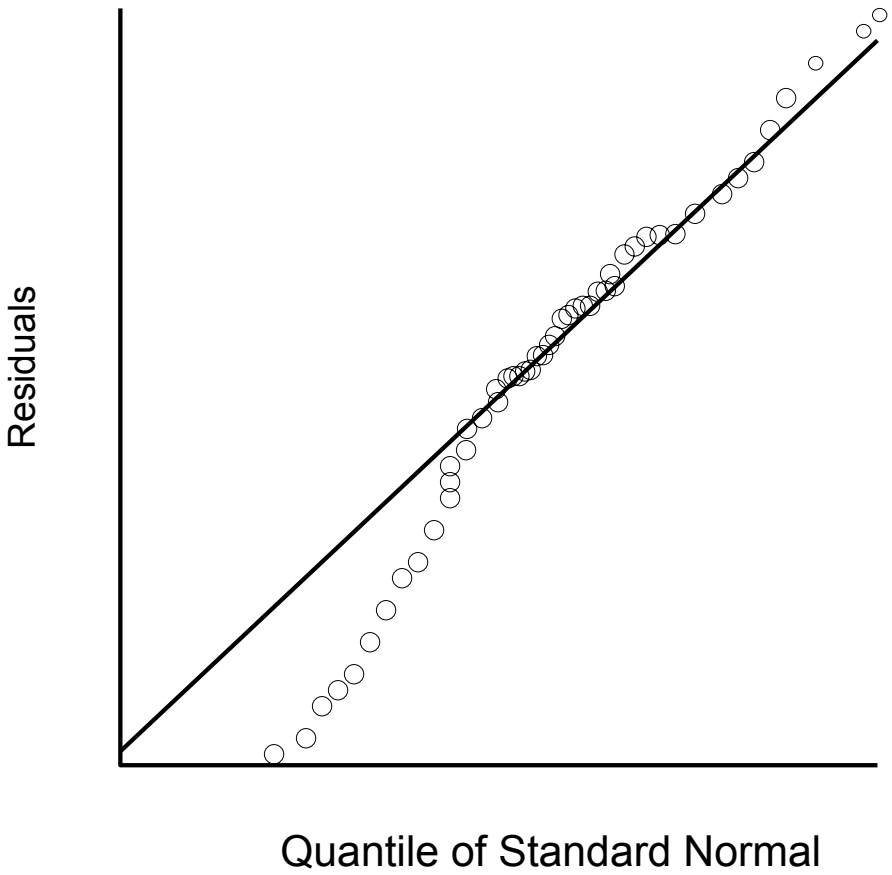
Evidence of Non-Constant Variance



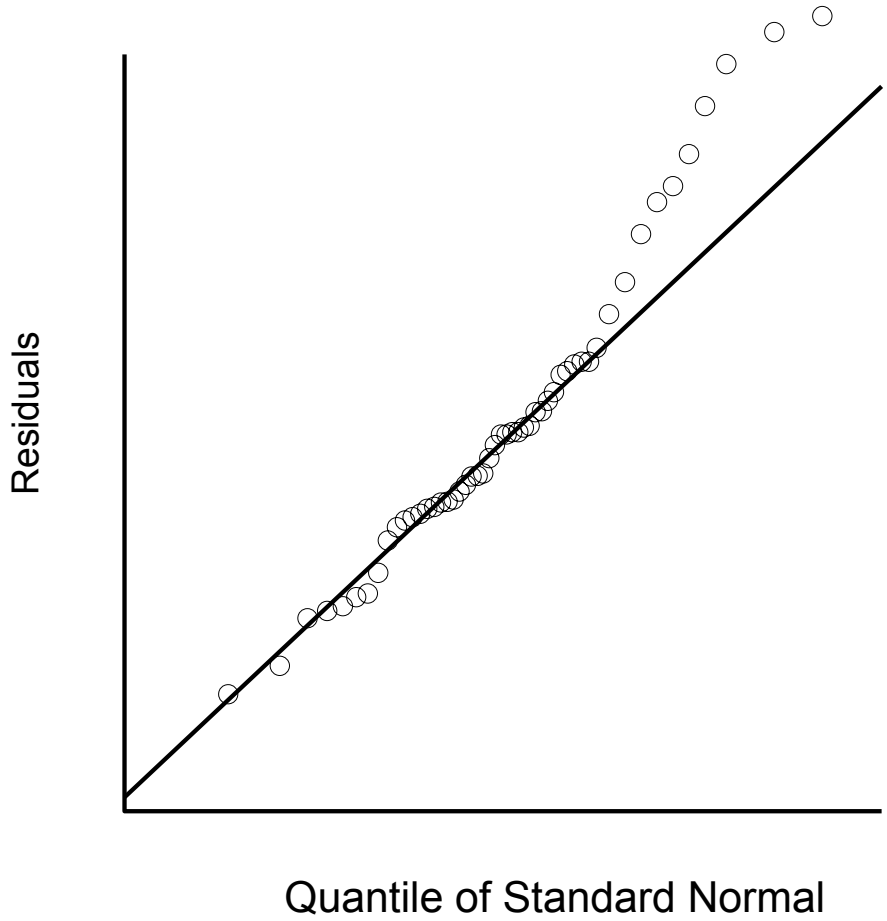


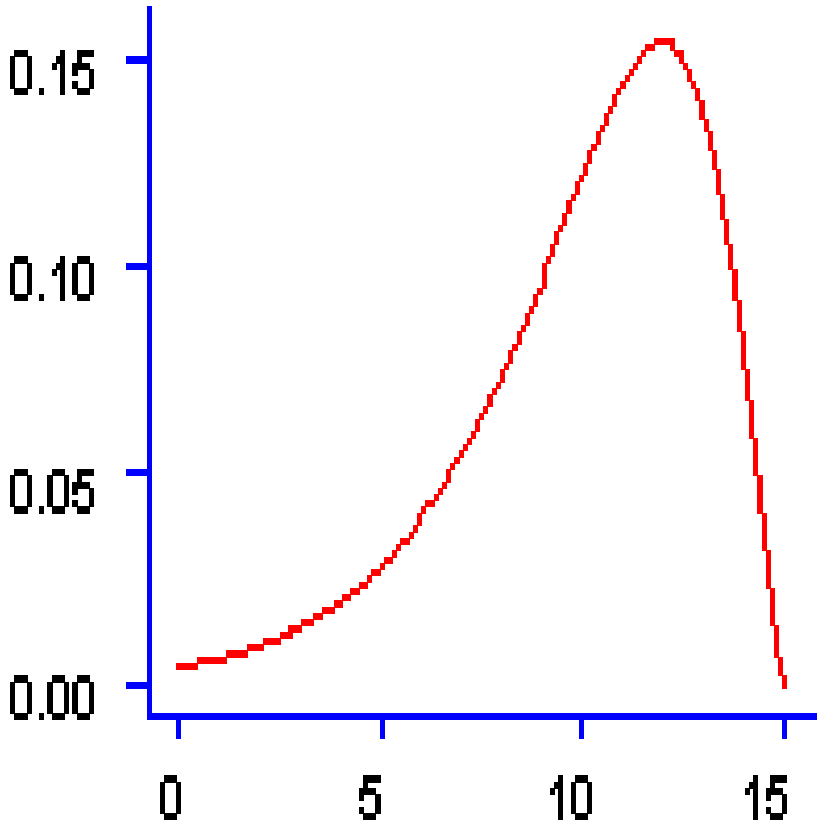
Evidence of Non-Normality

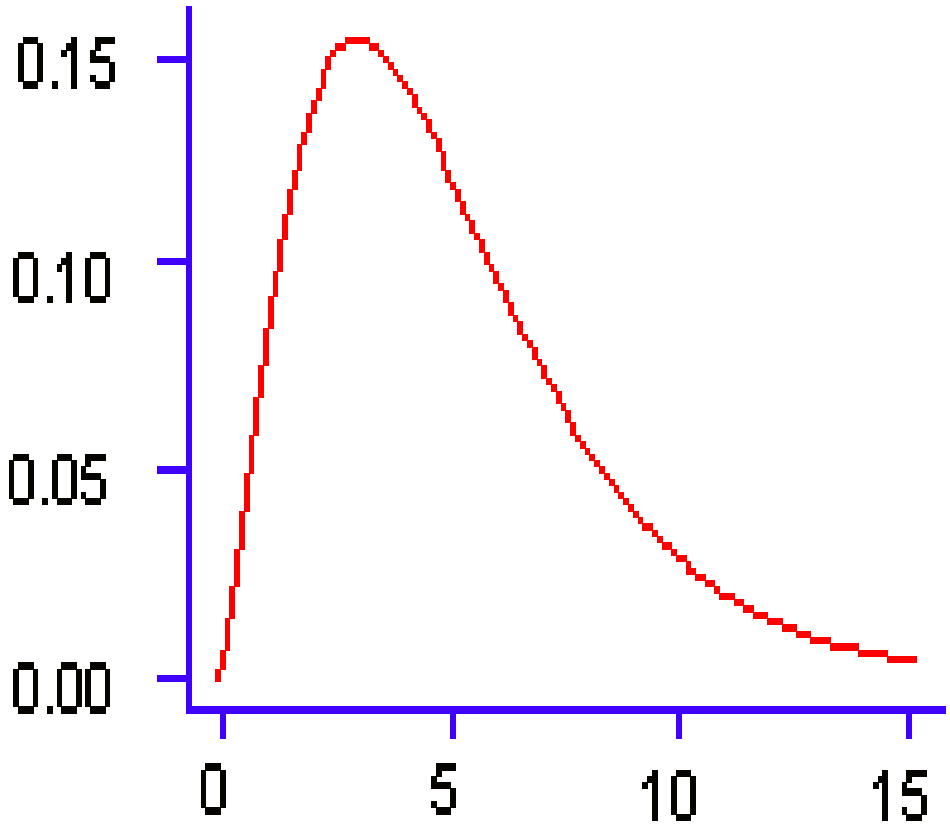
Skewed Left



Skewed Right





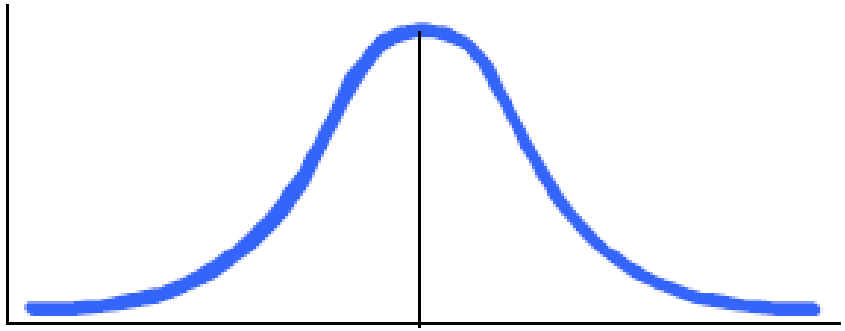


- negative skewness
- skew to the left
- mean < median
- lower tail longer than upper tail

- positive skewness
- skew to the right
- mean > median
- upper tail longer than lower tail

Skewness

Figure 2-13 (b)



Mode = Mean = Median

SYMMETRIC

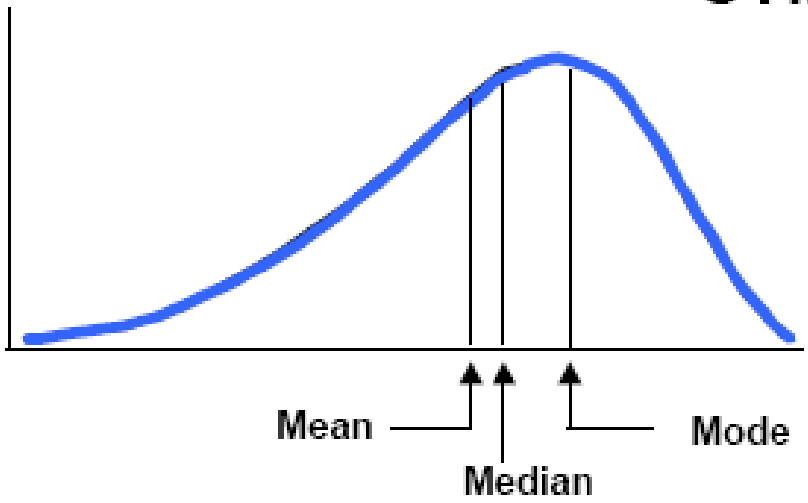
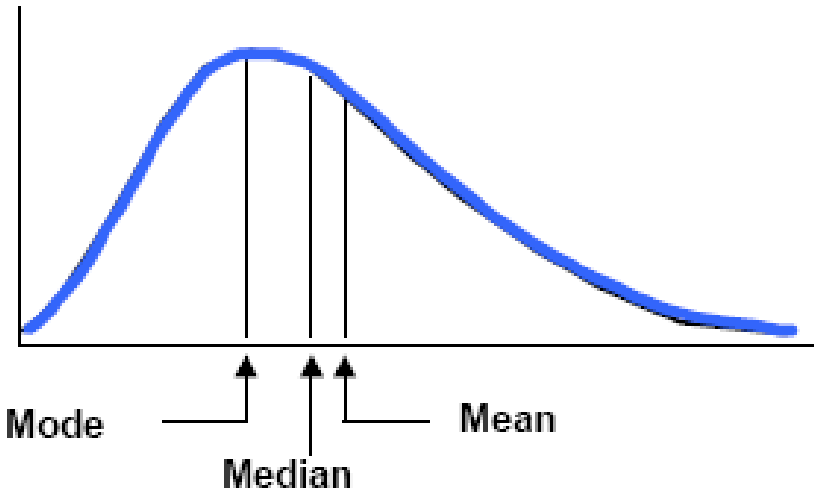


Figure 2-13 (a)

SKEWED LEFT
(negatively)



SKEWED RIGHT
(positively)

Figure 2-13 (c)

- Outliers

Outliers, either with respect to the response variable, or with respect to the independent variable can have a major influence on the value of the regression parameter estimate. Gross outliers should always be checked first for data authenticity. In terms of the response variable, if there is no legitimate reason to remove the offending observation(s) it may be informative to fit the regression model with and without the outlier(s). If statistical inference changes depending on the inclusion or the exclusion of the outlier(s), its is probably best to use a robust form of regression, such as median least squares or least absolute deviation regression. For the independent variable, if your sample of data is reasonably large, it is generally recommended to use some form of truncation that restricts the range of the independent variable.

Measure of Predictive Accuracy.

- The coefficient of determination (R^2) is commonly use as an index of predictive accuracy. The value of R^2 is simply the ratio of the regression sum of squares to the total sum of squares (SSR/SST) x 100%. The value of R^2 is interpreted as the percentage of the total variance in the outcome Y that is explained by the independent variable X.
- It is important to note that the value of R^2 will always increase when an additional independent variable is added to the regression model. In the multiple regression setting an adjusted R^2 is frequently used as the measure of predictive accuracy. The adjusted R^2 is the ratio of the regression sum of square to the total sum of squares adjusted by the number of degrees of freedom (p) associate with the sum of squares. ($R^2_{adjusted}$ = SSS/(p –1)/SST/(n-1)).

Example Computation of R^2 for the VO_2 Max Data.

Table 1. Regression ANOVA Table.

Source	df	SS	MS	F	P
Regression	1	817.52	817.52	75.10	<0.001
Error	8	87.09	10.89		
Total	9	904.61			

$$R^2 = \frac{SSR}{SST} = \frac{817.52}{904.61} = 0.90$$

The R^2 of 0.90 implies that approximately 90% of the variation in VO_2 Max could be explained by knowing the length of time that each subject had spent on the treadmill.

Example: Analysis of Variance (ANOVA) and Covariance (ANCOVA)
(An application of simple linear regression with a natural extension to multivariable regression)

Example: Is milk consumption (dietary calcium intake) associated with lower systolic blood pressure (SBP) [McCarron et al. *Science* 1982; 217-269, Abbott et al. *Stroke* 1996; 27:813-818]?

Consider the following population based sample of Japanese-American men (entries represent SBP in mm Hg):

	Milk consumption		
	None	Some (<8 oz/d)	Lots (>8 oz/d)
	155	150	135
	150	140	130
	140	135	120
	135	120	115
	145	130	120
Mean	145	135	124

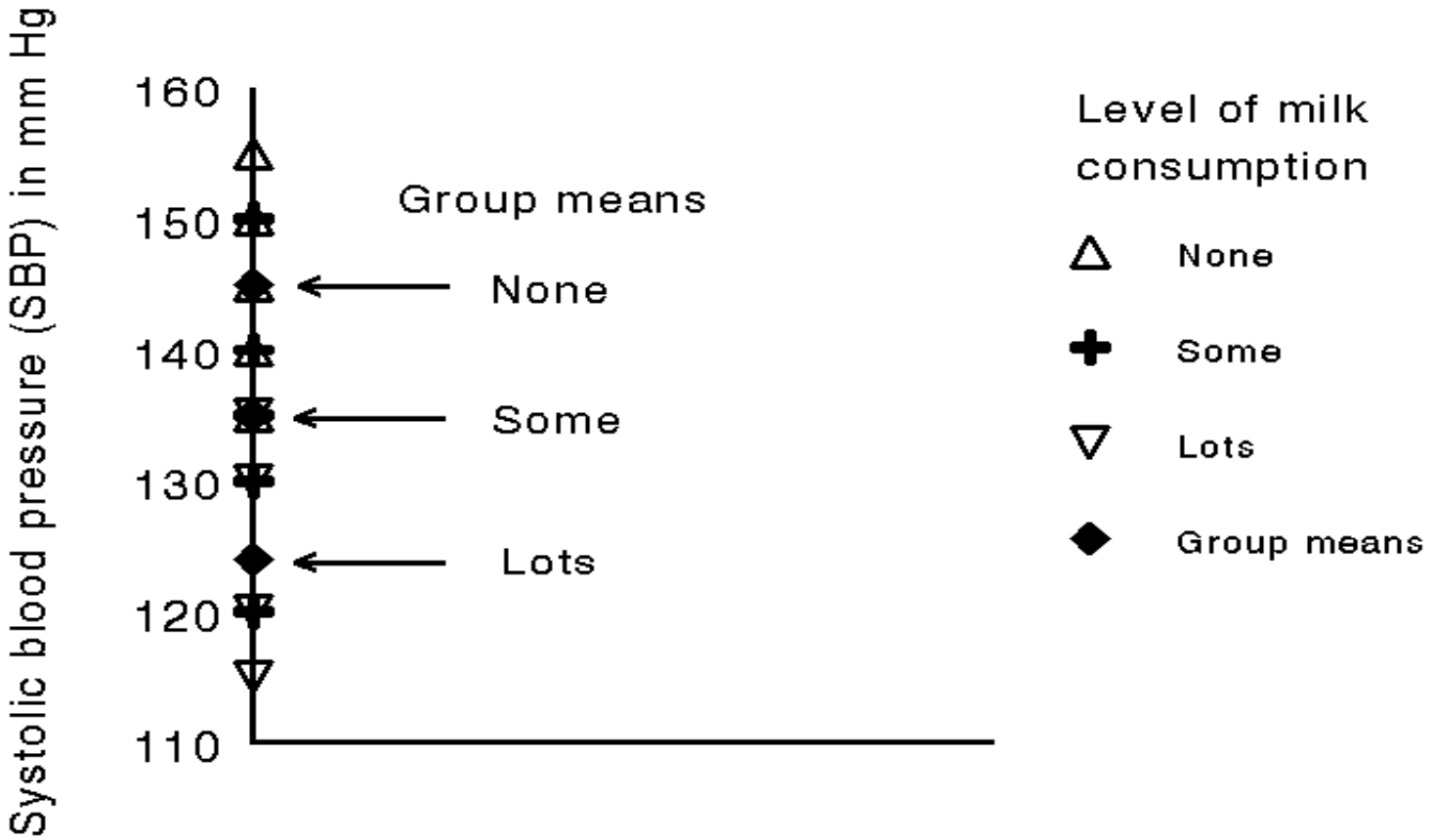
Note: This is like a regression problem, where we might like to know whether milk consumption (X), can be used to say something (or can be used to make predictions) about SBP (Y) among men. If X is related to Y, then perhaps promoting changes in X will provide an effective way of changing Y.

How do we determine if milk consumption is related to SBP?

First, we might note that milk intake is usually consumed in units, and it might be reasonable to compare the 3 groups as opposed to treating milk consumption as strictly a continuous variable (i.e., people don't report that they consume 1.234 oz/d).

To compare the three groups, we use a procedure called *Analysis of Variance*. It is a procedure which is intended to compare average responses for 3 or more groups. It is called *Analysis of Variance* because it helps us examine variability among group averages. If the variability is low, then groups are similar. If the variability is high, then groups differ.

A plot of the data appear as follows:



Conclusions:

	Lots	Some	None
Mean	124	135	145

Note: Group means with a common underline are not significantly different. Group means that do not share a common underline are significantly different.

Conclusion #1: Men who drink lots of milk have significantly lower blood pressures than men who drink no milk ($p<0.05$).

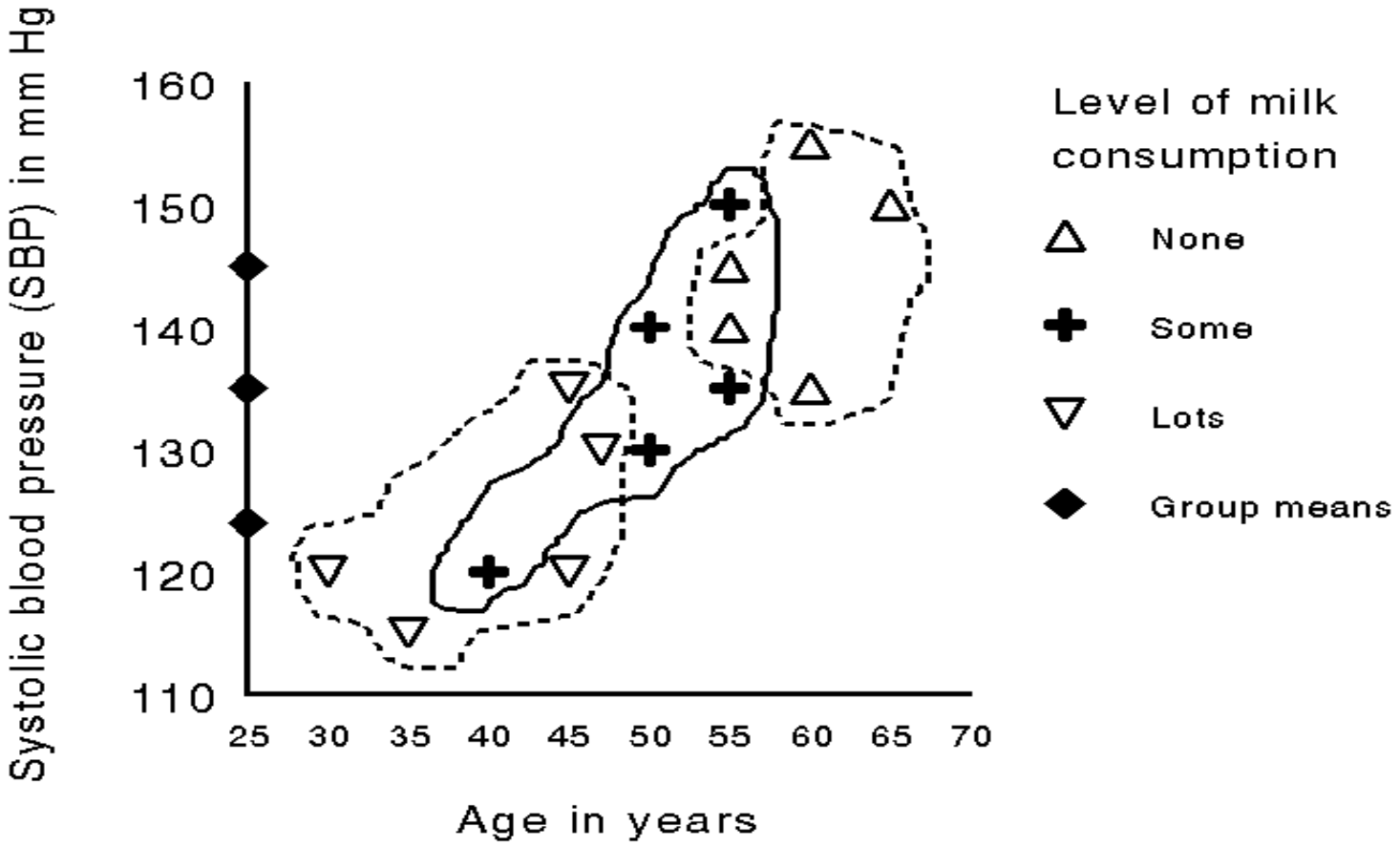
Conclusion #2: Differences in systolic blood pressure levels between adjacent groups of milk consumption are on the order of 10 mm Hg.

Conclusion #3: Drink milk!

Problem: Younger men in this group consume more milk than men who are older.

Question: Do consumers of milk have lower blood pressure because of the milk or because they are younger?

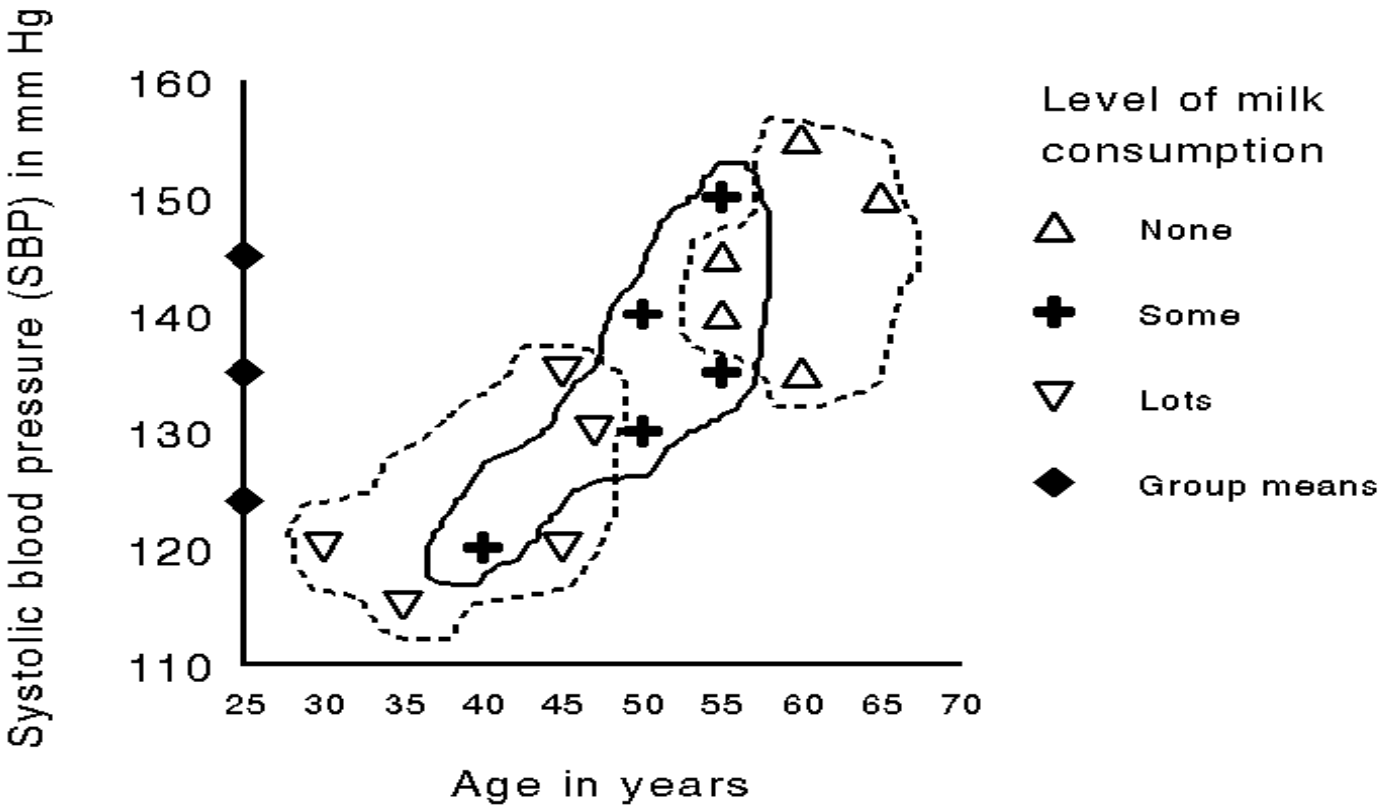
Suppose we now take the relationship between age and milk consumption into account as shown in the following graph:



Now, for each milk consumption group, let's fit a simple linear regression line; i.e.,

$$Y_I = a_I + b_I X_I + \text{error}$$

Where I = 1 for non drinkers, I = 2 for some, and I = 3 for lots. Graphically,



Interpretation: For a man in the *Ith* group of milk consumption who is age X', we predict his SBP to be

$$Y' = a_I + b_I X'$$

Note: Looking at the above picture, it seems hard to argue for different b_I (for both practical and statistical reasons - of course, we could always test to see if the slopes were significantly different); i.e.; it becomes easier (and for purposes of interpretation) to fit

$$Y_I = a_I + bX_I$$

Here, we assume all lines have the same slope ($b = b_1 = b_2 = b_3$).

Based on this model, for men who are the same age, say X', our best estimate of SBP according to amounts of milk consumed are given as follows:

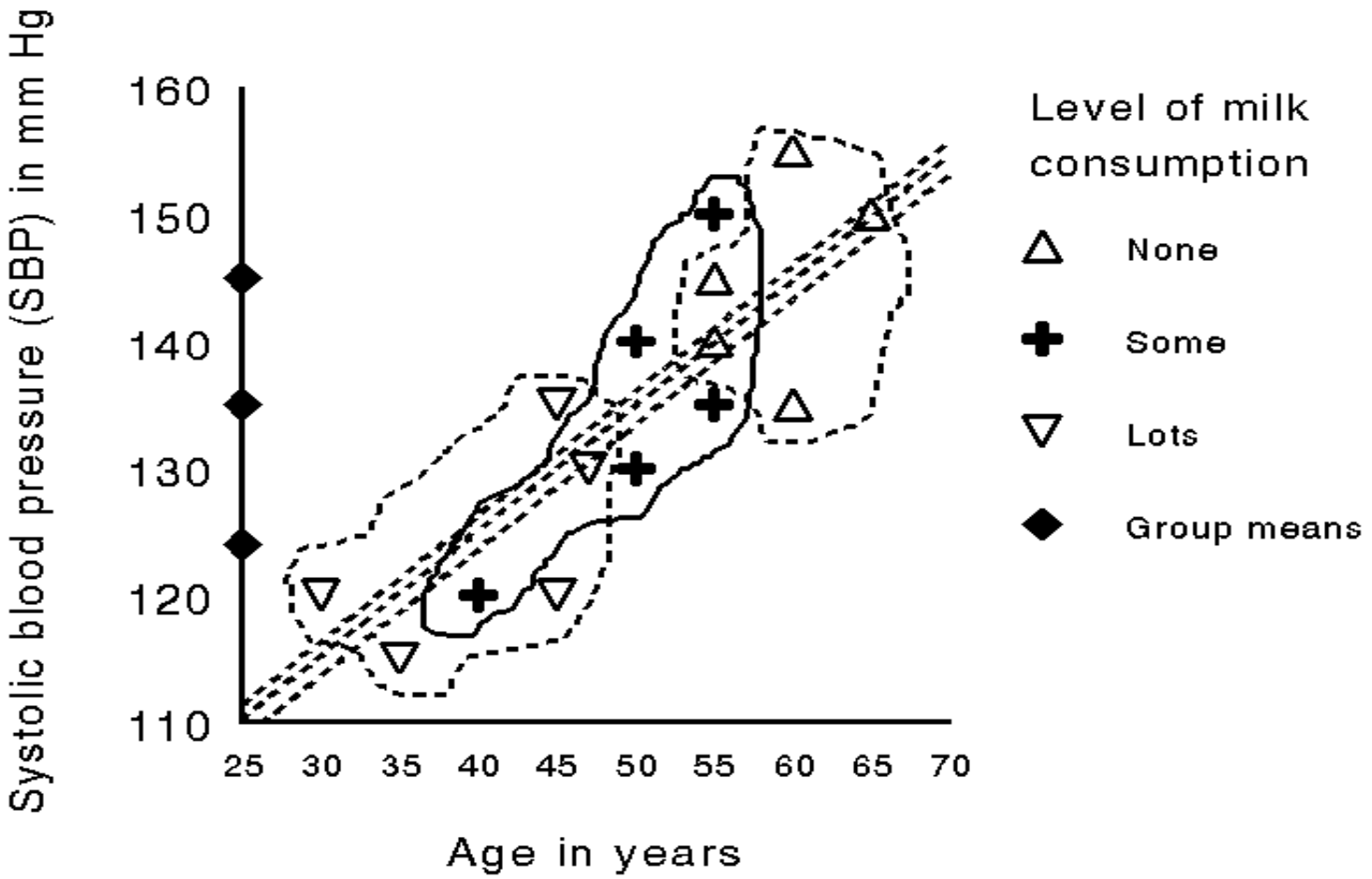
Amount consumed	Best estimate of SBP
None	$a_1 + bX'$
Some	$a_2 + bX'$
Lots	$a_3 + bX'$

Consequence: The difference in SBP between men of the same age who consume different amounts of milk is simply the difference between two intercepts.

For example, for men who are age X' = 50 (or any age for that matter), who consume no milk and lots of milk, the mean difference in SBP is

$$(a_1 + b50) - (a_3 + b50) = a_1 - a_3$$

Graphically, this is just the difference between the relative heights of the regression lines;



For our milk data, we have the following regression models:

Amount consumed	Best estimate of SBP
None	86.825 + 0.986(X')
Some	85.699 + 0.986(X')
Lots	84.165 + 0.986(X')

So, for the average aged man in our sample (X' = 50), we have the following estimates of SBP according to milk consumption levels.

Amount consumed	Best estimate of SBP
None	135.9
Some	134.8
Lots	133.3

Note: These new estimates are referred to as adjusted estimates (adjusted to the average age of the men in our sample). Note further: For different ages, adjusted estimates also differ, but the relative differences in SBP between milk consumption groups remain constant.

Summary of means comparisons

Milk strata	Unadjusted	Adjusted (to the average age)
None	145	135.9
Some	135	134.8
Lots	124	133.3

Notice that group differences have declined from about 10 mm Hg to about 1 to 1.5 mm Hg between adjacent milk consumption groups.

Note #1: This procedure is referred to as *Analysis of Covariance* where we adjust the usual *Analysis of Variance* with a covariate (age).

Note #2: Analysis of Covariance is actually a multivariable regression technique.

Definition: A multivariable regression technique is a method where we use two or more factors (milk and age) to predict an outcome (SBP). A univariable technique considers the relationship between one factor and an outcome.

Conclusion from the Analysis of Covariance

Conclusion #1: After adjusting for age, the effect of milk on SBP is no longer statistically significant.

Conclusion #2: The effect of milk on SBP appears to be derived through its association with age.

Least Squares Multivariable Regression

Multivariable regression by the method of least squares is an extension of the least squares simple linear regression model. The multivariate regression methods:

- Allow the interrelationship between the response and several independent variables to be evaluated simultaneously.

- Allow non-linear relationships between the response variable and the independent variables to be evaluated.

- Allow synergistic effects among the independent variables to be evaluated.

The Multivariable Least Squares Regression Equation.

In conventional notation.

$$y_i = \beta_o + \beta_1, x_{i,1} + \beta_2, x_{i,2} + \cdots + \beta_{p-1}, x_{i,p-1} + \varepsilon_i$$

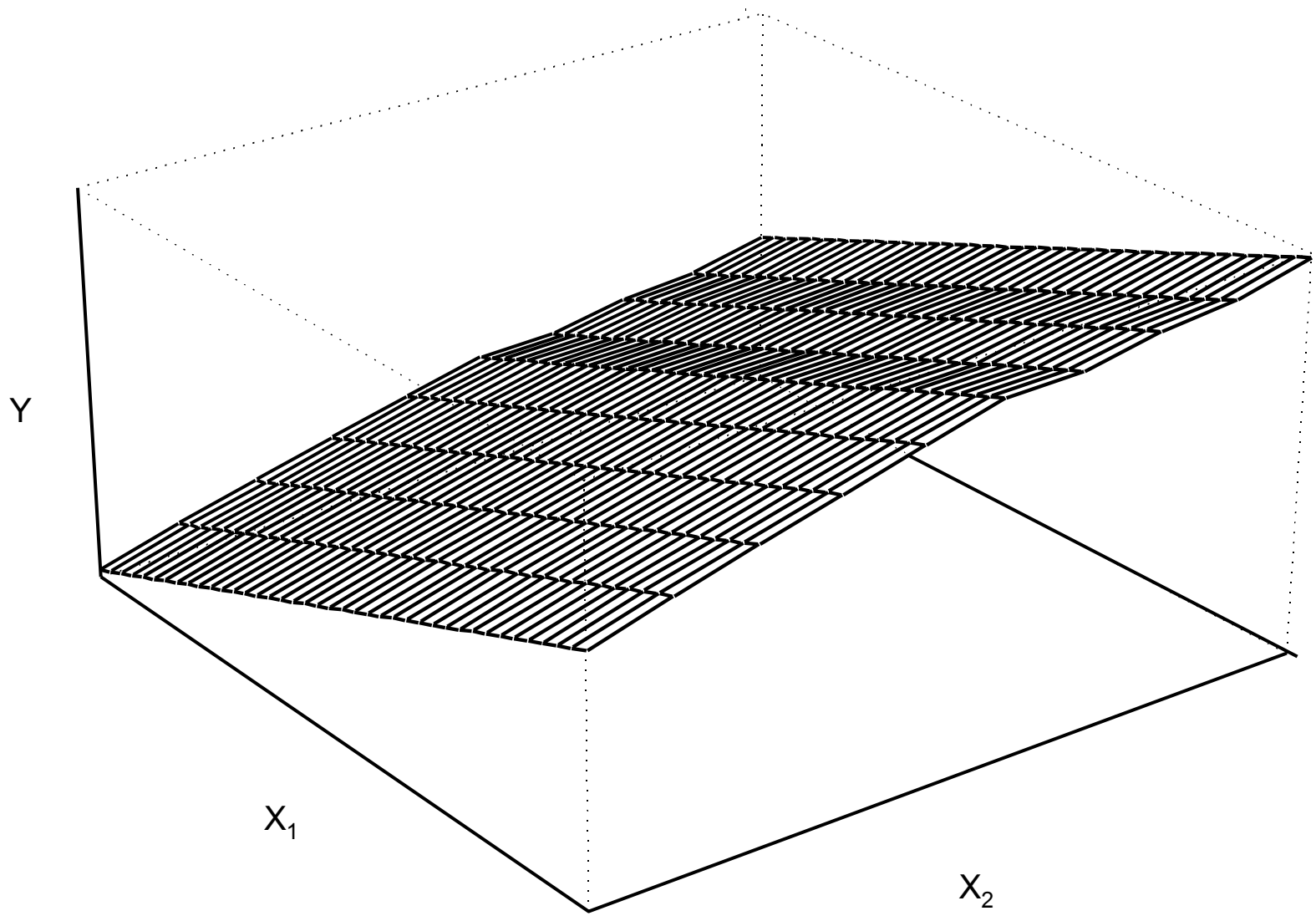
where

y_i is the ith response.

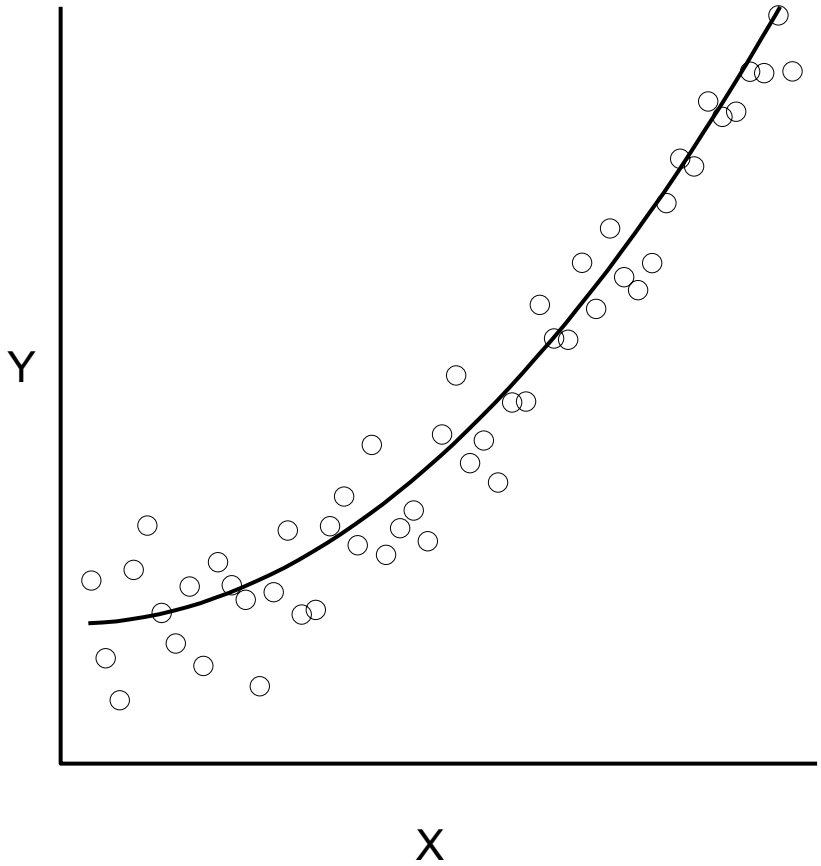
$\beta_o, \beta_1, \cdots \beta_{p-1}$ are the regression parameters.

$x_{i,1}, x_{i,2}, \cdots x_{i,p-1}$ are the ith individual's set of predictors.

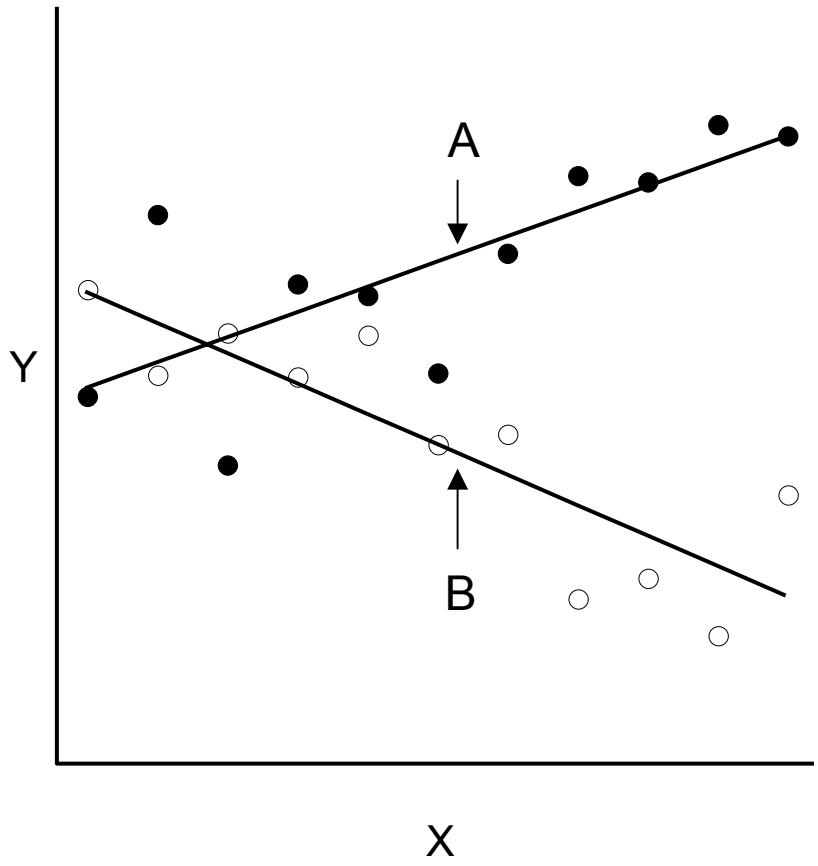
ε_i is the independent random error associated with the ith response, typically assumed to be distributed $N(0,\sigma)$.


Predicted Response Surface
Y
X_1
X_2

Non-Linear Trends



Interaction



Case Study

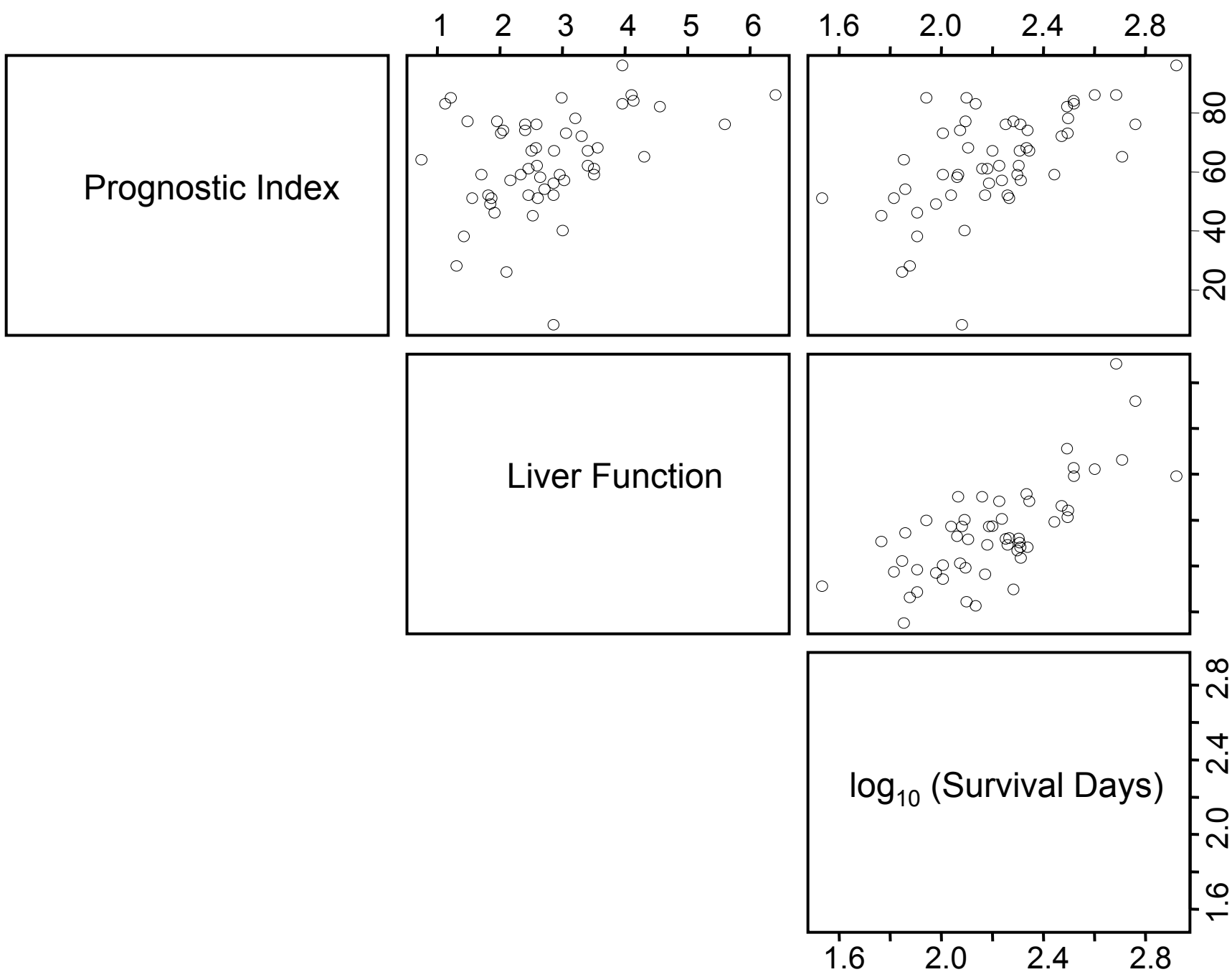
A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random sample of 54 patients was available. From each patient, the following information was extracted from the patients pre-operative records: 1) blood clotting score, 2) prognostic index, which includes the age of the patient, 3) enzyme function test score and 4) liver function test score. The goal of the study was to determine which of these factors were important in predicting survival.

Neter et al. (1996).

Liver Surgery Survival Data.

Table 1. Prognostic index and liver function and post-operative survival.

Subject	Prognostic Index	Liver Function Score	Survival Days	Survival $\log_{10}$
1	62	2.59	200	2.30
2	59	1.70	101	2.00
3	57	2.16	204	2.31
4	73	2.01	101	2.00
5	65	4.30	509	2.71
6	38	1.42	80	1.90
7	46	1.91	80	1.90
8	68	2.57	127	2.10
9	67	2.50	202	2.31
.	.	.	.	.
.	.	.	.	.
54	78	3.20	313	2.50


Prognostic Index
Liver Function
log10 (Survival Days)
1
2
3
4
5
6
1.6
2.0
2.4
2.8
20
40
60
80

Least Squares Model.

$$E(\log(\text{Survival}_i)|x_i) = \beta_o + \beta_1(\text{Prognostic Index}_i) + \beta_2(\text{Liver Fun}_i)$$

Parameter	Estimate
Intercept	1.481
Prog. Index.	0.006
Liver Fun.	0.150

Regression Equation

$$E(\log(\text{Surv})_i|x_i)=1.481 - 0.006(\text{Prog. Index}_i) + 0.150(\text{Liver Fun}_i)$$

Tests of Statistical Inference.

Table 2. Global test of no association between Y and X.

Source	df	SS	MS	F_{obs}	$P(F>F_{obs})$
Regression	2	2.581	1.290	47.269	<0.001
Error	51	1.392	0.027		
Total	53	3.973			

Table 3. Individual tests of no association between Y and x_j.

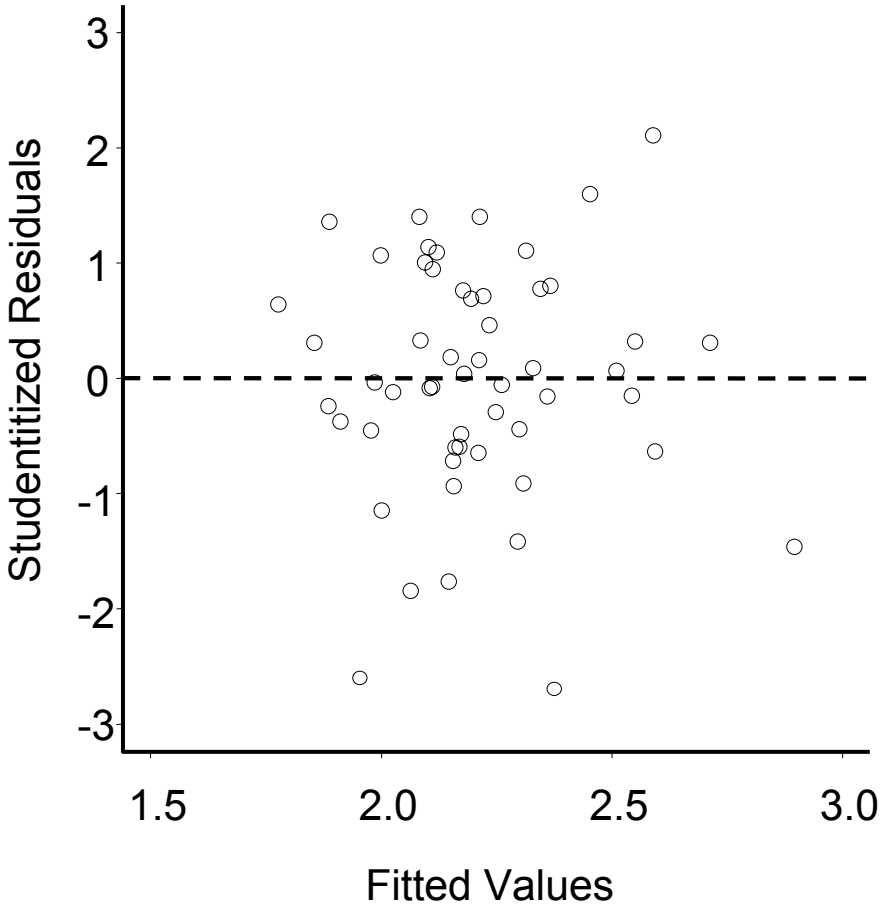
Parameter	Estimate b	SE b	t_{obs}	$P(T>t_{obs})$
Intercept	1.389	0.084		
Prognostic Index	0.006	0.001	5.089	<0.001
Liver Function	0.139	0.022	6.303	<0.001

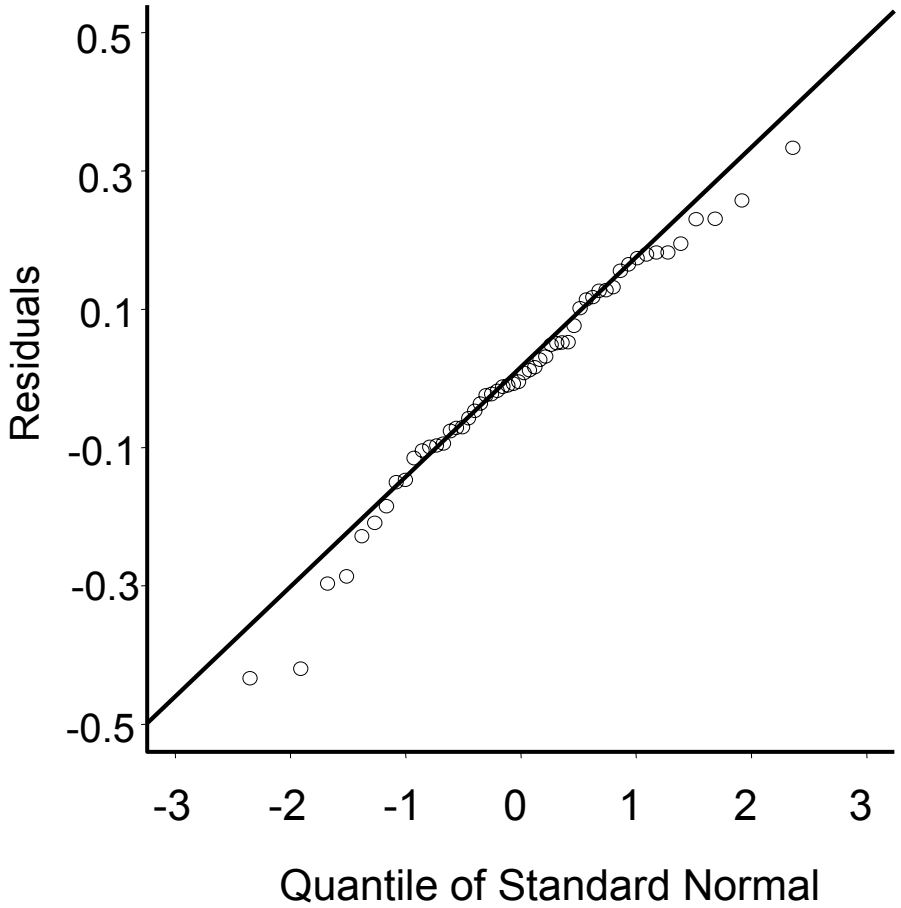
Confidence Intervals for the β_j (s).

Table 4. 95% confidence intervals of the regression parameters

Parameter	Estimate b	SE b	df	$t_{(51,0.95)}$	Lower 95%CL	Upper 95%CL
Intercept	1.481	0.092				
Prog Index	0.006	0.001	51	2.00	0.004	0.008
Liver Fun.	0.150	0.023	51	2.00	0.104	0.196

Residual Diagnostics





Regression Analysis Summary

The patient's pre-operative prognostic index value ($p<0.001$) and the patient's pre-operative liver function score ($p<0.001$) were determined to be positively associated with the patient's log transformed post-operative survival time.

The effect of a one unit increase in the pre-operative prognostic index was to increase the patient's post-operative survival time on the log scale by 0.006 units [95%CL(0.004,0.008)], while the effect of a one unit increase in the pre-operative liver function score, was to increase the patient's post-operative survival time on the log scale by 0.150 units [95%CL(0.104, 0.196)].

Case Study

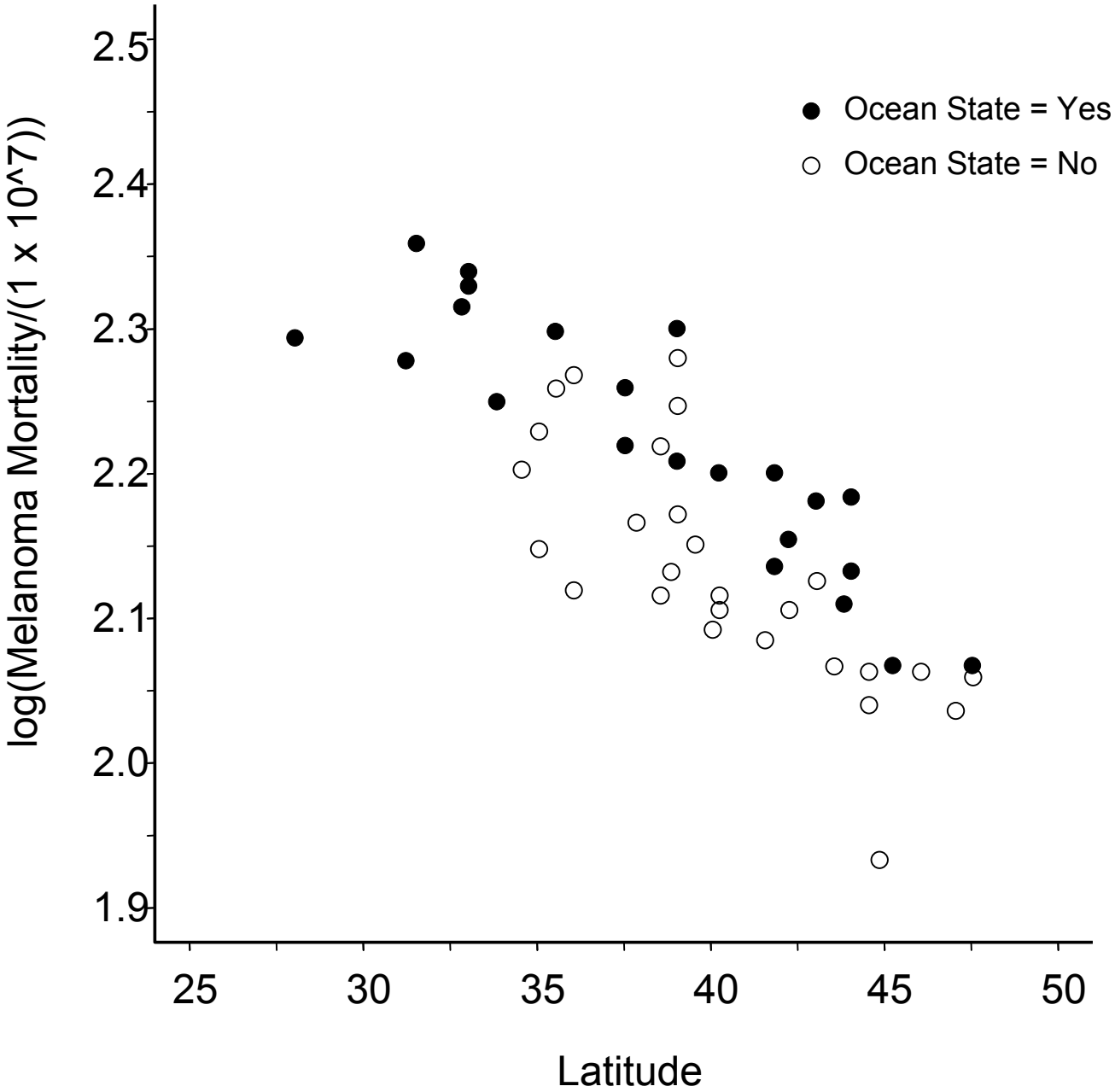
Data on the mortality due to malignant melanoma of the skin of white males were collected during the period of 1950-1969 from each state in the United States as well as the District of Columbia. No mortality data were available for Alaska and Hawaii for this period. The goal of the study was to assess whether the incidence of melanoma was related to the states' latitude and the states' proximity to an ocean (ocean state; yes or no).

Fisher et al. (1993)

Melanoma Data.

Table 1. Melanoma mortality data from 48 states in US.

State	Mortality (Per 1 x 10^7)	Latitude (Degrees)	Ocean State
Alabama	219.0	33.0	yes
Arizona	160.0	34.5	no
Arkansas	170.0	35.0	no
California	182.0	37.5	yes
Colorado	149.0	39.0	no
Connecticut	159.0	41.8	yes
Delaware	200.0	39.0	yes
DC	177.0	39.0	no
Florida	197.0	28.0	yes
	.	.	.
	.	.	.
Wyoming	134.0	43.0	no


log(Melanoma Mortality/(1 x 10^7))
2.5
2.4
2.3
2.2
2.1
2.0
1.9
25
30
35
40
45
50
Latitude
Ocean State = Yes
Ocean State = No

Least Squares Separate Slopes Model.

$$E(\log(\text{Mortality}_i)|x_i) = \beta_o + \beta_1(\text{Latitude}_i) + \beta_2(z_{i,2}) + \beta_3(\text{Latitude}_i * z_{i,2})$$

where $z_{i,2}$ = 1 if Ocean State=yes, else $z_{i,2}$ =0

Parameter	Estimate
Intercept	-2.796
Ocean State =yes	-0.016
Latitude	-0.021
Latitude x Ocean State = yes	-0.002

Regression Equations

$$E(\text{Mortality}_i|\text{Ocean State=no}) = -2.796 - 0.021(\text{Latitude}_i)$$
$$E(\text{Mortality}_i|\text{Ocean State=yes}) = -2.812 - 0.019(\text{Latitude}_i)$$

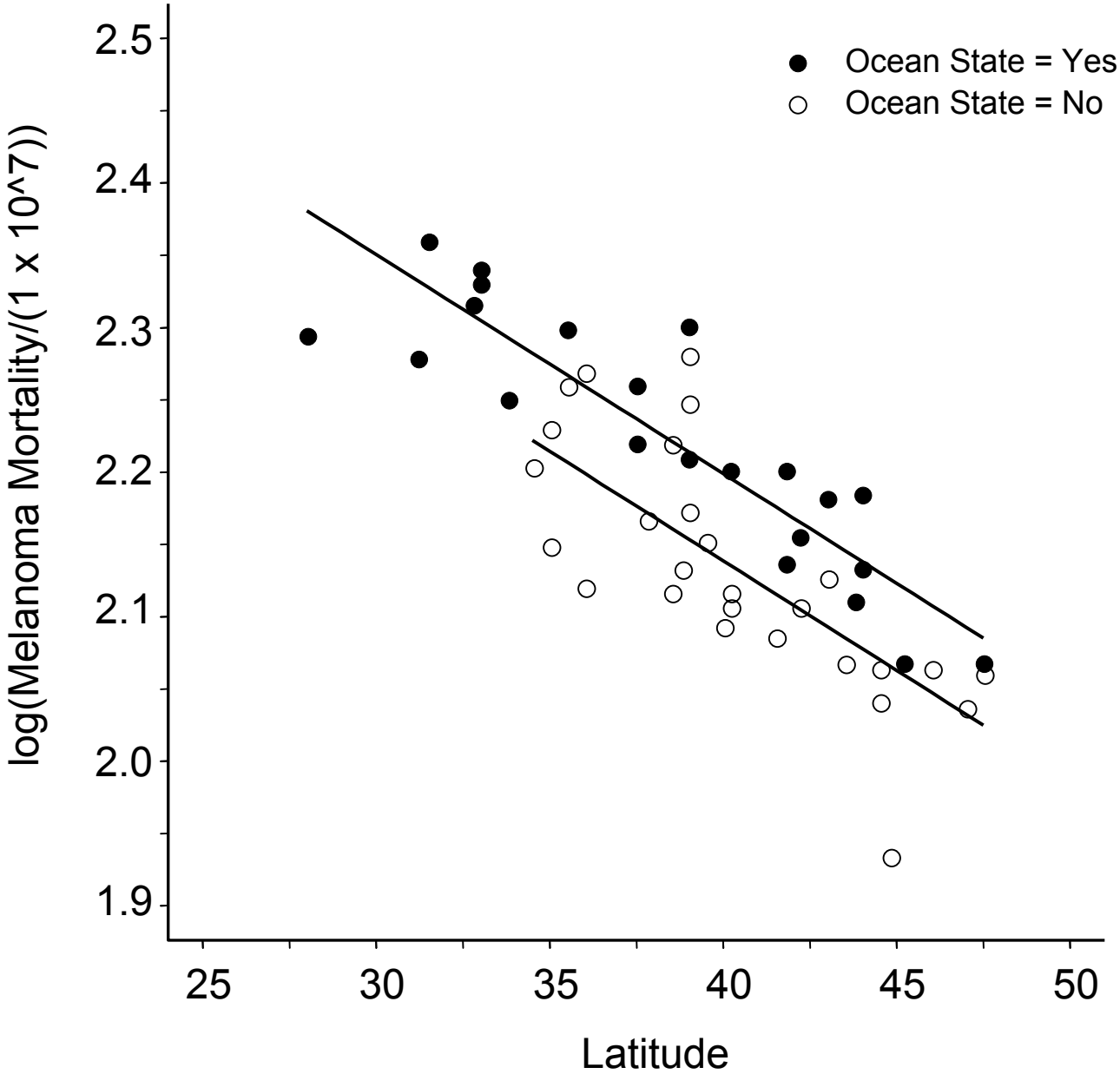
Tests of Statistical Inference for the Separate Slopes Model.

Table 2. Global test of no association between Y and X.

Source	df	SS	MS	F_{obs}	$P(F>F_{obs})$
Regression	3	0.322	0.107	44.757	<0.001
Error	45	0.108	0.002		
Total	48	0.431			

Table 3. Individual tests of no association between Y and x_j.

Parameter	Estimate	SE	t_{obs}	$P(T>t_{obs})$
Intercept	2.796	0.102		
Ocean=yes	-0.021	0.128	-0.166	0.868
Latitude	-0.016	0.002	-6.517	<0.001
L x O=yes	0.002	0.003	0.642	**0.524**


Ocean State = Yes
Ocean State = No
log(Melanoma Mortality/(1 x 10^7))
2.5
2.4
2.3
2.2
2.1
2.0
1.9
25
30
35
40
45
50
Latitude

Least Squares Common Slope Model

$$E(\log(\text{Mortality}_i)|x_i) = \beta_o + \beta_1(\text{Latitude}_i) + \beta_2(z_{i,2})$$

where $z_{i,2}$ = 1 if Ocean State=yes, else $z_{i,2}$ =0

Parameter	Estimate
Intercept	2.745
Ocean State (=yes)	0.060
Latitude	-0.015

Regression Equations

$$E(\log \text{Mortality}_i|\text{Ocean State=no}) = 2.745 - 0.015(\text{Latitude}_i)$$
$$E(\log \text{Mortality}_i|\text{Ocean State=yes}) = 2.805 - 0.015(\text{Latitude}_i)$$

Tests of Statistical Inference for the Common Slope Model

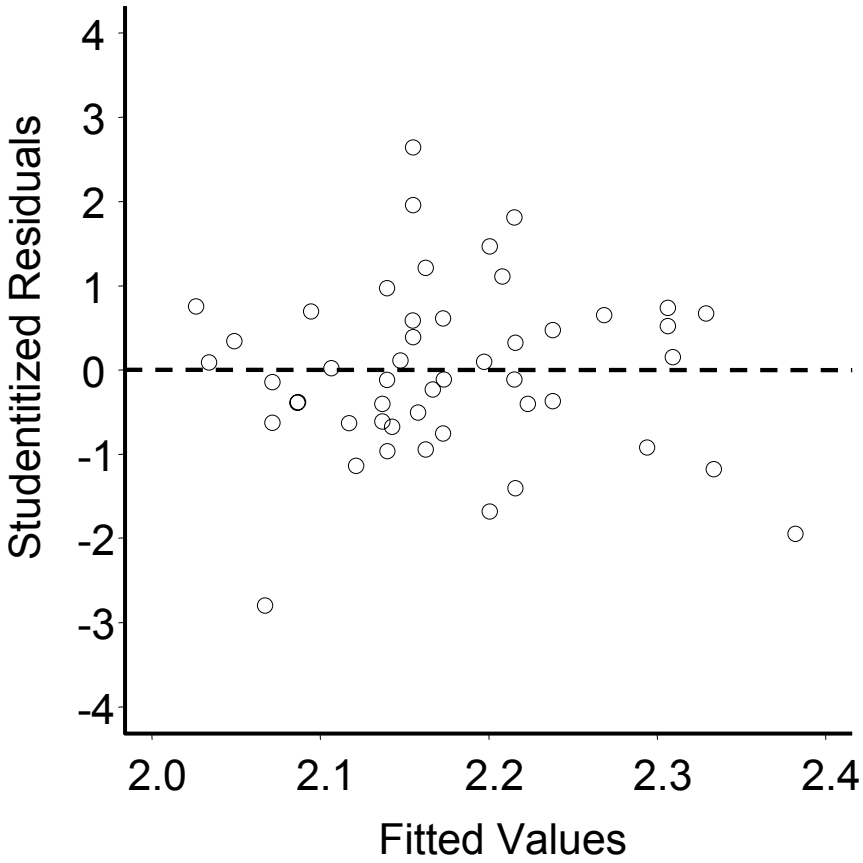
Table 4. Global test of no association between Y and X.

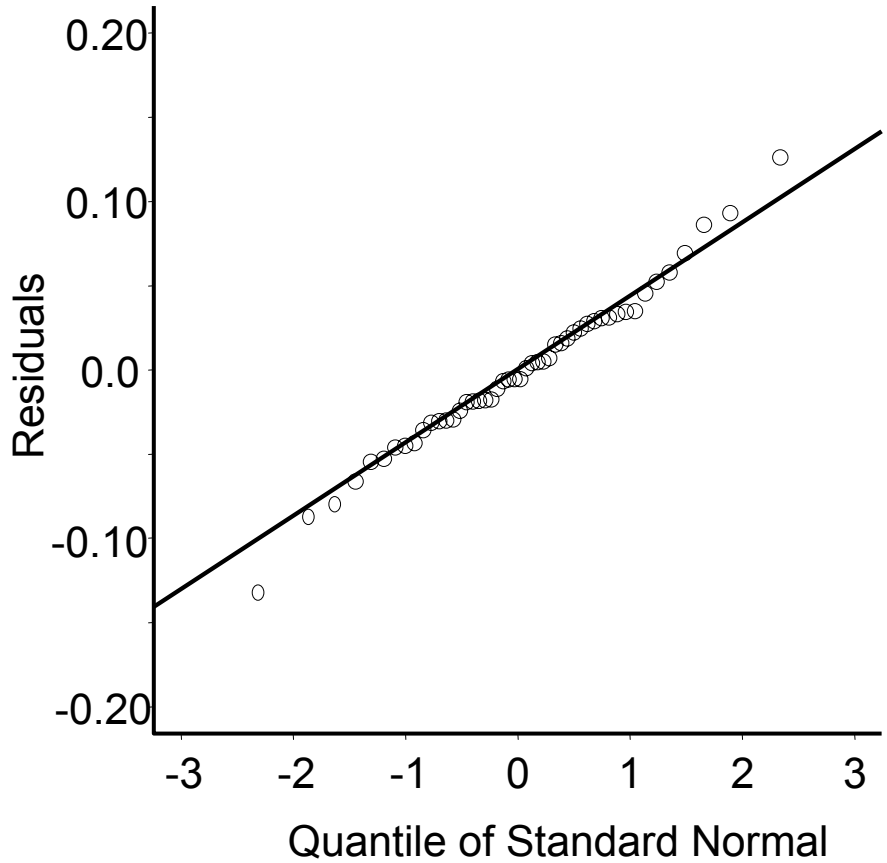
Source	df	SS	MS	F_{obs}	$P(F>F_{obs})$
Regression	2	0.323	0.161	67.794	<0.001
Error	46	0.109	0.002		
Total	48	0.431			

Table 5. Individual tests of no association between Y and x_j.

Parameter	Estimate	SE	t_{obs}	$P(T>t_{obs})$
Intercept	2.745	0.0630		
Ocean=yes	0.060	0.0143	4.222	<0.001
Latitude	-0.015	0.0014	-9.793	<0.001

Residual Diagnostics





Regression Analysis Summary

Melanoma mortality on the log scale was negatively related to latitude ($p<0.001$) and was higher among those states that bordered an ocean ($p<0.001$). The predicted melanoma mortality increased by 0.15 [95%CL(0.12, 0.18)] on the logarithmic scale for each 10 degree reduction in latitude. Melanoma mortality increased by 0.60 [95%CL(0.031,0.088)] units on the logarithmic scale for those states bordering and ocean.

for each 10 degree reduction in latitude, the predicted melanoma mortality is increased by a factor of exp(0.15) or 1.16(1.13,1.20). Melanoma mortality increased by a factor of exp(0.06) or 1.06(1.03, 1.09) for those states bordering an ocean.

Thank You !