# Comprehensive Introduction to Clinical Investigation Biostatistics
# Session # 5

## “Regression and Correlation”

Jim Patrie, MS

Division of Biostatistics and Epidemiology

July 17, 2001

# 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.

I) 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.

## II) Different Types of Regression.

There are several different types of regression methods:

- General linear regression.
- Non-linear regression.
- Robust regression.
- Non-parametric regression.
- Generalized linear regression.
- Parametric and non-parametric survival regression.

# Part I
# Simple Linear Regression

## III) The Simple General Linear Model Setting.

- The data consist of N paired measurements. Each pair consisting of a measurement from a response variable Y, and a measurement from an 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 they can 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

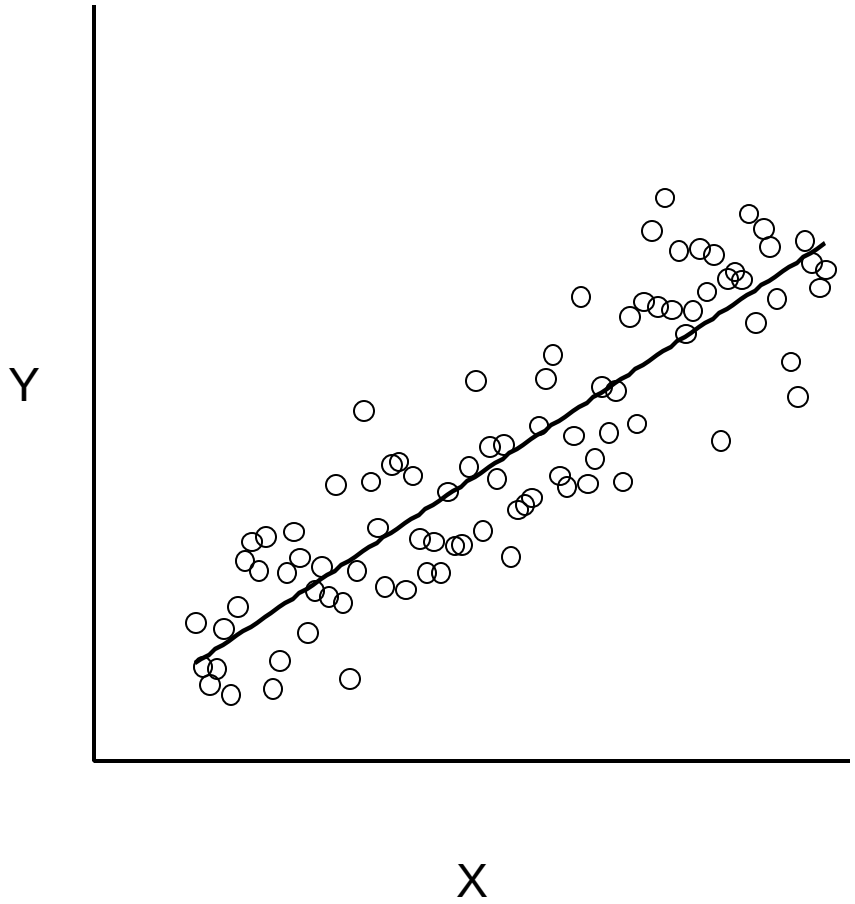
- When the independent variable X has a continuous scale of measure the functional relationship between X and Y is summarized by a simple linear equation in which the expected value of $y_i$ $(i = 1,2\cdots,n)$ is estimated as a linear function of $x_i$.
- When the independent variable X is continuous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

## X-Categorical

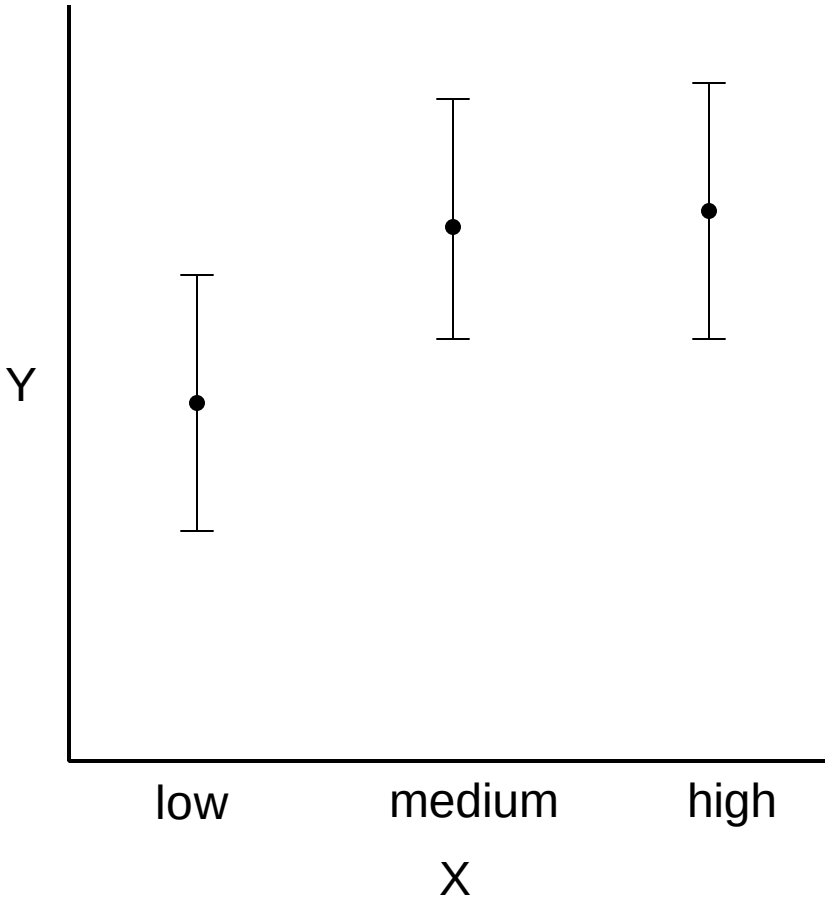
- When the independent variable X is categorical the functional relationship between X and Y is typically summarized by a comparison of the mean of Y across the categories of X.
- When X is categorical the analysis is referred to as “One-Way ANOVA”.

# Simple-Linear Regression and One-Way ANOVA Scenarios

## Simple-Linear Regression




## One-Way ANOVA



## IV) The Simple-Linear Regression Equation.

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

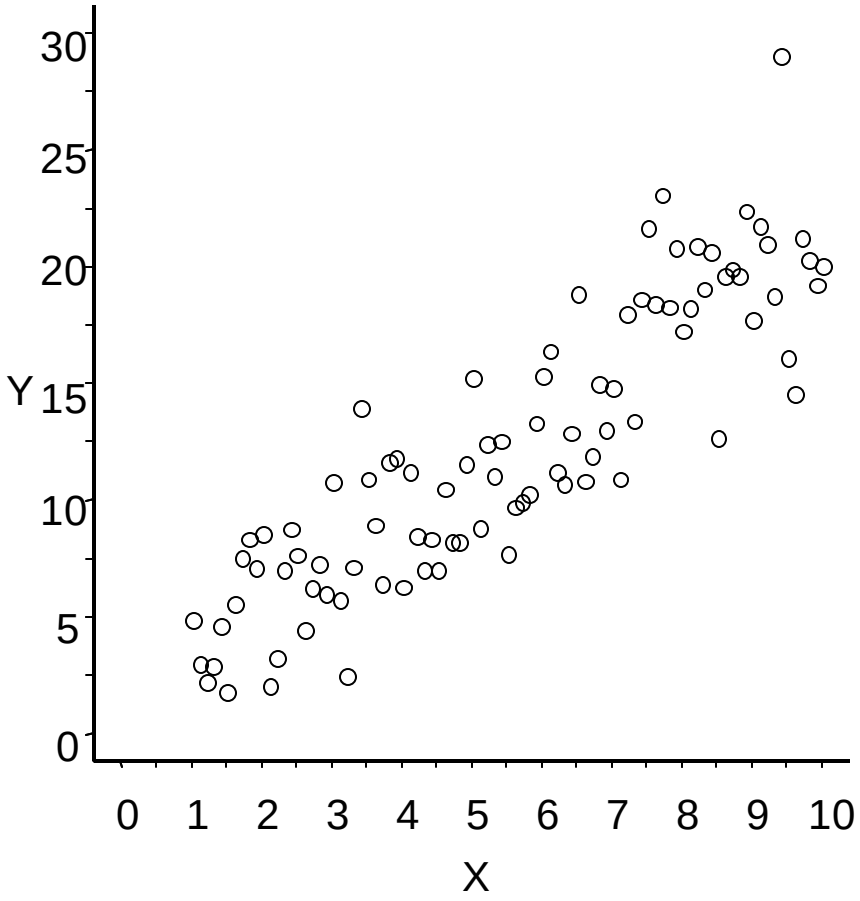
$$(i = 1, 2, \cdots, n)$$

where

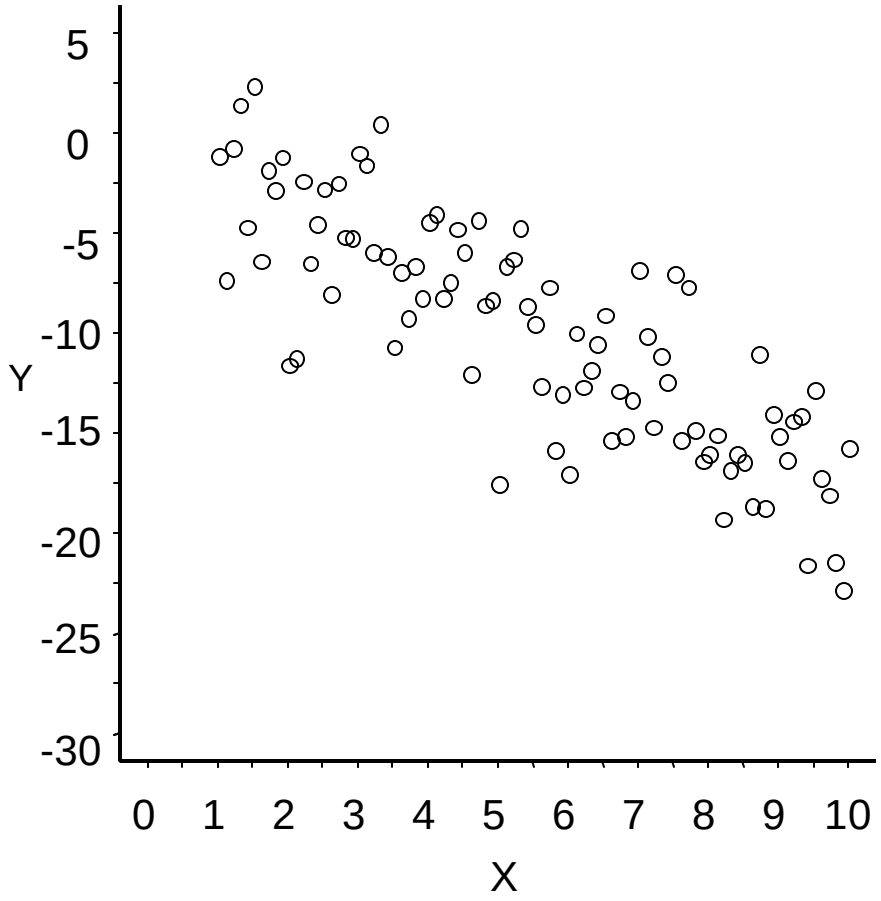
$y_i$ = the i*th* value of the response.
$x_i$ = the i*th* value of the independent variable.
$\alpha$ = the intercept parameter (y intercept).
$\beta$ = the slope parameter ($\Delta y/\Delta x$).
$e_i$ =the random error associated with the i*th* response value.

The $e_i$(s) are assumed to be independent identically distributed normal random variables with mean 0 and variance $\sigma^2$.

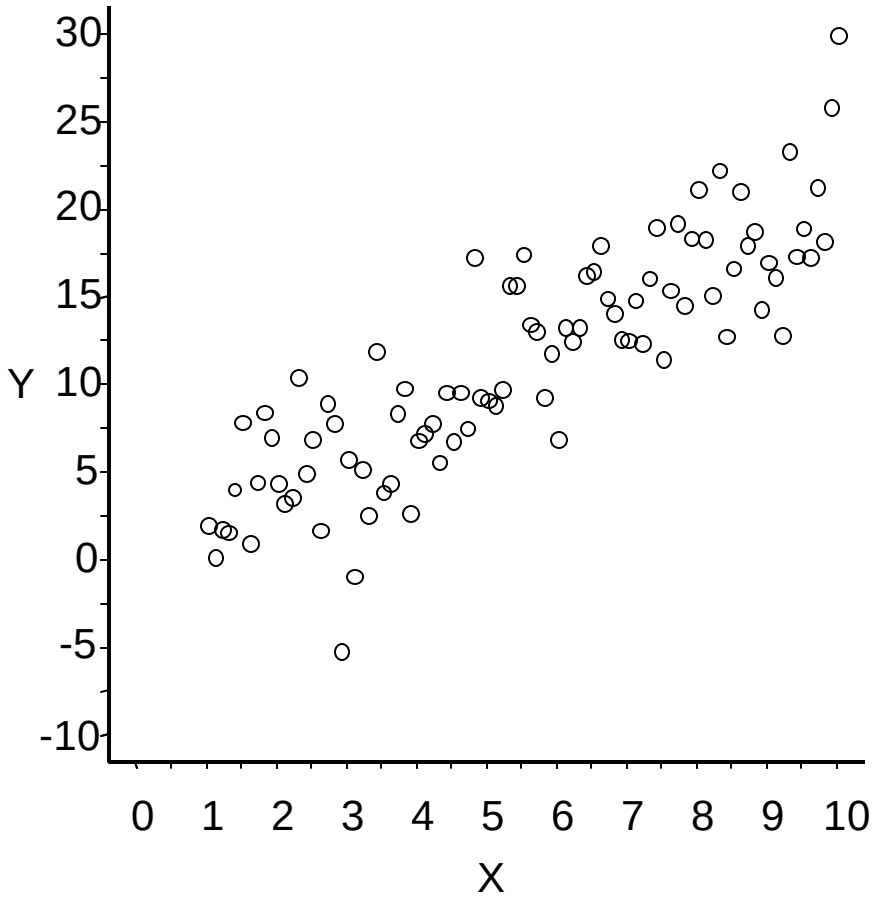
# Examples of the Assumed Underlying Data Generating Process



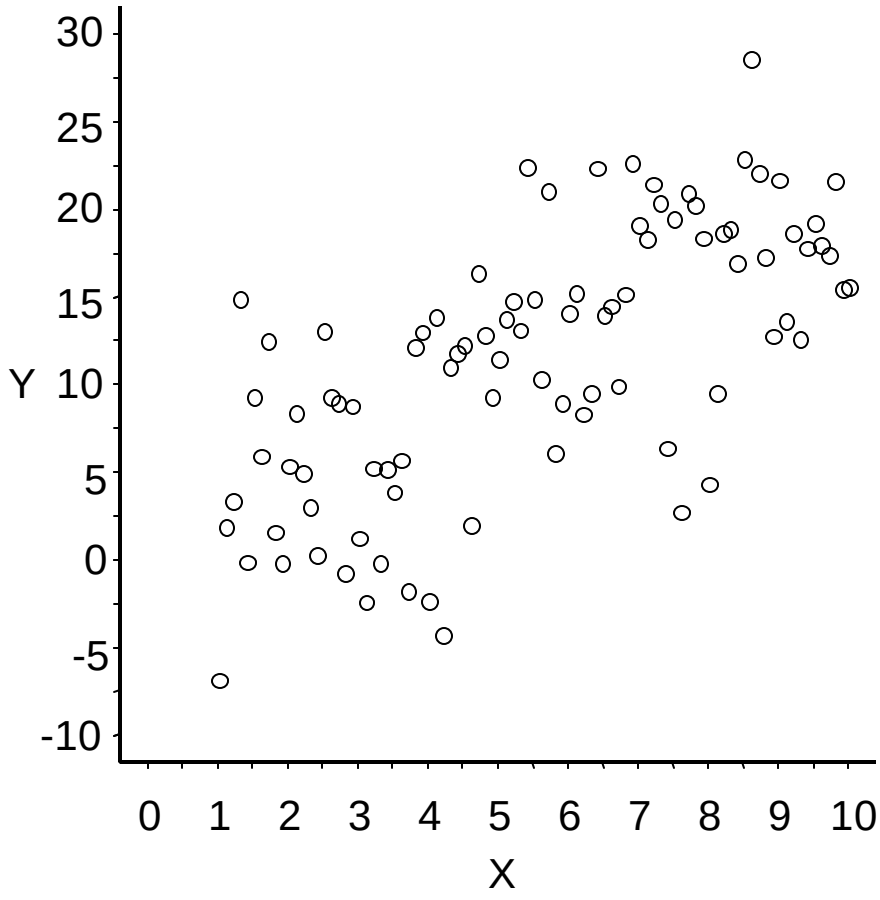

$y = 1 + 2x + \varepsilon$ where $\varepsilon \sim N(0,3)$




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




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




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

V) 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 $x_i$.

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

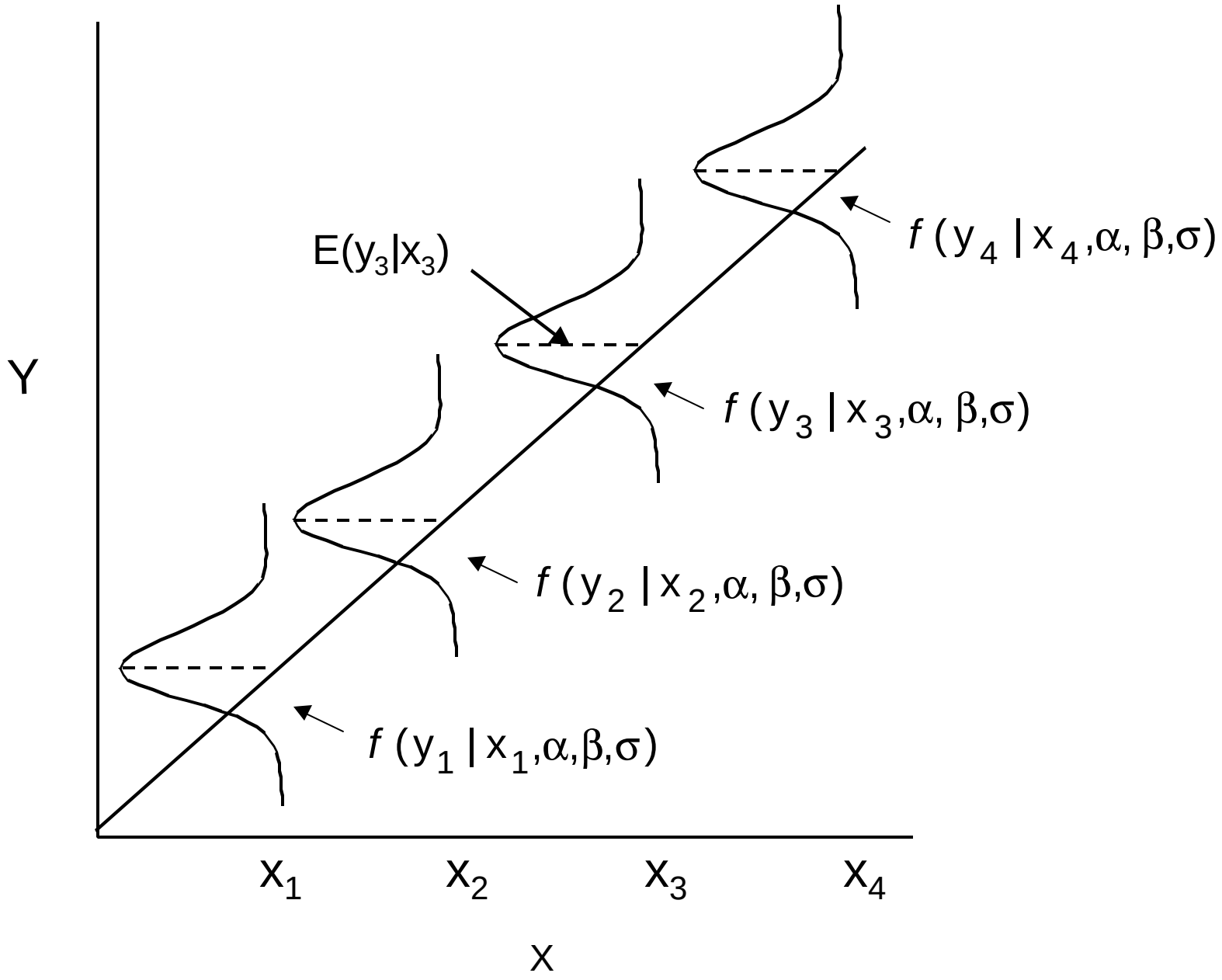
$\alpha$ = the intercept parameter (y intercept).

$\beta$ = the slope parameter ($\Delta$ y/$\Delta$x).

VI) 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 $\sigma^2$.
- The $y_i$ (*s)* are independent.

# Model Assumptions



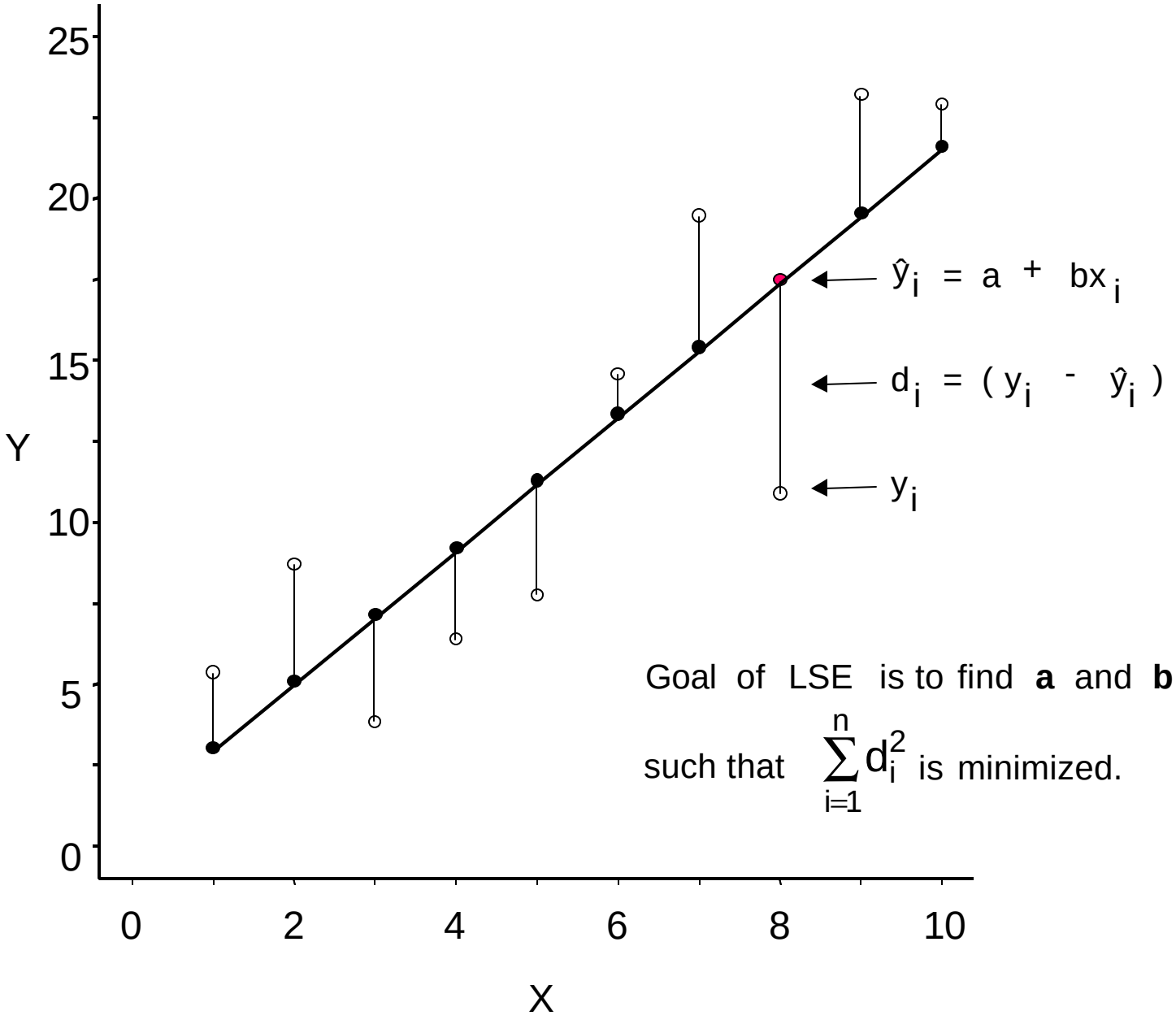
## VII) Least Squares Parameter Estimation.

### a) The Goal of Least Squares Estimation

- The goal of least squares parameter estimation is to estimate the value of $\alpha$ and $\beta$ 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.

b) The Properties of the Least Squares Estimators.

When the least squares simple-linear regression model assumptions are valid the estimators for $\alpha$ and $\beta$ have the following mathematical properties:

- The estimators are unbias.

- The estimators have uniformly minimum variance among all unbias estimators of $\alpha$ and $\beta$.

c) Estimating $\alpha$ and $\beta$ 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 â.

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

d) 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 $\alpha$ and $\beta$. 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.

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

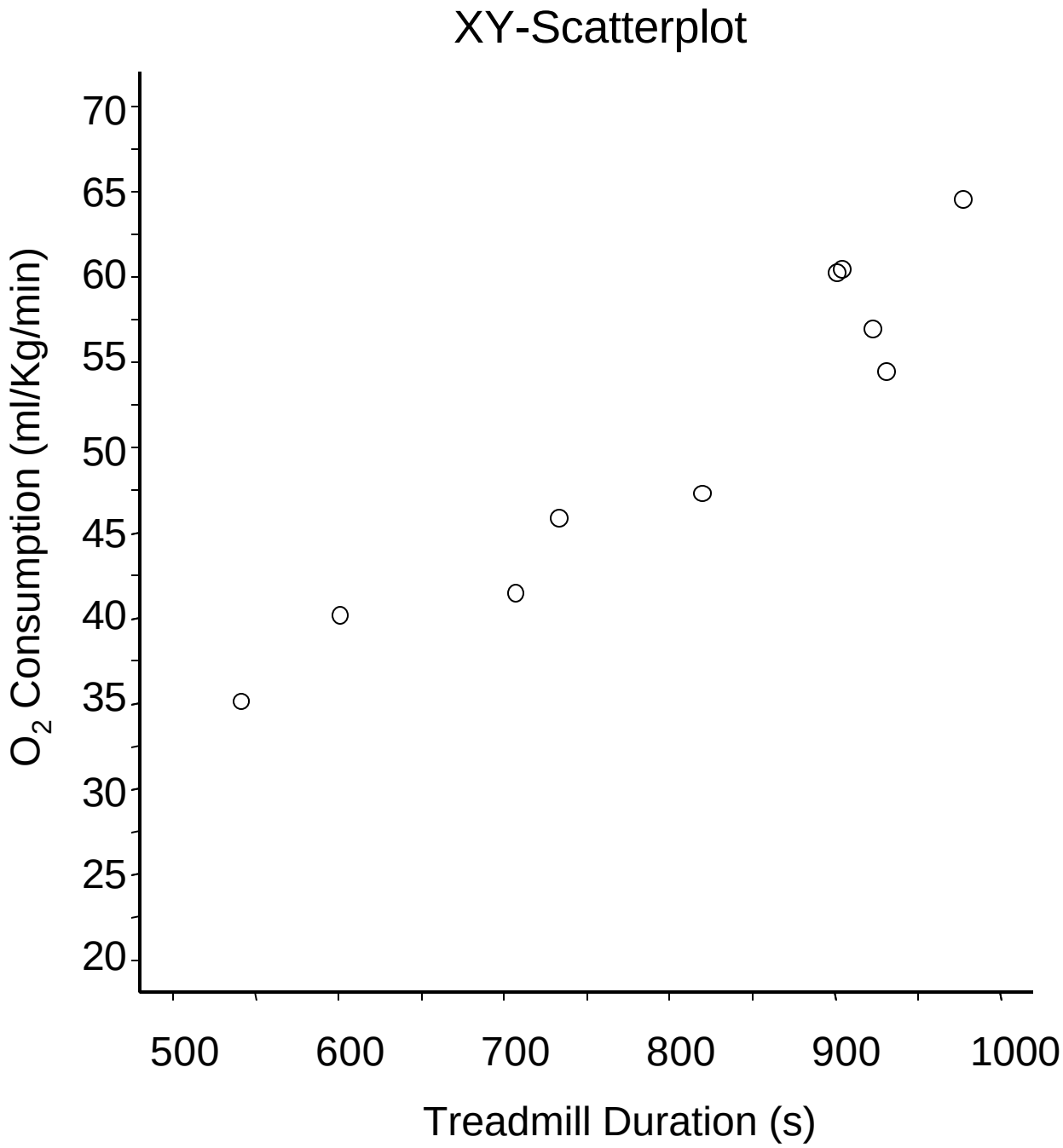
$$SE(a) = \sqrt{s^2_{y.x}\left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right)} \quad \text{where} \quad s^2_{y.x} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n - p)$$

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

# VIII) Example: Treadmill Exercise Data.

Exercise Data from 10 Healthy Active Males.

| Treadmill Duration (s) | $VO_2$ MAX (ml/kg/min) |
|---|---|
| 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
70
65
60
55
50
45
40
35
30
25
20
O2 Consumption (ml/Kg/min)
500
600
700
800
900
1000
Treadmill Duration (s)

a) The Least Squares Estimates for $\alpha$ and $\beta$.

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

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

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

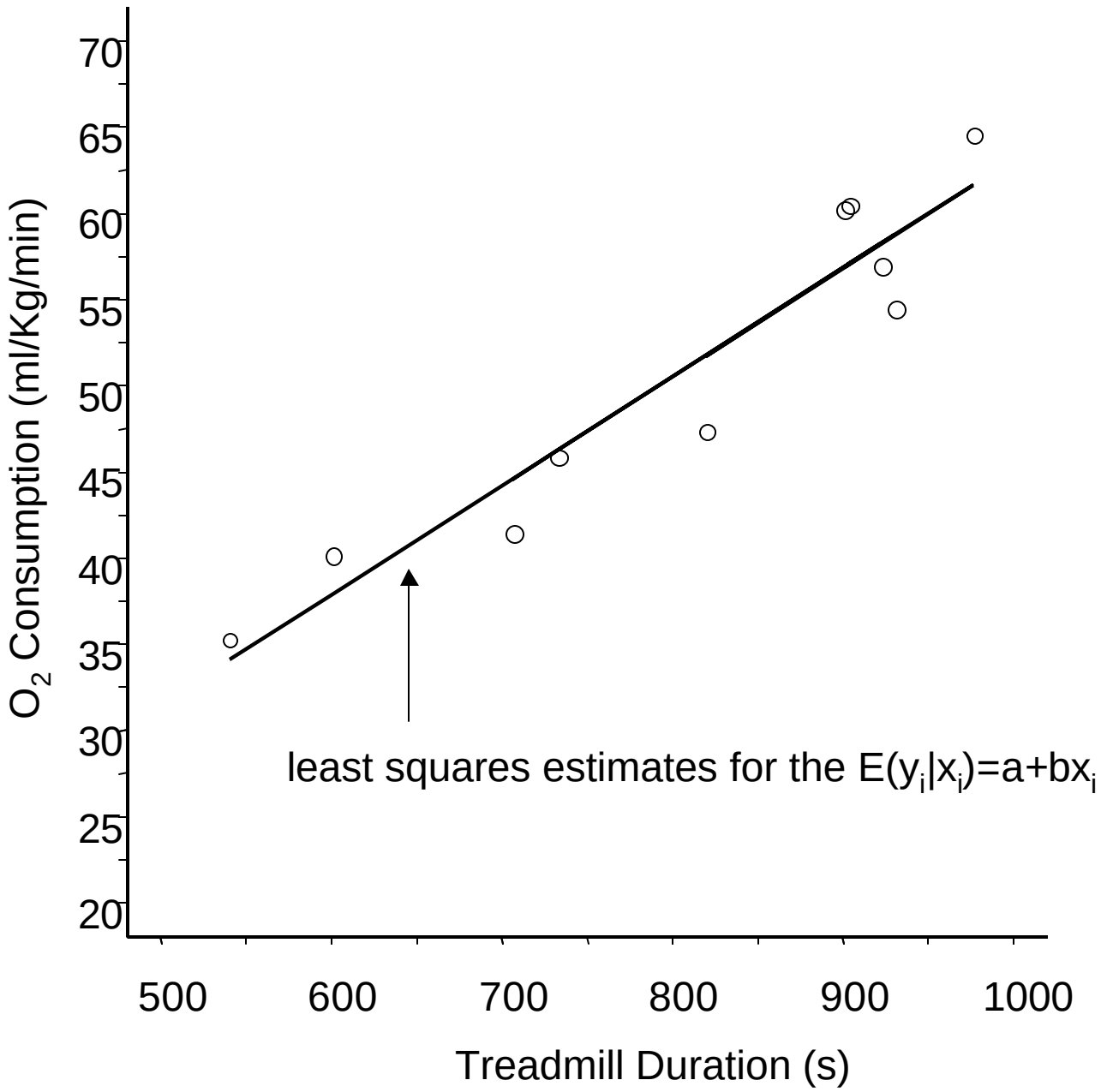
$$L_{xy} = (706 - 802.8)(41.5 - 50.7) + \cdots + (540 - 802.8)(35.2 - 50.7) = 12936.6$$

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

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

Regression Equation

$$E(y_i \mid x_i) = -0.026 + 0.063 \frac{O_2 \text{ (ml/Kg/min)}}{\text{s}} (\text{treadmill duration}_i \text{ (s)})$$


70
65
60
55
50
45
40
35
30
25
20
$O_2$ Consumption (ml/Kg/min)
500
600
700
800
900
1000
Treadmill Duration (s)
least squares estimates for the $E(y_i|x_i)=a+bx_i$

b) The Least Squares Estimate of SE for a and b.

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

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

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

## IX) 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 ($y_i$) of the outcome and the value ($x_i$) of the independent variable, or equivalently that the value of $\beta$ 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 less consistent with the null hypothesis of no association.

a) Sum of Squares Principle.

- Note that the linear estimator for $E(y_i|x_i)$ can be expressed as:

$$\hat{y}_i = a + bx_i$$
$$= \bar{y} - b\bar{x} + bx_i$$
$$= \bar{y} - b(x_i - \bar{x})$$

- Under the null hypothesis, we would therefore choose $\hat{y}_i = \bar{y}$ as our estimator of the $E(y_i|x_i)$ at all values $x_i$. This leads to the concept of "sum of squares decomposition", the mathematical principle upon which the statistical inference of the least squares regression model is based.

b) Sum of Squares Decomposition.

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

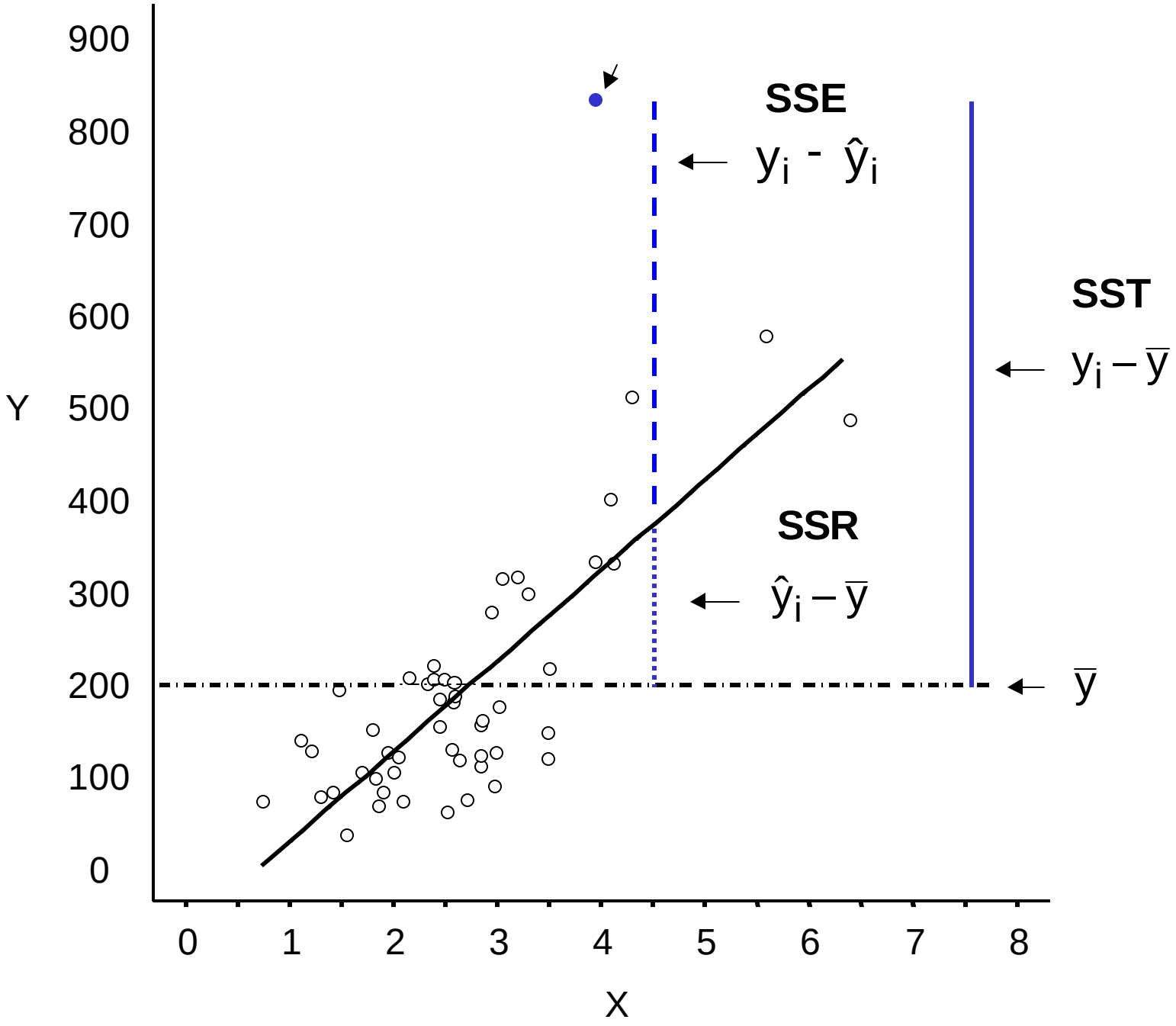
SS Total = SS Regression + SS Error

Where

SST = the sum of the squared difference between the $y_i$(s) and the mean of Y.

SSR = the sum of the squared difference between the predicted values of the $y_i$(s) and the mean of Y.

SSE = the sum of the squared difference between the $y_i$(s) and the predicted values of the $y_i$(s).


900
800
700
600
500
400
300
200
100
0
Y
0
1
2
3
4
5
6
7
8
X
SSE
$y_i - \hat{y}_i$
SST
$y_i - \bar{y}$
SSR
$\hat{y}_i - \bar{y}$
$\bar{y}$

- The criterion for the global assessment of the model's fit is based on a standardized ratio of the regression sum of squares (SSR) to the residual sum of squares (SSE). A large ratio indicates a good fit, whereas a small ratio indicates a poor fit.
- The components of the sum of squares decomposition and the remaining information that is required to assess goodness of fit are presented in tabular form in what is referred to as the regression ANOVA table.

Table 1. Regression ANOVA table.

| Source | SS | DF | MS | F |
|---|---|---|---|---|
| Regression | SSR | p-1 | SSR/(p-1) | MSR/MSE |
| Error | SSE | n-p | SSE/(n-p) | |
| Total | SST | n-1 | | |

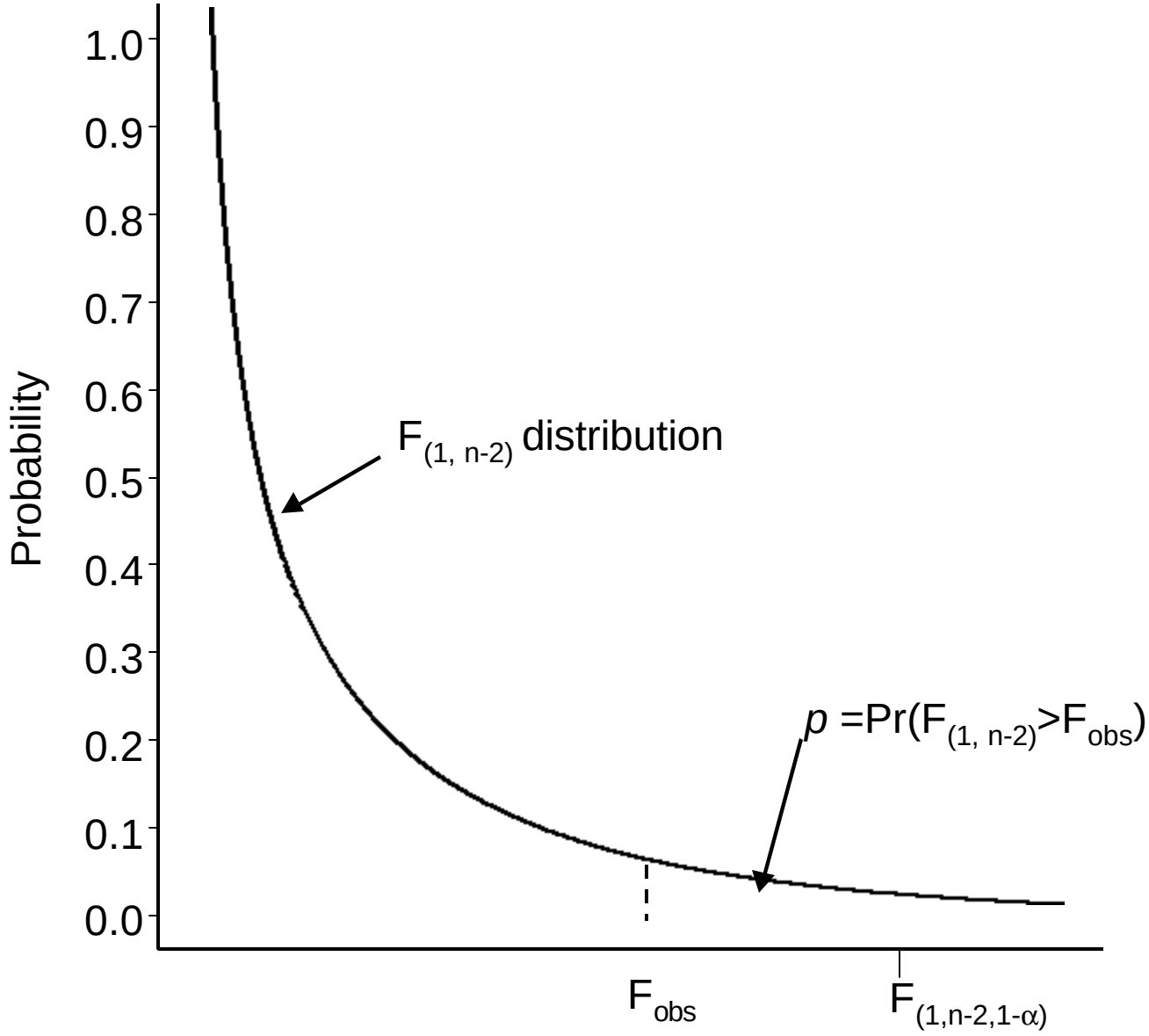
c) F-Test for the Hypothesis of No Association Between Y and X.

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

$$F_{obs} = \frac{MSR}{MSE} = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/(p-1)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p)}$$

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

# F Null-Distribution



d) Example: Regression of $VO_2$ Max onto Duration of Exercise.

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

$$F_{obs} = \frac{MSR}{MSE} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/(p-1)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p)} = \frac{817.5/(2-1)}{87.1/(10-2)} = 75.1$$

For a two-sided test with significance level 0.05

$F_{(1,8,0.95)} = 5.32$

$F_{obs} > F_{(1,8,0.95)}$, which $\Rightarrow$ we should reject Ho: $\beta = 0$.

e) An Equivalent Test for the Hypothesis of No Association.

- We can also use the t-Test as an alternative method of testing the null hypothesis Ho: $\beta = 0$, versus $H_a$: $\beta \neq 0$.

- The estimator b for $\beta$ is a linear function of Y. If Y is normally distributed with mean $\mu$ and variance $\sigma^2$ then it can be shown that the sampling distribution of b will be normal with mean $\beta$, and variance $\sigma^2/L_{xx}$.

- Note that if $\sigma^2$ were known, then under the null hypothesis $b/\sqrt{s^2/L_{xx}}$ would be a standard normal random variable with mean 0 and variance 1.

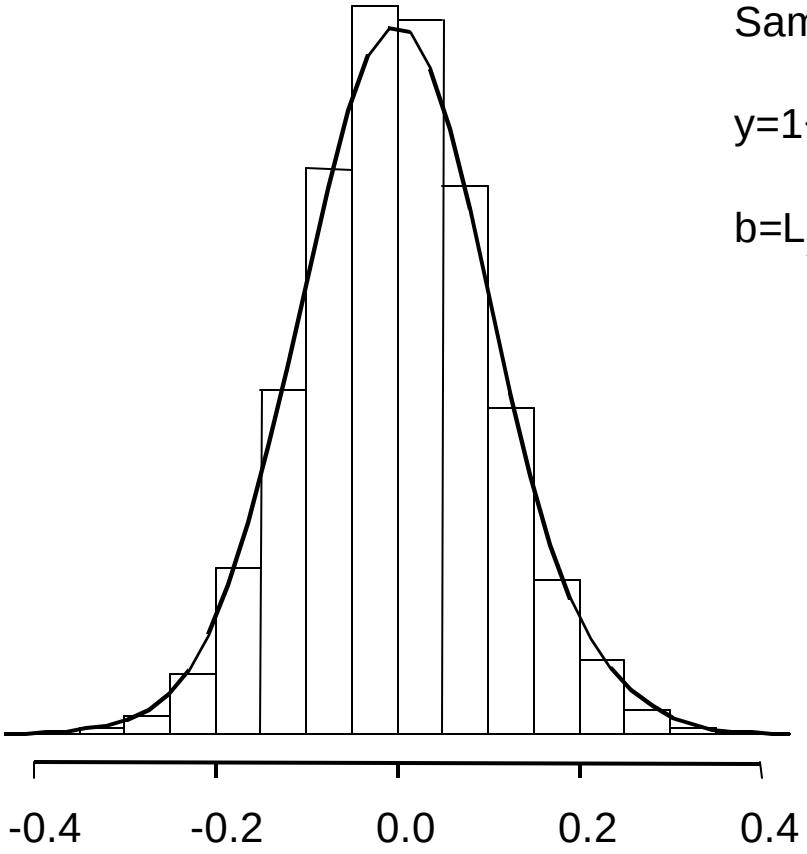
- Since we have to estimate $\sigma^2$ from the data, under the null hypothesis the ratio $b/\sqrt{\hat{s}^2/L_{xx}}$ follows at *t*-distribution with n-p degrees of freedom.

Simulation=5000

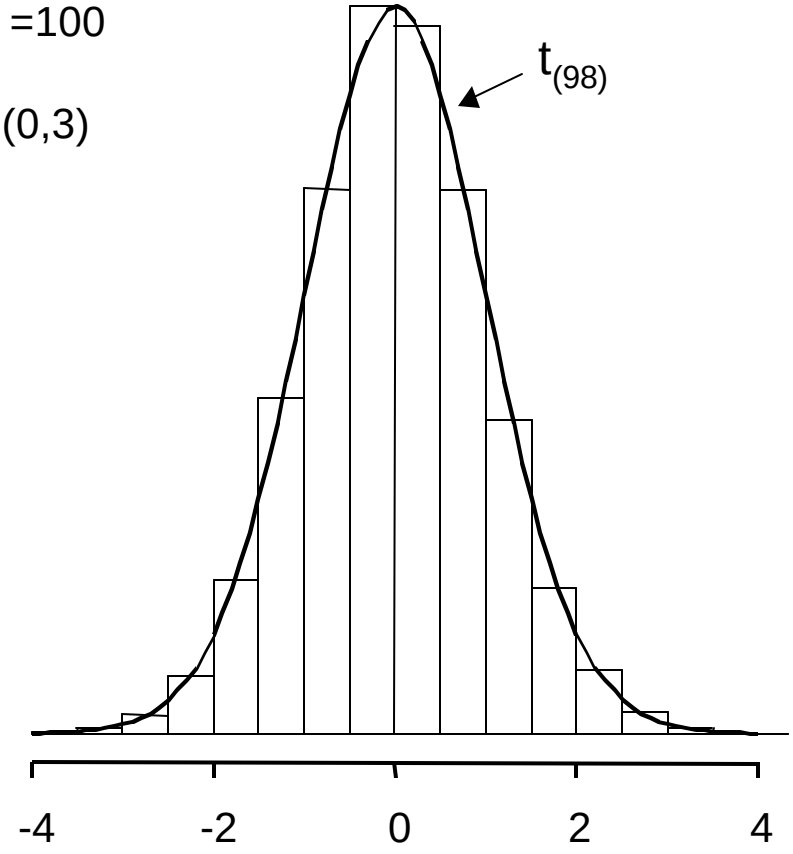
Sample Size =100

$y=1+0x+e\sim N(0,3)$

$b=L_{xy}/L_{xx}$




Sampling Distribution of b




Null-Distribution of b/(SE(b))

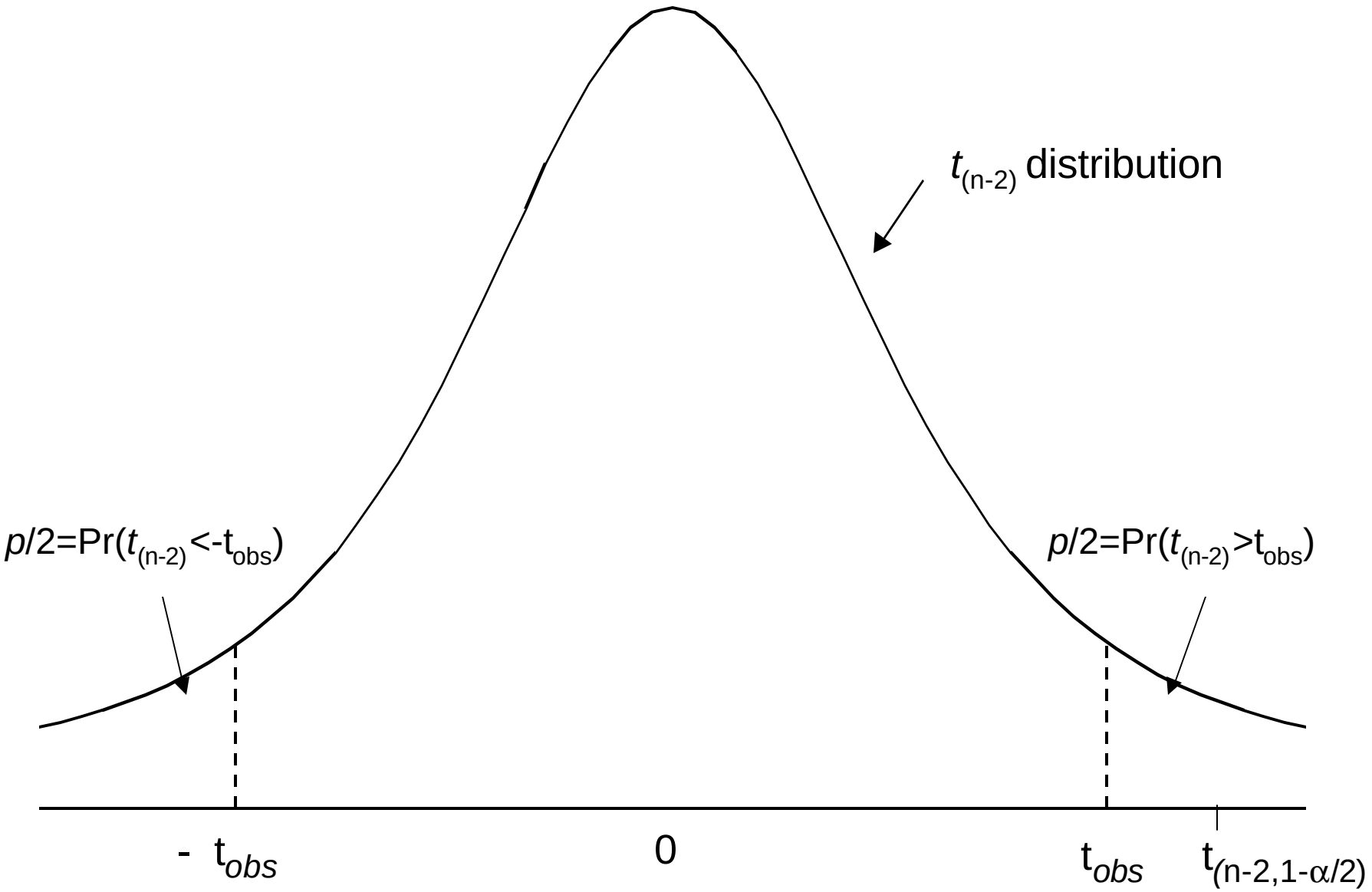
f) *t*-Test for the Hypothesis of No Association between Y and X.

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

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

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

# *t* Null-Distribution



g) 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: $\beta = 0$.

## h) 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 |

## X) Interval Estimation for $\alpha$ and $\beta$.

- From the sample of data that we observed we can compute an estimate for $\alpha$ and $\beta$. If we were to collect a second sample of data from the same reference population and we were to again estimate $\alpha$ and $\beta$ by the methods of least squares we would expect that these estimates would differ in magnitude somewhat from those that we estimated from the first sample of data. It is highly improbable that the least squares parameter estimates from any single sample of data would give the exact value of $\alpha$ and $\beta$ for the population. We are thus limited to defining a range of plausible values for $\alpha$ and $\beta$. When the data generating process adheres to the assumptions of the least squares simple-linear regression model we can define a range of plausible values for both $\alpha$ and $\beta$ with a specified level of confidence. These two ranges are referred to as the confidence intervals for $\alpha$ and $\beta$.

- 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 $\beta$ and $\alpha$ 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 $\beta$ and $\alpha$ 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 $\beta$ and $\alpha$ by the preceding criterion, we would expect 100%(1-$\alpha^*$) of these confidence intervals to contain the true value of the respective parameter.

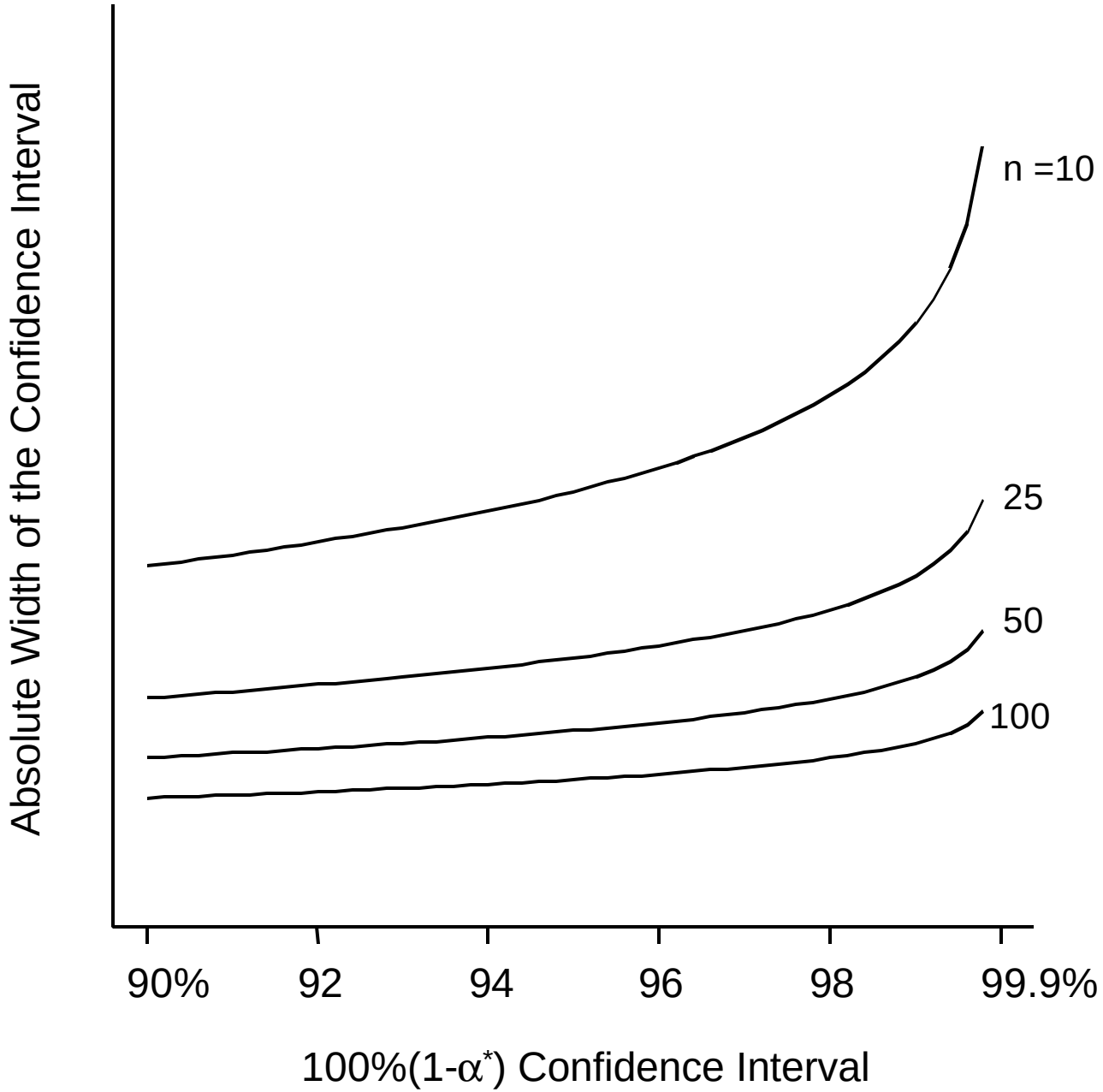
a) Example: Regression of $VO_2$ Max onto Duration of Exercise

95% CI for $\alpha$.

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

95% CI for $\beta$.

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


Absolute Width of the Confidence Interval
n =10
25
50
100
90%
92
94
96
98
99.9%
100%(1-α*) Confidence Interval

## XI) Interval Estimation for the $E(y_i|x_i)$ and $y_i$.

- There are two $100\%(1-\alpha^*)$ intervals that are commonly used to estimate a range of plausible values for the expected value of $y_i$ at $x_i$. The first is referred to as the $100\%(1-\alpha^*)$ confidence interval (CI), which is utilized when the goal is to estimate the value of the expected value of $y_i$ at a single x within the observed range of the $x_i(s)$. The second is referred to as the $100\%$ $(1-\alpha^*)$ simultaneous confidence band (CB), which is utilized when the goal is to estimate the expected value of $y_i$ at all values of X within the range of the observed $x_i(s)$.

- When the goal is to predict the value y at an unobserved x within the range of the observed $x_i(s)$, we use a $100\%(1-\alpha^*)$ prediction interval (PI).

Formulas for Computing the $100\%(1-\alpha^*)$ CI, CB, and PI.

a) $100\%(1-\alpha^*)$ confidence interval (CI) for the $E(y_i \mid x_i)$ for one x.

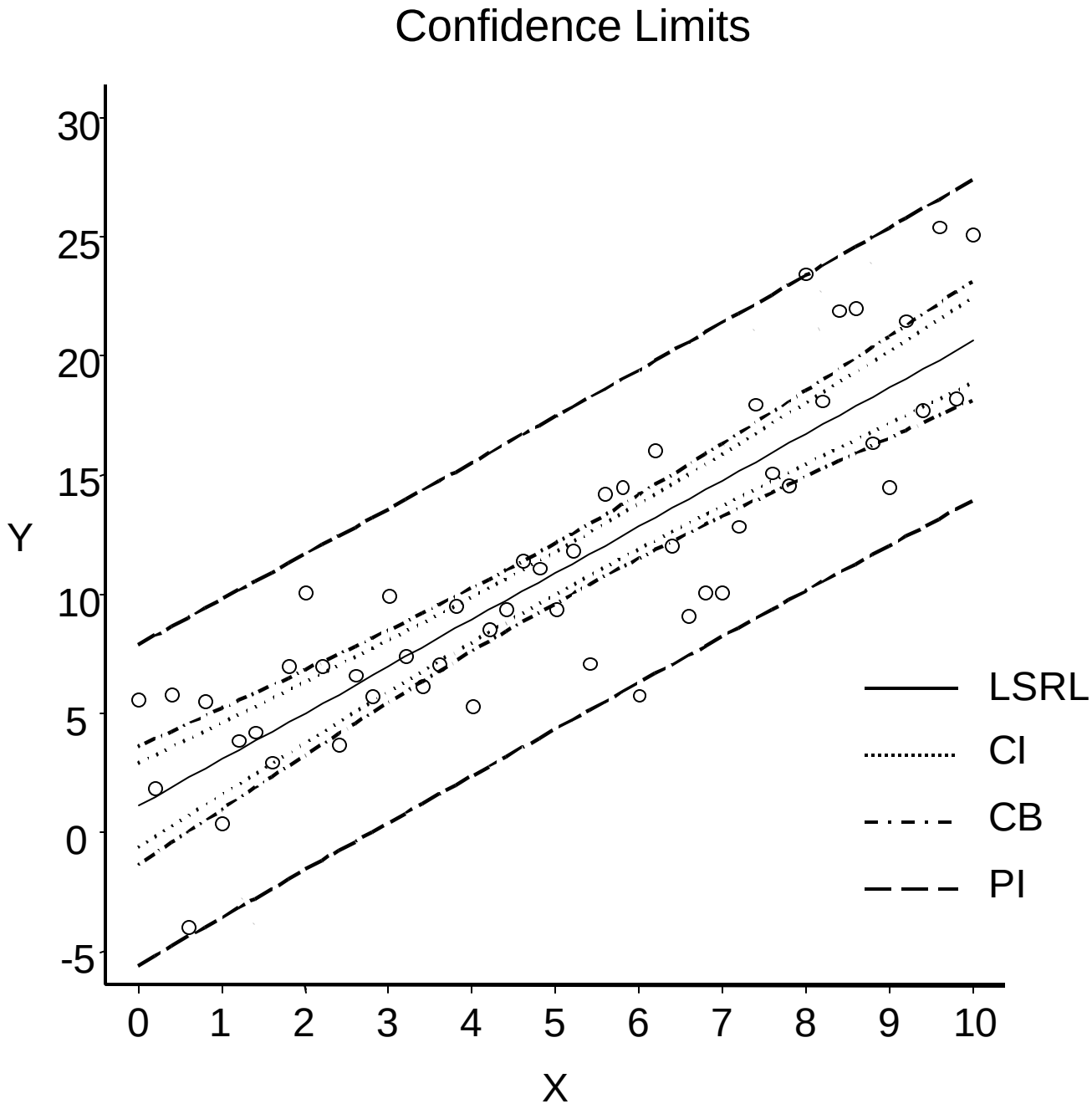
$$\hat{y}_i \pm t_{(n-2,1-\alpha^*/2)} \sqrt{s^2_{x.y}\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{L_{xx}}\right)}$$

b) $100\%(1-\alpha^*)$ simultaneous confidence band (CB) for the $E(y_i \mid x_i)$ at all X.

$$\hat{y}_i \pm \sqrt{2F_{(1-\alpha^*,2,n-2)}\left(s^2_{x.y}\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{L_{xx}}\right)\right)}$$

c) The $100\%(1-\alpha^*)$ prediction Interval (PI) for one y at one x.

$$\hat{y}_i \pm t_{(n-2,1-\alpha^*/2)} \sqrt{s^2_{y.x}\left(1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{L_{xx}}\right)}$$


Confidence Limits
30
25
20
15
10
5
0
-5
Y
0
1
2
3
4
5
6
7
8
9
10
X
LSRL
CI
CB
PI

## XII) 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) The model's stability.

- 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.

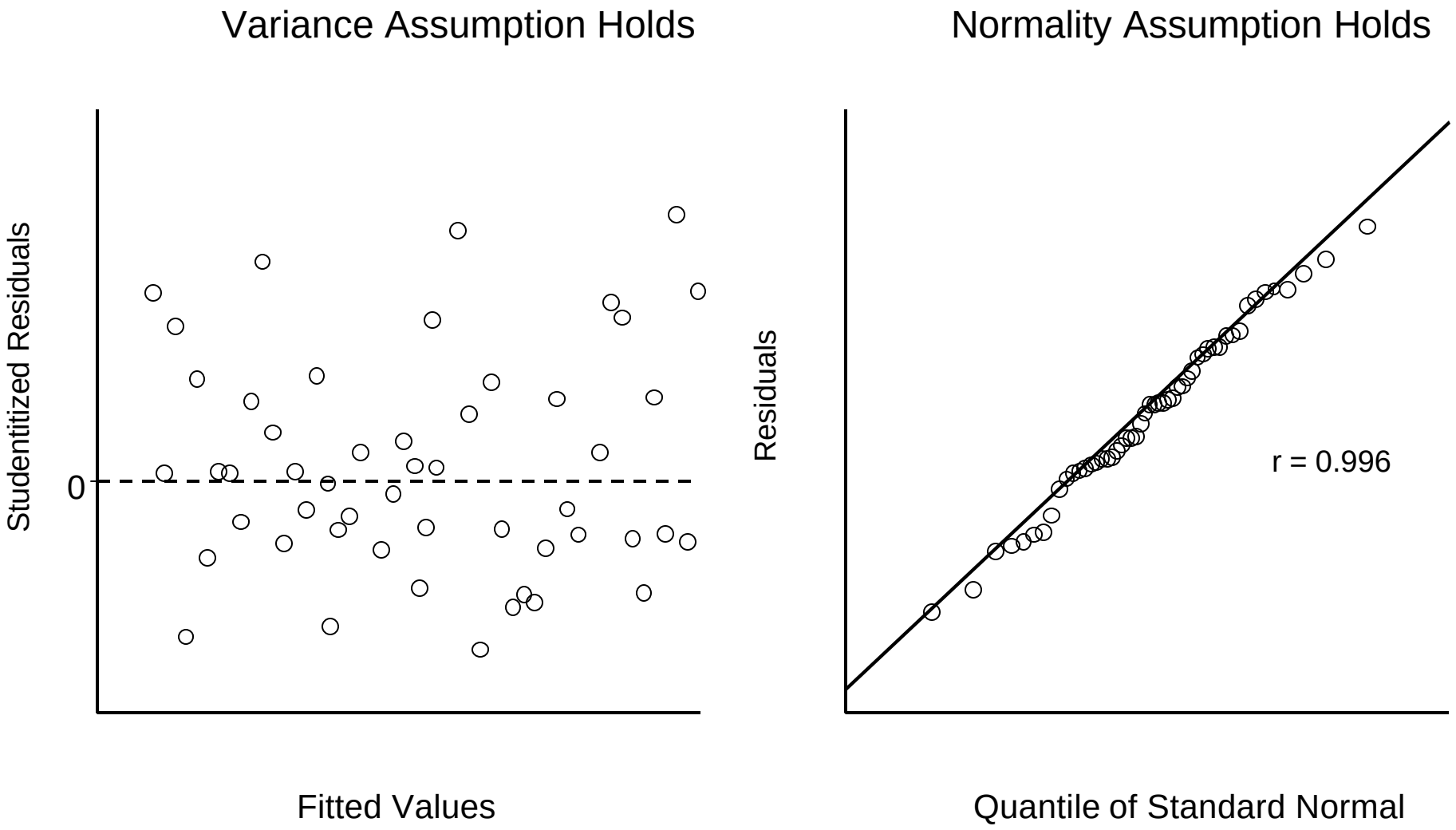
b) Checking the assumption of normality.

- 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.

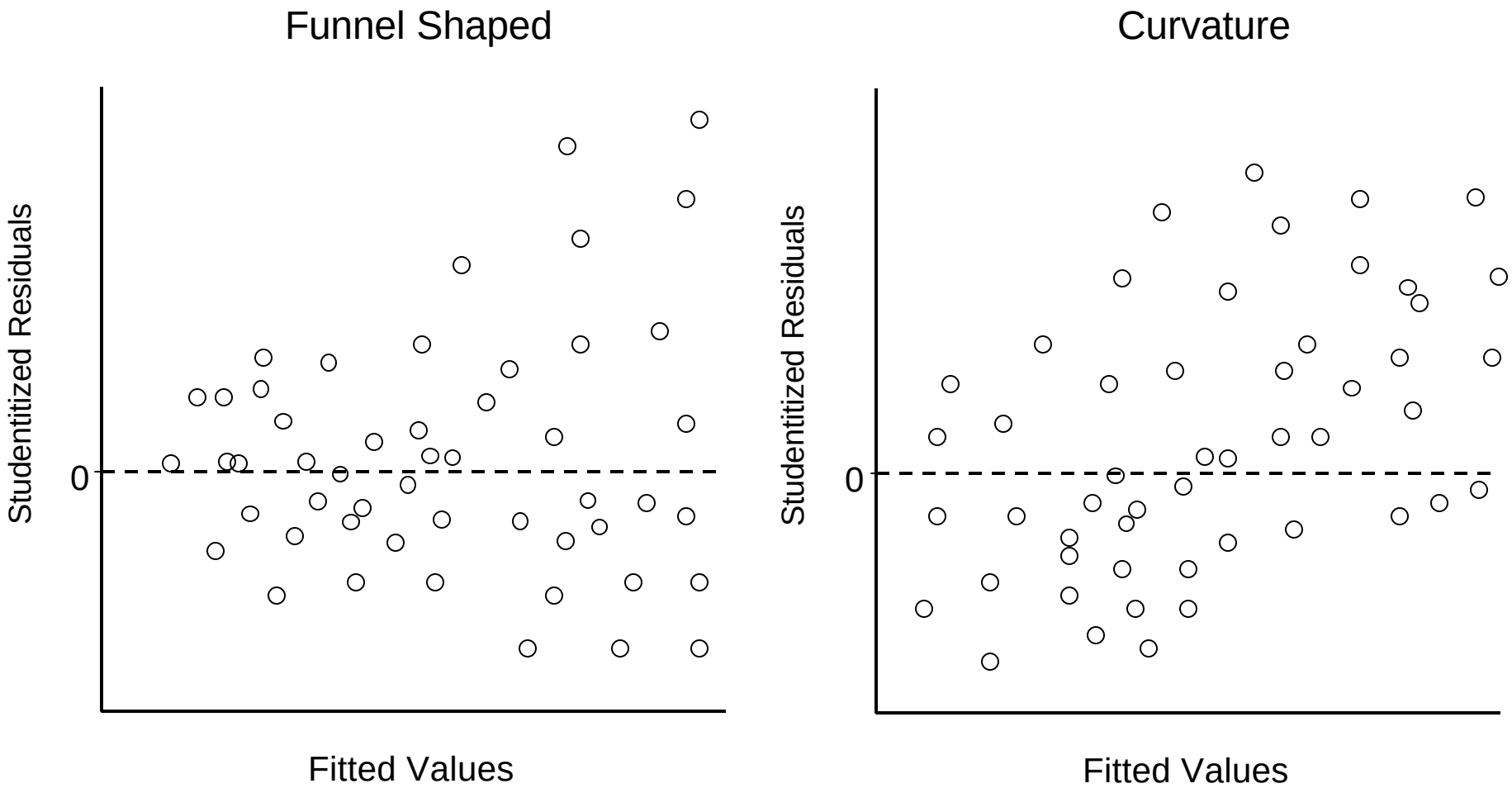
c) Checking for the correctness of functional form.

- 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

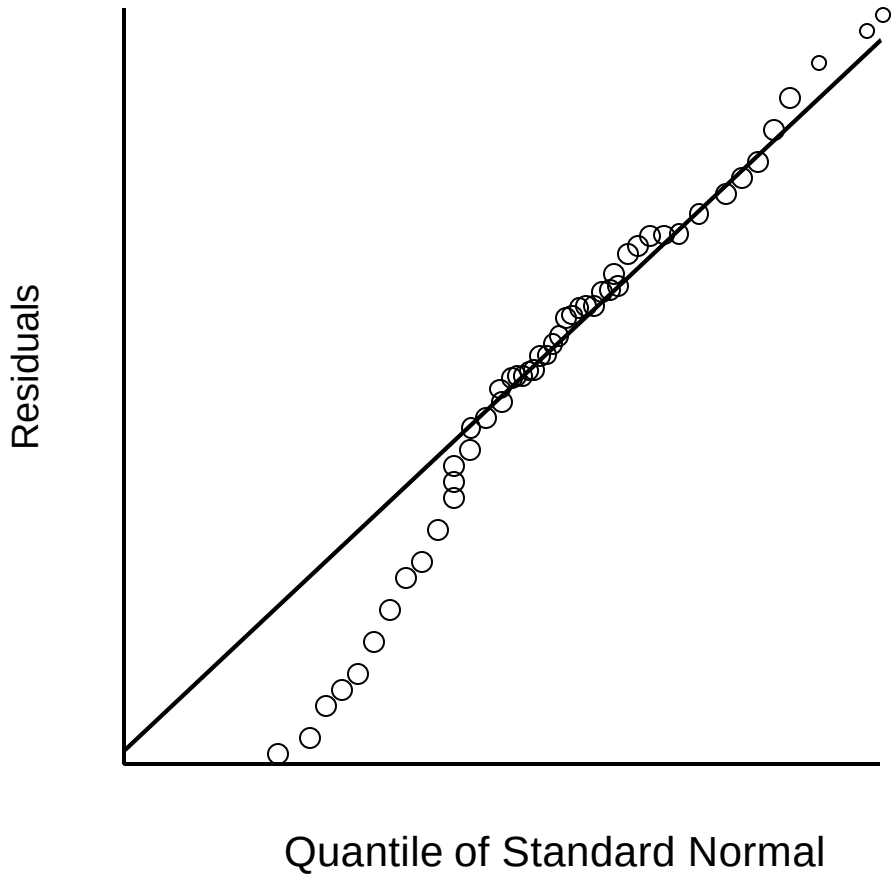


# Evidence of Non-Constant Variance

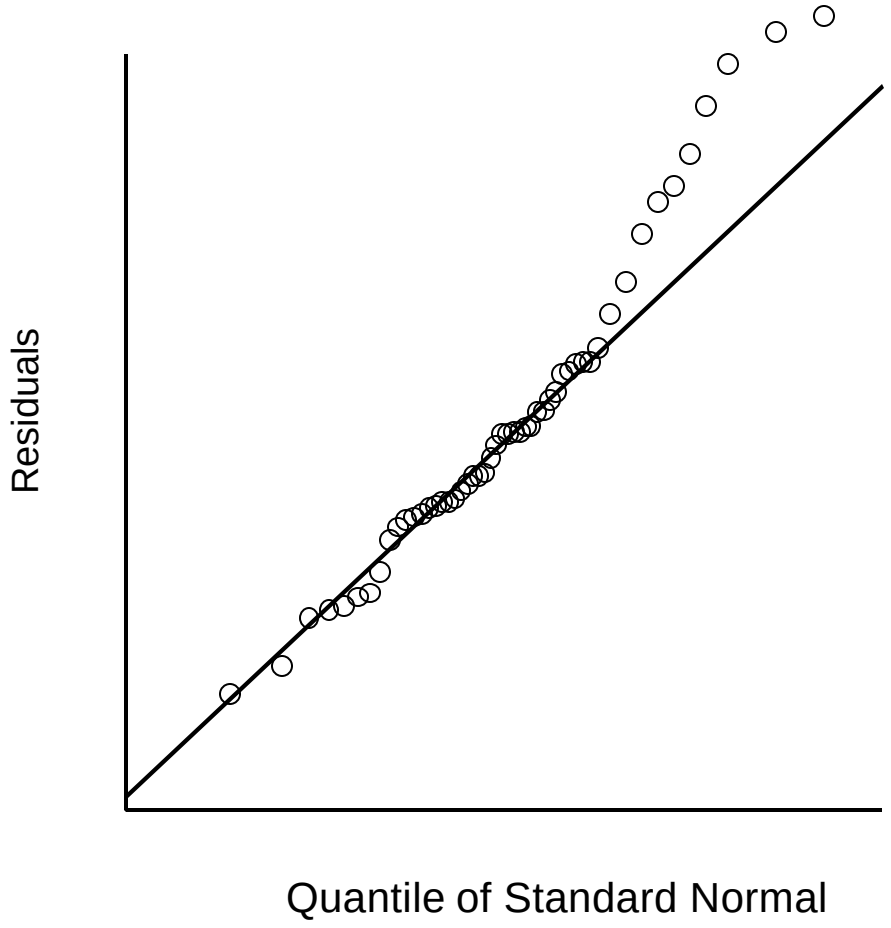


# Evidence of Non-Normality

## Skewed Left




## Skewed Right



d) Checking model stability.

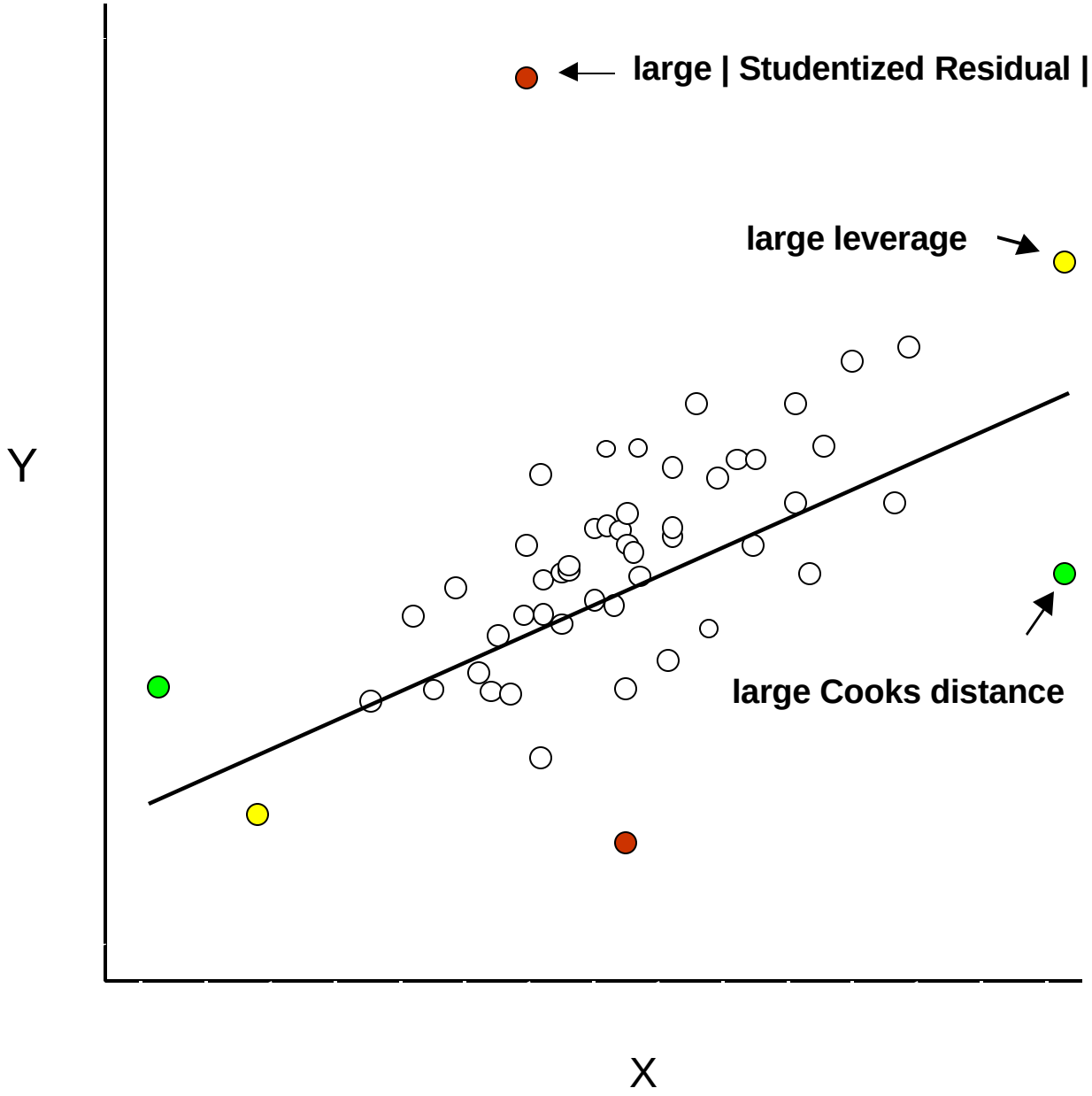
There are measures such as the Cook's distance, leverage, and the studentized residual that can be utilized to determine whether one or more observations are overly influential with regard to determining the value of the regression parameter estimate.

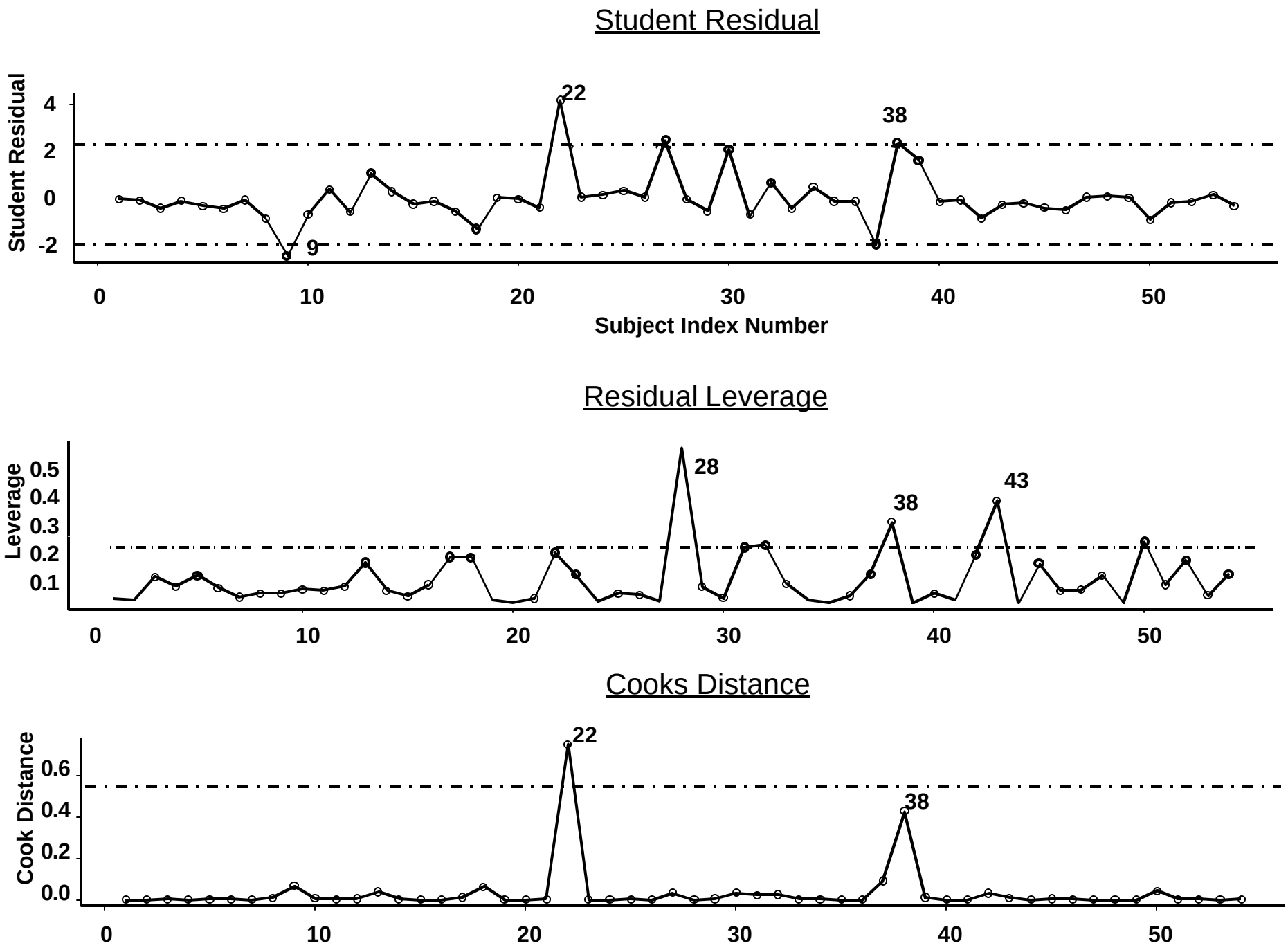
- Cook's distance provides a measure the magnitude by which the predicted values of the regression model change if the i*th* observation is removed from sample of data.

- Leverage provides a measure of how extreme the value $x_i$ is relative to the remaining values of X.

- The studentized residual provides a measure of how extreme the value of $y_i$ is relative to the remaining values of Y.

# Influence Measures




Student Residual
Student Residual
4
2
0
-2
22
38
9
0
10
20
30
40
50
Subject Index Number
Residual Leverage
Leverage
0.5
0.4
0.3
0.2
0.1
28
38
43
0
10
20
30
40
50
Cooks Distance
Cook Distance
0.6
0.4
0.2
0.0
22
38
0
10
20
30
40
50

XIII) Remedial Measures.

- Non-Constant Variance

A transformation of the response variable to a new scale (e.g. log) is often helpful in attaining equal residual variation across the range of the fitted values. If there is no variance stabilizing transformation which rectifies the situation an alternative approach is to use a more robust estimator, such as iterative weighted least squares.

- Non-Normality

Generally, non-normality and non-constant variance go hand in hand. An appropriate variance stabilizing transformation will more than likely also be remedial in attaining normally distributed residual error.

- 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.

## XIV) 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). The value 100% x $R^2$ 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 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.

## XV) Model Validation

- Model validation is utilized to determine how accurately your model will predict when your regression equation is applied to a new sample of data. It is important to note that when the regression parameters are estimated from your sample of data the parameter estimates are optimized to fit your sample of data. When you apply your model to a new sample of data, the value of the measure of predictive accuracy will be less than the value that was computed from the original regression fit because the model parameters estimates from the original fit were not optimized to predict the values of the response from the new sample of data.

There are several methods of model validation.

a) External-Validation

The least squares regression model equation is applied to a new sample of data that is independent from the sample that was utilized to estimate the model parameters. The measure of predictive accuracy is computed base the discrepancy between the observed value of the outcome from the external sample and the predicted value of the outcome when the predicted value is derived from the original regression equation.

b) Cross-Validation

The least squares regression model parameters are estimated from a randomly selected subset of the sample of observations. The regression model equation is then applied to the subset of the observations that were withheld from the model building process, and the measure of predictive accuracy is computed.

c) Bootstrap-Validation

Alternatively, the estimate of the measure of predictive accuracy from the original fit of the regression model can be adjusted for over optimism by a bootstrap resampling procedure. The bootstrap validation provides a estimate of the optimism that is induced by optimizing the model's fit to the observed sample of data. The optimism is then subtracted from the the original index of predictive accuracy.

# Part II
# Multiple Linear Regression

# Least Squares Multivariate Regression

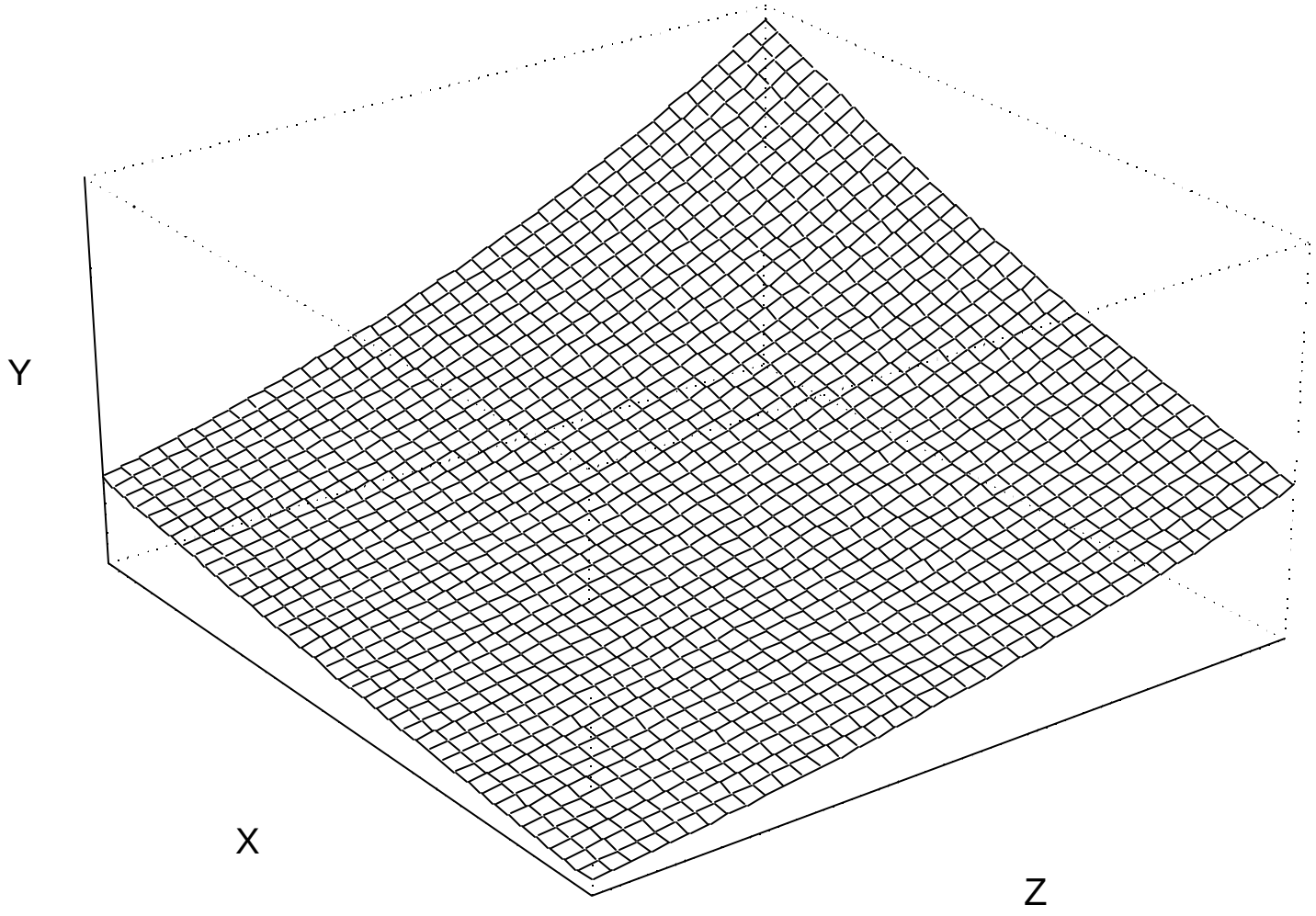
Multivariate regression by the method of least squares is an extension of the least squares simple linear regression model. The multivariate regression methods:

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

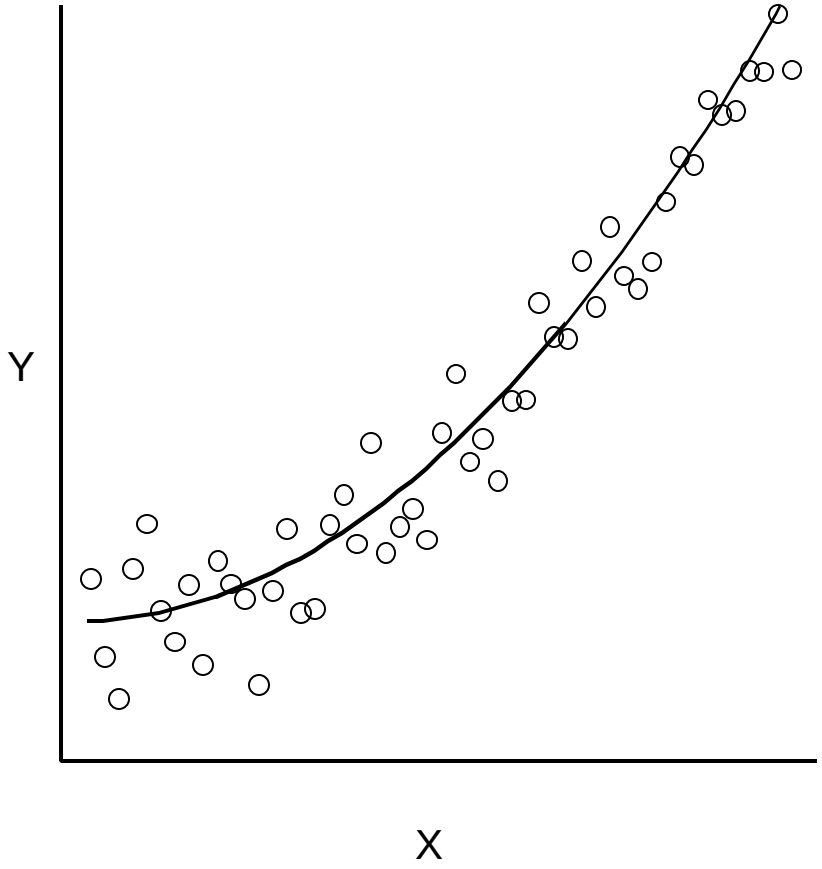
- Allow non-linear relationships between the response variable and the independent variables to be evaluated.

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

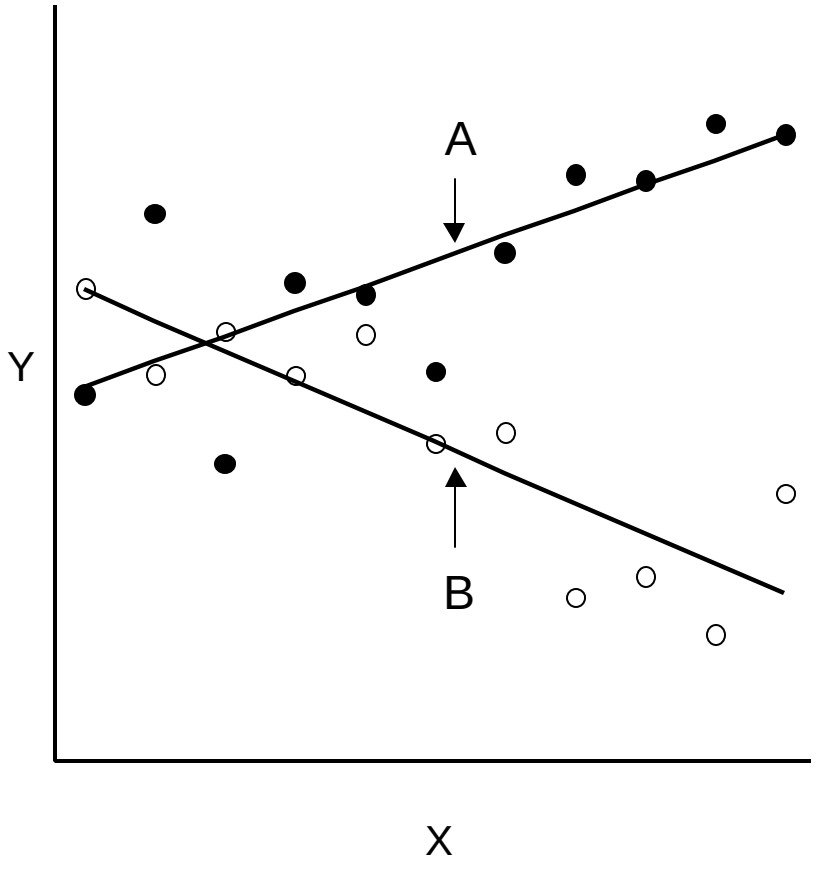
# Response Surface



Non-Linear Trends




Interaction



# I) The Multivariate Least Squares Regression Equation.

## a) In conventional notation.

$$y_i = \hat{a}_o + \hat{a}_1, x_{i,1} + \hat{a}_2, x_{i,2} + \cdots + \hat{a}_{p-1}, x_{i,p-1} + \mathring{a}_i$$

where

$y_i$ is the ith response.

$\hat{a}_o, \hat{a}_1, \cdots \hat{a}_{p-1}$ are the regression parameters.

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

$\mathring{a}_i$ is the independent random error associated with the ith response, typically assumed to be distributed $N(0, \acute{o})$.

b) In matrix notation.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$\mathbf{y}$ = n x 1 vector of response values.

$\mathbf{X}$ = n x p matrix of known constants (predictors).

$\boldsymbol{\beta}$ = p x 1 vector of regression parameters.

$\boldsymbol{\varepsilon}$ = n x 1 vector of identically distributed random errors, typically assume to be distributed $N(0,\sigma^2)$.

$$y_{(n \times 1)} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X_{(n \times p)} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,n} \end{bmatrix}$$

$$\hat{a}_{(p \times 1)} = \begin{bmatrix} \hat{a}_o \\ \hat{a}_1 \\ \vdots \\ \hat{a}_{p-1} \end{bmatrix} \qquad \mathring{a} = \begin{bmatrix} \mathring{a}_1 \\ \mathring{a}_2 \\ \vdots \\ \mathring{a}_n \end{bmatrix}$$

## II) The Multivariate Least Squares Regression Model.

### a) In conventional notation.

$$E(y_i \mid x_{i,1,..,p-1}) = \hat{a}_o + \hat{a}_1, x_{i,1} + \hat{a}_2, x_{i,2} + \cdots + \hat{a}_{p-1}, x_{i,p-1}$$

where

$E(y_i \mid x_{i,1,...,p-1})$ is the expected value of ith response.

$\hat{a}_o, \hat{a}_1, \cdots \hat{a}_{p-1}$ are the regression parameters.

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

b) In matrix notation.

$$E(Y|X) = X\mathbf{b}$$

where

E(y|X) = n x 1 vector of predicted values.

X = n x p matrix of known constants (predictors).

$\beta$ = p x 1 vector of regression parameters.

## c) The y and X model components

The column vector y consists of a set of random observations with a continuous scale of measure, while the columns of X may consist of a combination of the following:

- continuous measures, or functions of continuous measures (e.g. polynomial).

- dichotomous measures (e.g. gender).

- nominal (e.g. male female) or ordinal classification variables (e.g. age class).

- product terms computed between the values of two or more of the columns of X; referred to as interaction terms.

## III) Parameter Estimation.

To estimate the $\hat{a}(s)$ we minimize

$$Q = (y - X\hat{a})^T (y - X\hat{a})$$

by solving $\frac{\partial Q}{\partial \hat{a}} = 0$ with repect to $\hat{a}$.

The resulting estimator for the vector $\hat{a}$ is

$$b = (X^T X)^{-1} X^T y.$$

IV) Parameter Interpretation.

a) When $x_j$ is continuous, $\beta_j$ represents the change in the expected value of the response for a one unit change in $x_j$ with all remaining independent variables (x) held constant.

Note that

$$
\begin{aligned}
E(y_i \mid x_{i,j}+1) - E(y_i \mid x_{i,j}) &= \hat{a}_0 + \hat{a}_1 x_{i,1} + \cdots \hat{a}_j (x_{i,j}+1) + \cdots + \hat{a}_{p-1} x_{i,p-1} \\
&\quad - \hat{a}_0 + \hat{a}_1 x_{i,1} + \cdots \hat{a}_j (x_{i,j}) + \cdots + \hat{a}_{p-1} x_{i,p-1} \\
&= \hat{a}_j (x_{i,j}+1) - \hat{a}_j (x_{i,j}) \\
&= \hat{a}_j
\end{aligned}
$$

b) When $x_j$ is dichotomous, we create an indicator variable $z_j$ (j=1,2) such that $z_{i,j} = 0$ if $x_{i,j}$ equals a specified reference level of $x_j$, and $z_{i,j} = 1$ otherwise. $\beta_j$ represents the change in the expected value of Y from the reference level of $x_j$ ($z_j$ =0) to the non-reference level of $x_j$ ($z_j$ =1) with all remaining independent variables (x) held constant.

Note that

$$
\begin{aligned}
E(y_i \mid z_{i,j}+1) - E(y_i \mid z_{i,j}=0) &= \hat{a}_0 + \hat{a}_1 x_{i,1} + \cdots \hat{a}_j (z_{i,j}=1) + \cdots + \hat{a}_{p-1} x_{i,p-1} \\
&\quad - \hat{a}_0 + \hat{a}_1 x_{i,1} + \cdots \hat{a}_j (z_{i,j}=0) + \cdots + \hat{a}_{p-1} x_{i,p-1} \\
&= \hat{a}_j(1) - \hat{a}_j(0) \\
&= \hat{a}_j
\end{aligned}
$$

c) When $x_j$ is categorical with c categories we create binary indicator variables $z_k$ (k=1,2,…c-1) . For each indicator variable $z_k$ we let $z_{i,k}$ = 1 if $x_{i,j}$ = category $c_k$, zero otherwise. $\beta_k$ then represents the change in the expected value of the response when moving from the reference category $c_c$ of $x_j$ to category $c_{k.}$ with all remaining independent variables held constant.

Note that

$$E(y_i \mid x_{i,k} = c_k) - E(y_i \mid x_{i,k} = c_{k+1}) = E(y_i \mid z_{i,k} = 1) - E(y_i \mid z_{i,k} = 0)$$

$$= \hat{a}_0 + \cdots + \hat{a}_{k-1} x_{i,k-1} + \hat{a}_k (z_{i,k} = 1) + \cdots + \hat{a}_{p-1} x_{i,p-1}$$

$$- \hat{a}_0 + \cdots + \hat{a}_{k-1} x_{i,k-1} + \hat{a}_k (z_{i,k} = 0) + \cdots + \hat{a}_{p-1} x_{i,p-1}$$

$$= \hat{a}_k (1) - \hat{a}_k (0)$$

$$= \hat{a}_k$$

## V) Hypothesis Tests for the Multiple Linear Regression Model.

- In the multiple linear regression model setting, under the null hypothesis we assume that there is no linear association between the response variable Y and the independent variables $x_1, x_2, \cdots, x_{p-1}$.

- The global test of no association is once again based on the F-test.

- If the global test of association is rejected, the individual associations between Y and $x_1, x_2, \cdots, x_{p-1}$ can be evaluated by the common t-Test.

a) F-Test for the Hypothesis of No Association between Y and X.

Hypothesis: Ho: $\beta_1=\beta_{2,}\cdots=\beta_{p-1}=0$ versus $H_a$: some $\beta_j \neq 0$

$$F_{obs} = \frac{MSR}{MSE} = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/(p-1)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p)}$$

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

b) *t*-Test for the Hypothesis of No Association between Y and $x_j$.

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

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

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

## V) Extra Sum of Squares F-test.

- Often we want to compare two regression models, one of which contains a subset of the terms that are included in the other. We refer to the model which includes all the independent variables as the full model while we refer to the model which contains only a subset of the independent variables as the reduced model. To test whether the additional terms of the full model provide significant additional information about the variation in the response we use what is referred to as an extra sum of squares F-Test. This test requires fitting both the full and the reduced model.

Extra Sum of Squares F - test

$$E(Y \mid X) = \hat{a}_0 + \hat{a}_1 x_1 + \cdots + \hat{a}_{k-1} x_{k-1} + \hat{a}_k x_k + \cdots + \hat{a}_{p-1} x_{p-1}$$

$$Ho : \hat{a}_k = \hat{a}_{k+1} = \cdots = \hat{a}_{p-1} = 0$$

$$Ha : \text{not all of the } \hat{a}_k \cdots \hat{a}_{p-1} = 0$$

$$F^* = \frac{\frac{SSE_R - SSE_F}{df_R - df_F}}{\frac{SSE_F}{df_F}}$$

$F^*$ is distributed $F_{(df_R - df_F, df_F)}$

If $F^* \geq F_{(df_R - df_F, df_F, 1-\acute{a}^*)}$ conclude Ha.

## VII) Interval Estimation for $\beta_j$.

If $b_i$ is the least squares estimator for $\beta_j$ and $SE(b_j)$ is the estimated standard error of $b_i$ then the $100\%(1-\alpha^*)$ confidence interval for $\beta_j$ is given by $b_j \pm t_{(n-p,1-\alpha^*/2)}SE(b_j)$.

VIII) Interval Estimation for the $E(y_i|x_{i,1,..p-1})$ and $y_i$

a) $100\%(1-\alpha^*)$ confidence interval (CI) for the $E(y_i \mid x_{i,1,..p-1})$.

$$\hat{y}_i \pm t_{(n-p,1-\alpha^*/2)} \sqrt{MSE\,[x_i^T (X^T X)^{-1} x_i]}$$

b) $100\%(1-\alpha^*)$ simultaneous confidence band (CB) for the $E(y_i \mid x_{i,1..p-1})$.

$$\hat{y}_i \pm \sqrt{pF_{(1-\alpha^*,p,n-p)} MSE[x_i^T (X^T X)^{-1} x_i]}$$

c) The $100\%(1-\alpha^*)$ prediction Interval (PI) for one $y_i$.

$$\hat{y}_i \pm t_{(n-2,1-\alpha^*/2)} \sqrt{MSE(1+\frac{1}{n}+x_i^T(X^TX)x_i)}$$

## IX) Residual Diagnostics for Checking Goodness of Fit

As in the case of the simple-linear regression model, the things to check once you have fit your initial multiple linear regression model included:

a) The assumption of constant variance
b) The assumption of normality.
c) The assumption of additivity.
d) The correctness of functional form.
e) The model's stability.

The preceding diagnostic checks can be conducted with the same graphical methods as were utilized in the simple linear model setting.

# Case Study #1

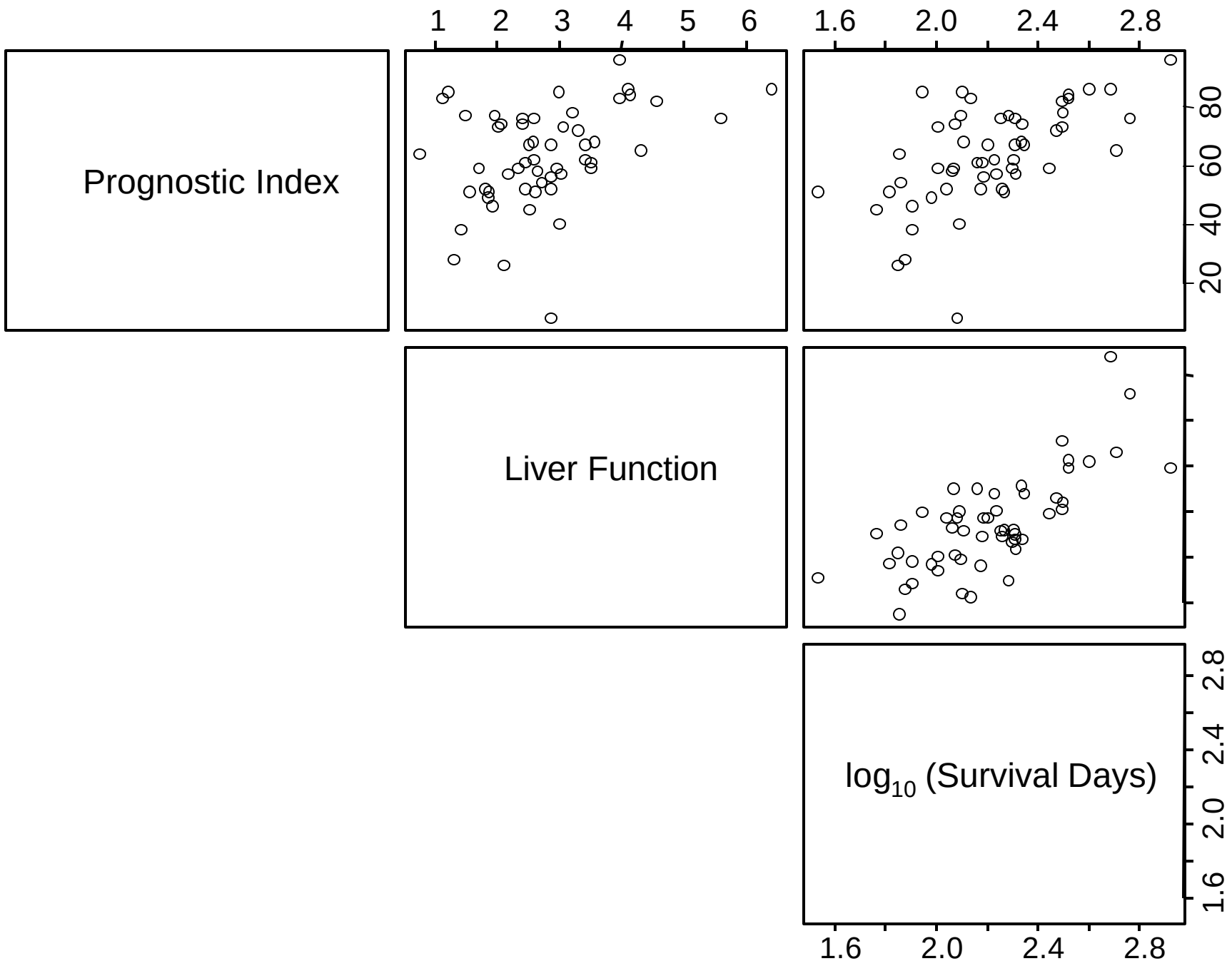
A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random sample of 54 patients was available. From each patient, the following information was extracted from the patients pre-operative records: 1) blood clotting score, 2) prognostic 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 | 1.408 |
| Prog. Index. | 0.006 |
| Liver Fun. | 0.150 |

Regression Equation

$$E(\log(\text{Surv})_i|x_i)=1.408 + 0.006(\text{Prog. Index}_i) + 0.150(\text{Liver Fun}_i)$$

# 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 $x_j$.

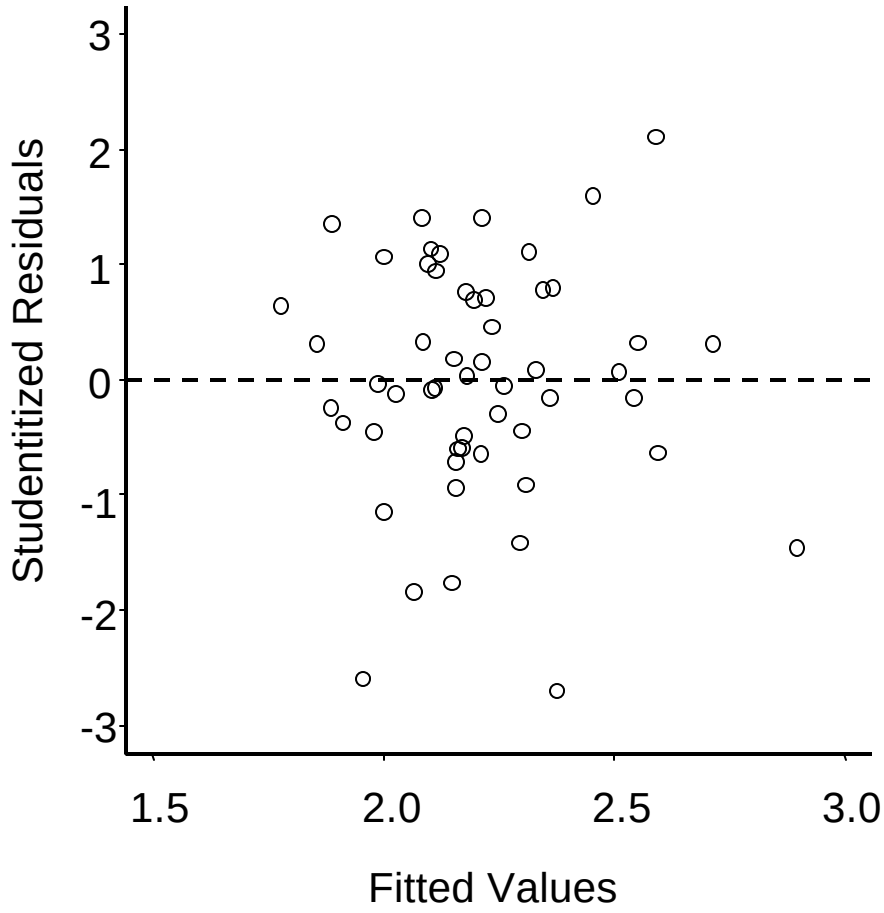
| Parameter | Estimate b | SE b | $t_{obs}$ | $P(T>t_{obs})$ |
|---|---|---|---|---|
| Intercept | 1.408 | 0.084 | | |
| Prognostic Index | 0.006 | 0.001 | 4.217 | <0.001 |
| Liver Function | 0.150 | 0.023 | 6.586 | <0.001 |

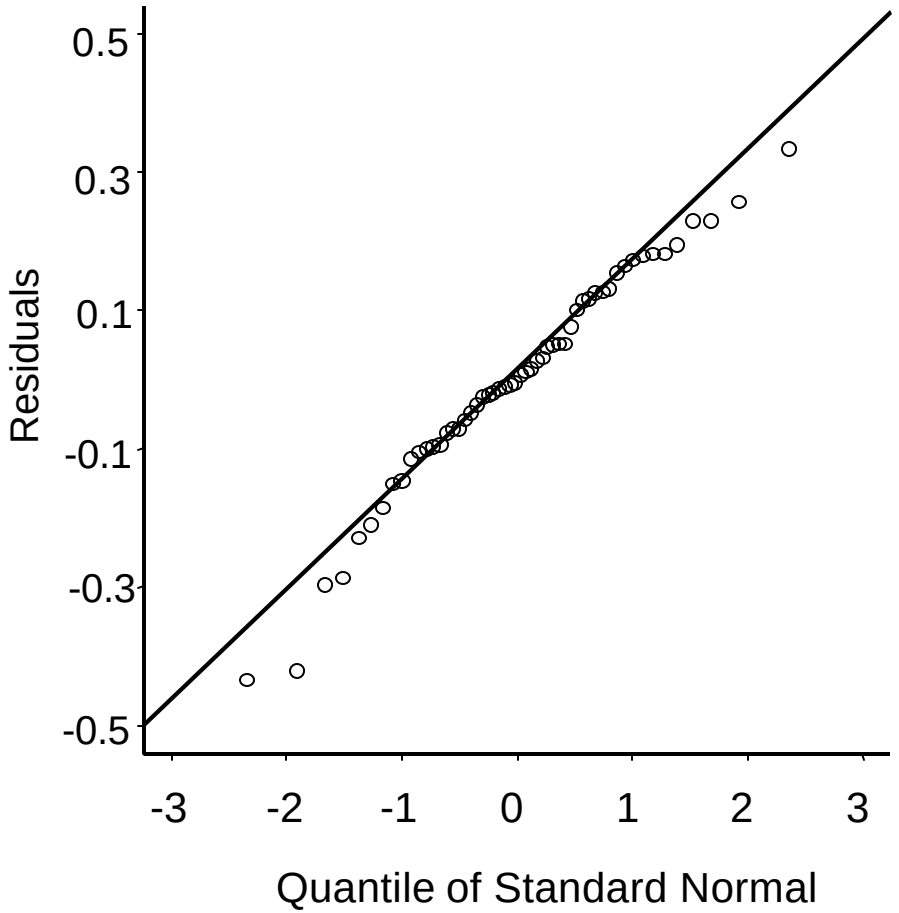
## Confidence Intervals for the $\beta_j$ (s).

Table 4. 95% confidence intervals of the regression parameters

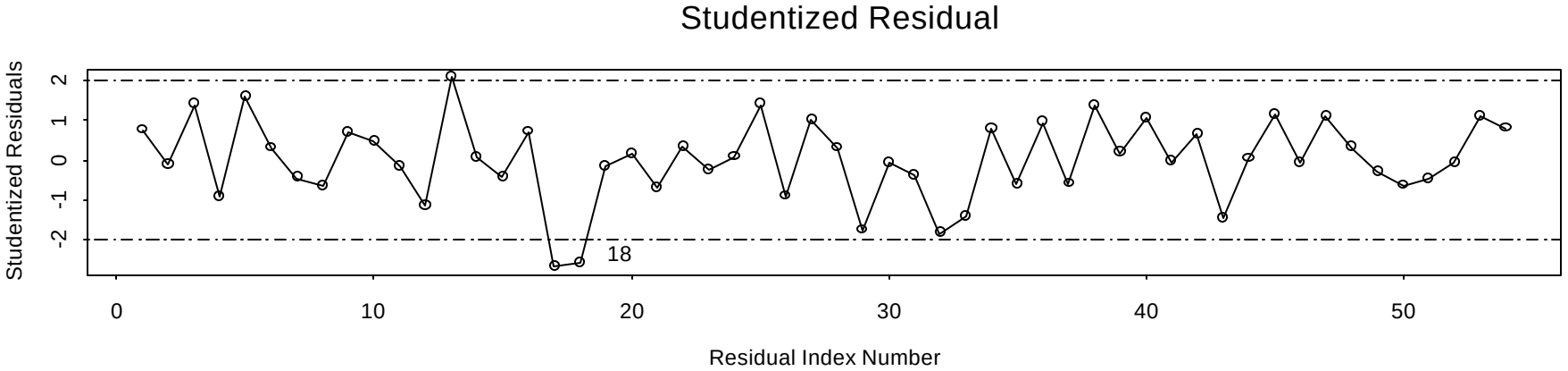
| Parameter | Estimate b | SE b | df | $t_{(51,0.975)}$ | Lower 95%CL | Upper 95%CL |
|---|---|---|---|---|---|---|
| Intercept | 1.408 | 0.092 | | | | |
| Prog Index | 0.006 | 0.001 | 51 | 2.00 | 0.004 | 0.008 |
| Liver Fun. | 0.150 | 0.023 | 51 | 2.00 | 0.104 | 0.196 |

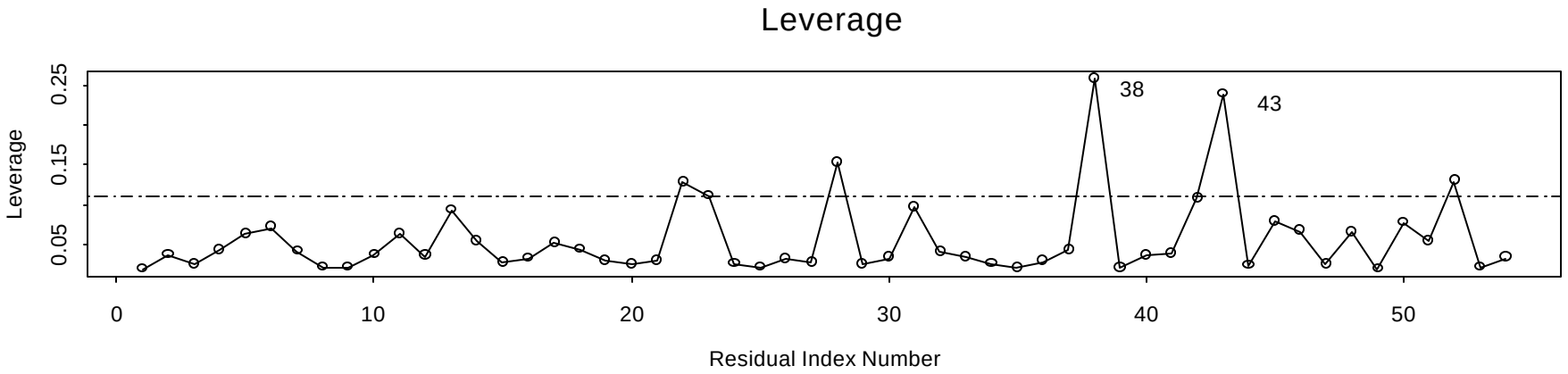
# Residual Diagnostics

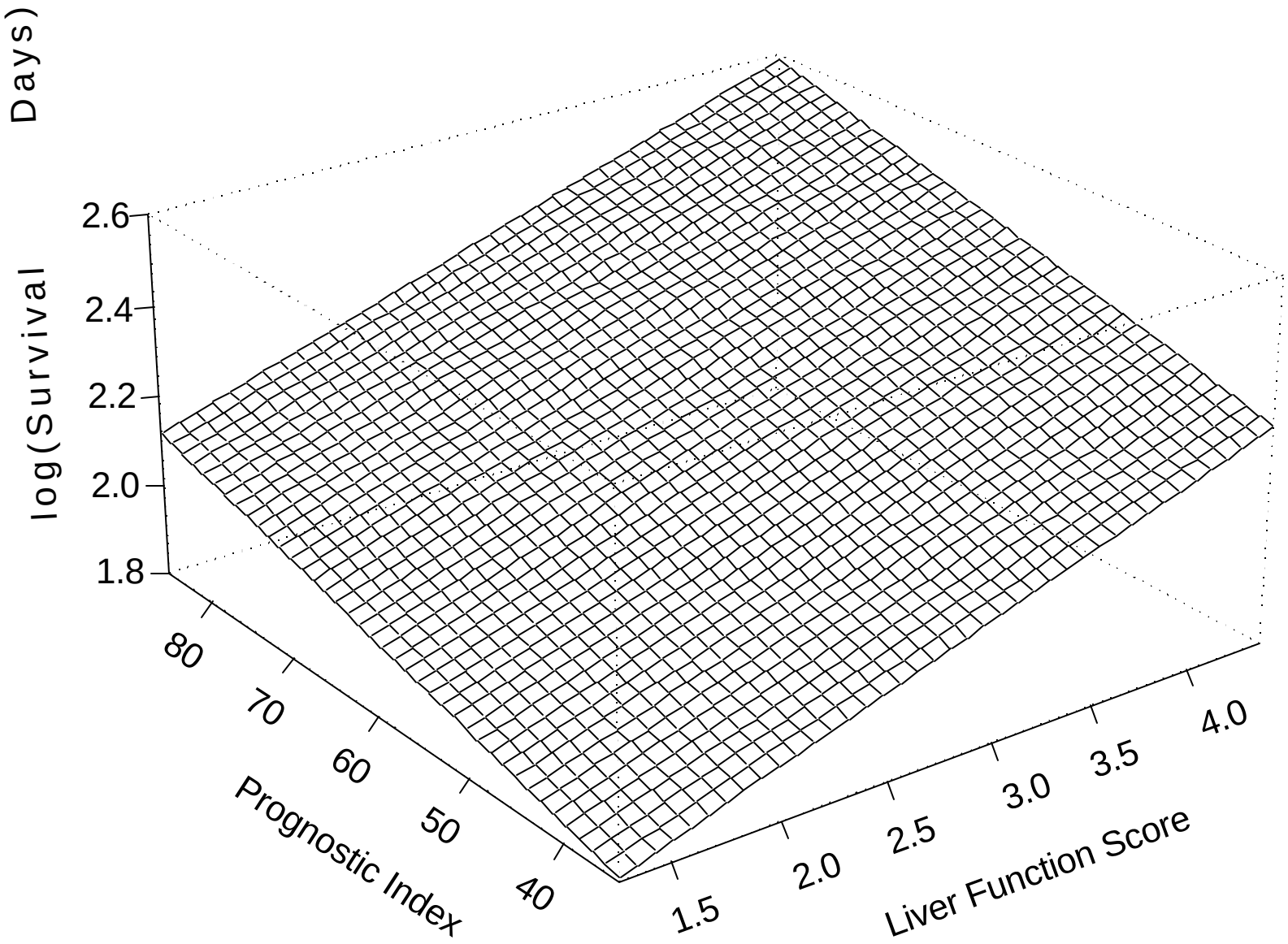


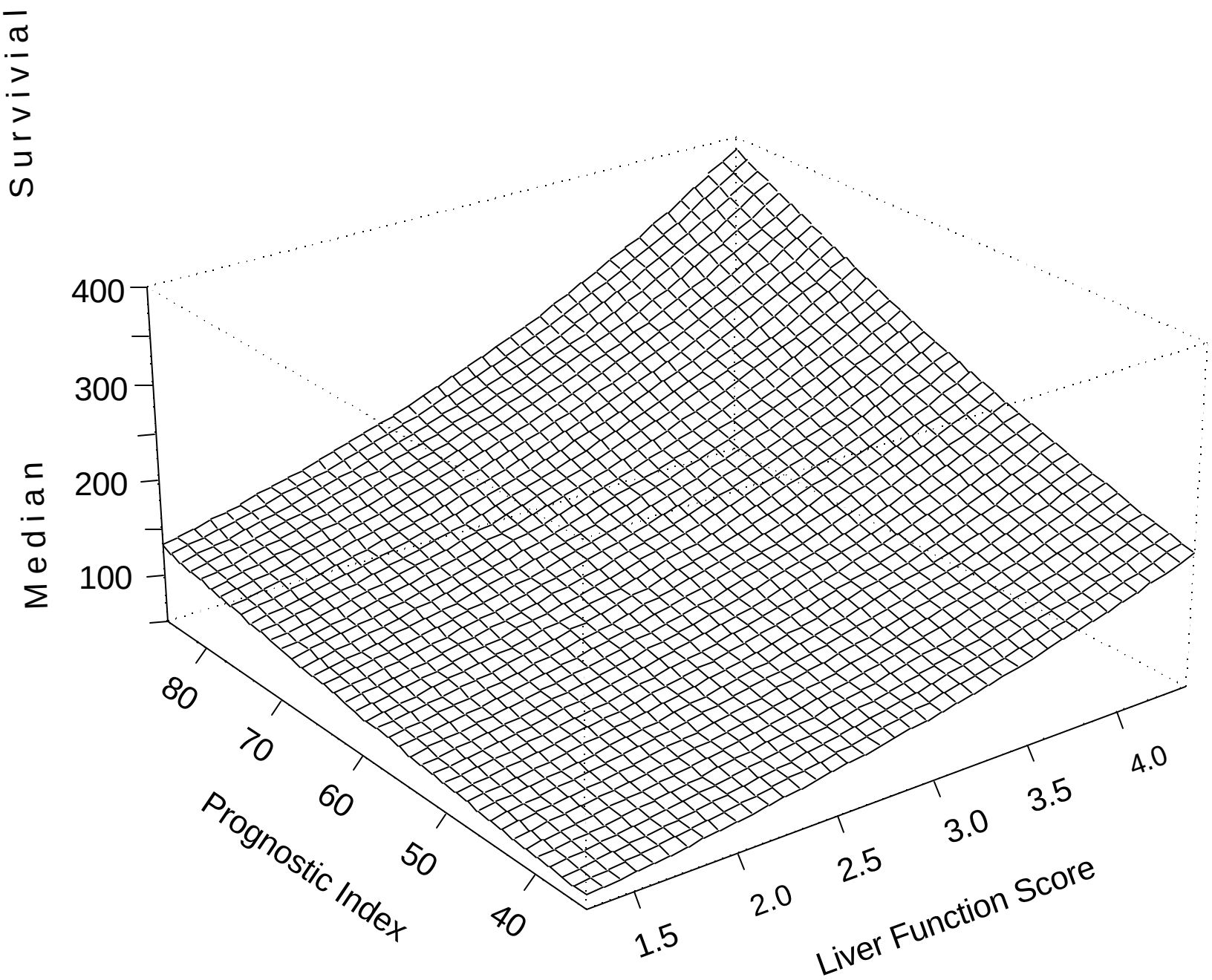




Cooks Distance
Value of Cook Distance
0.0
0.10
0.20
38
43
0
10
20
30
40
50
Residual Index Number


Studentized Residual
Studentized Residuals
-2
-1
0
1
2
18
0
10
20
30
40
50
Residual Index Number


Leverage
Leverage
0.05
0.15
0.25
38
43
0
10
20
30
40
50
Residual Index Number


log(Survival Days)
2.6
2.4
2.2
2.0
1.8
80
70
60
50
40
Prognostic Index
1.5
2.0
2.5
3.0
3.5
4.0
Liver Function Score


Survivial
Median
400
300
200
100
80
70
60
50
40
Prognostic Index
1.5
2.0
2.5
3.0
3.5
4.0
Liver Function Score

## 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 # 2

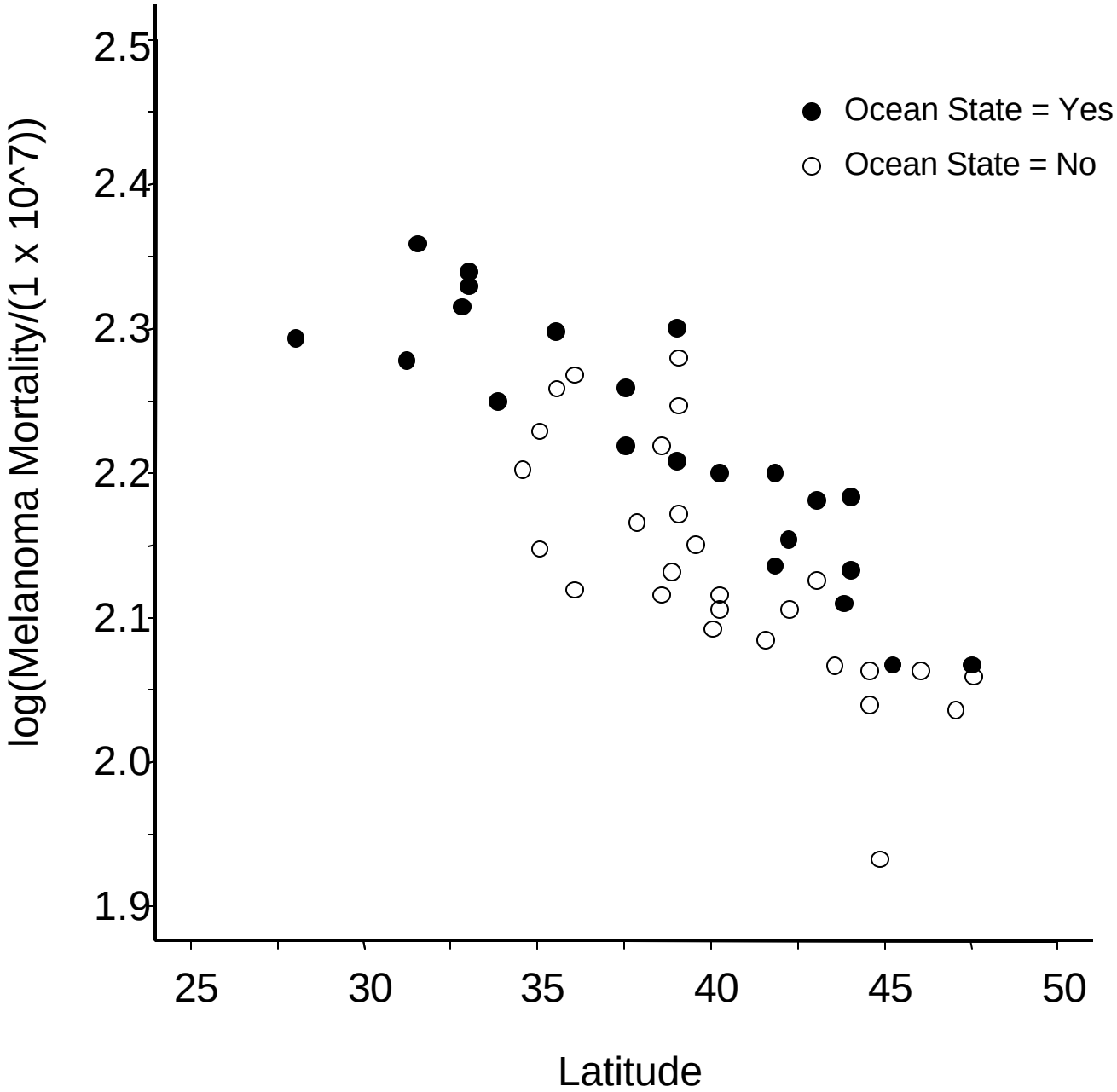
Data on the mortality due to malignant melanoma of the skin of white males were collected during the period of 1950-1969 from each state in the United States as well as the District of Columbia. No mortality data were available for Alaska and Hawaii for this period. The goal of the study was to assess 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 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 |

Least Squares Separate Slopes Model.

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

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** |

An Equivalent Test by the Extra Sum of Squares F-test.

$Ho: \hat{a}_3 = 0$ versus $Ha: \hat{a}_3 \neq 0$

$$F^* = \frac{\frac{SSE_R - SSE_F}{df_R - df_F}}{\frac{SSE_F}{df_F}} = \frac{\frac{1.09 - 1.08}{46 - 45}}{\frac{1.08}{45}} = 0.417$$

$F_{(1,45,0.95)} = 4.06$

$F^* < F$ which $\Rightarrow$ we fail to reject Ho.

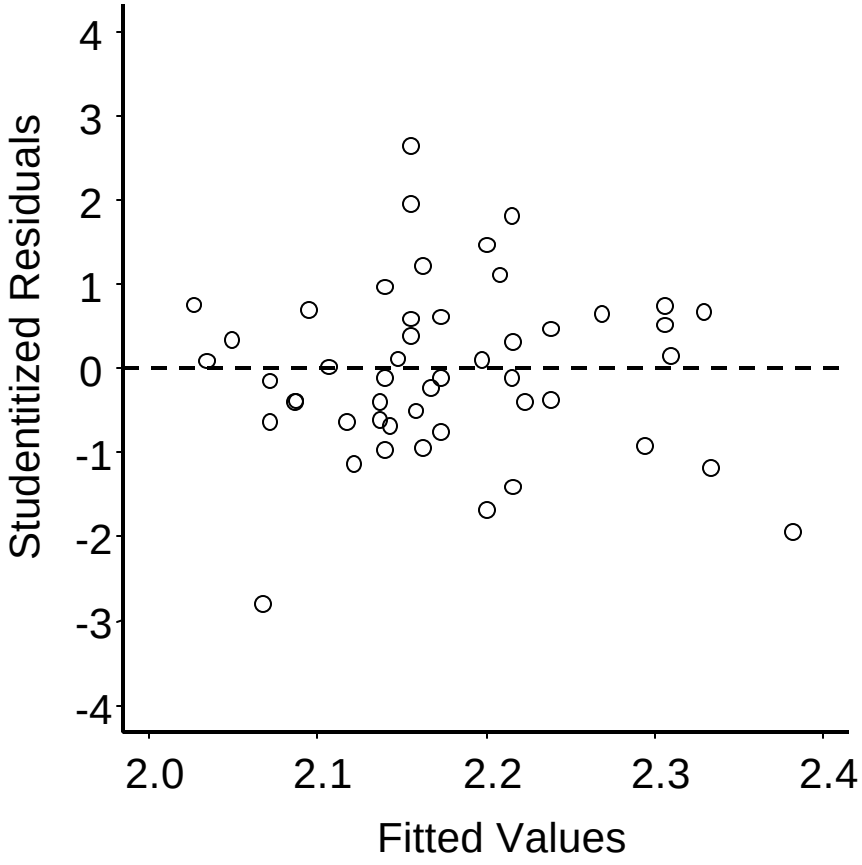
Note that when $df_R - df_F$ of freedom equals1 that $(t_{obs})^2 = F_{obs}$ which $\Rightarrow$ the pvalue associated with the t - Test is equal to the p value associated with the F test.

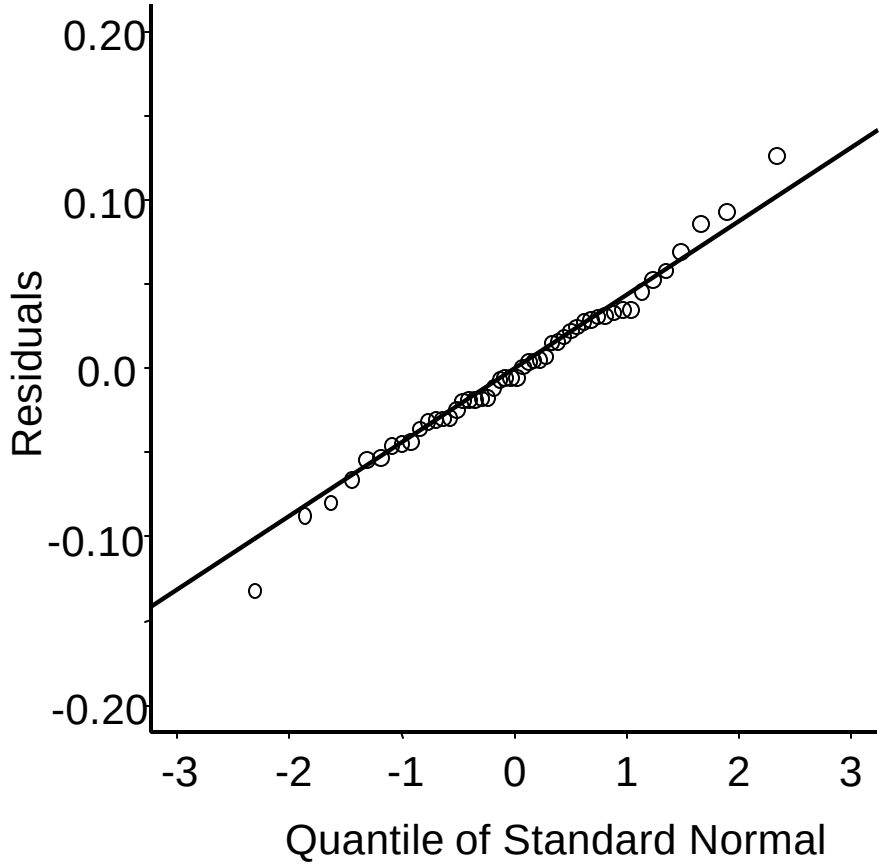
# Confidence Intervals for the $\beta_j$(s) of the Common Slope Model

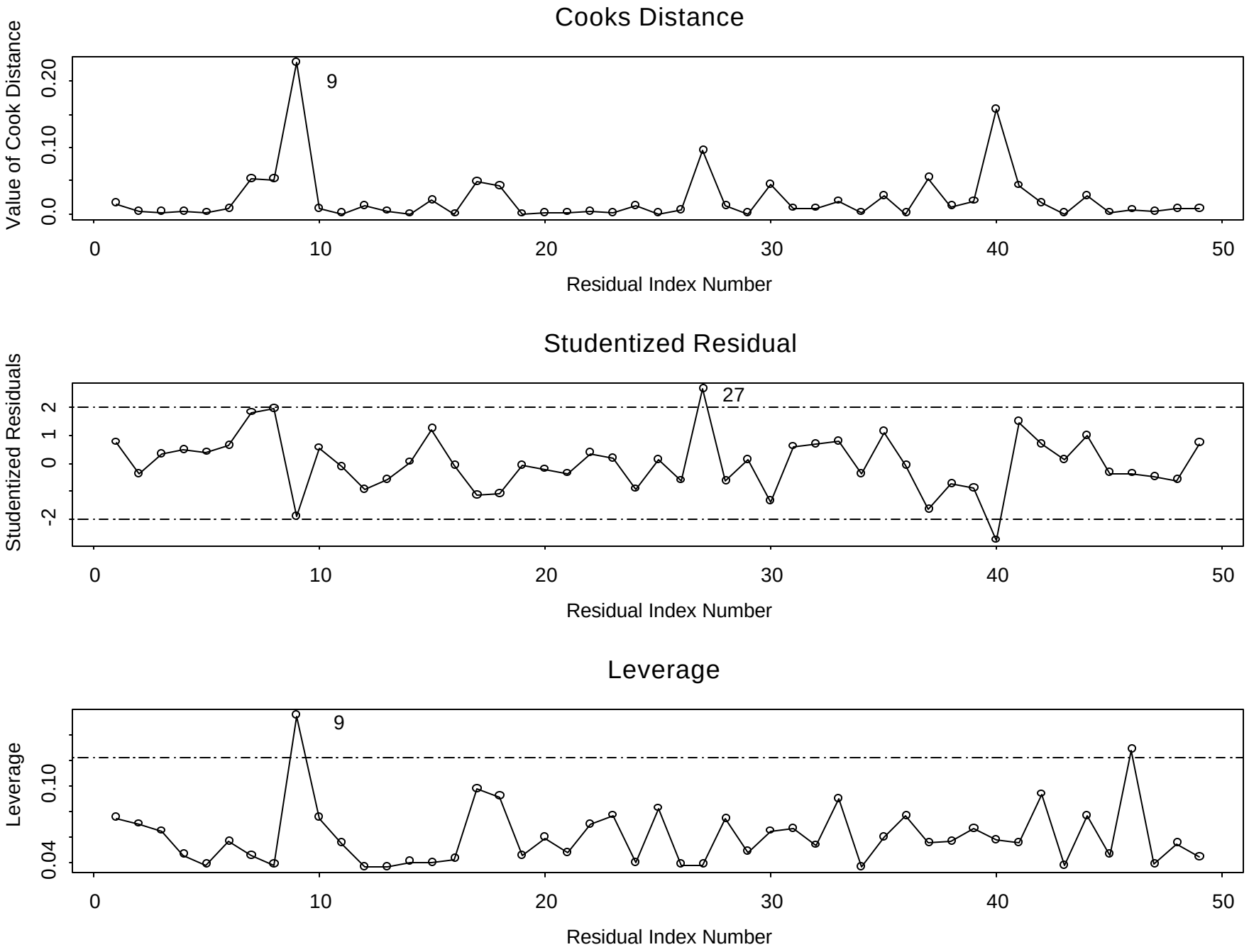
Table 6. 95% confidence intervals of the regression parameters.

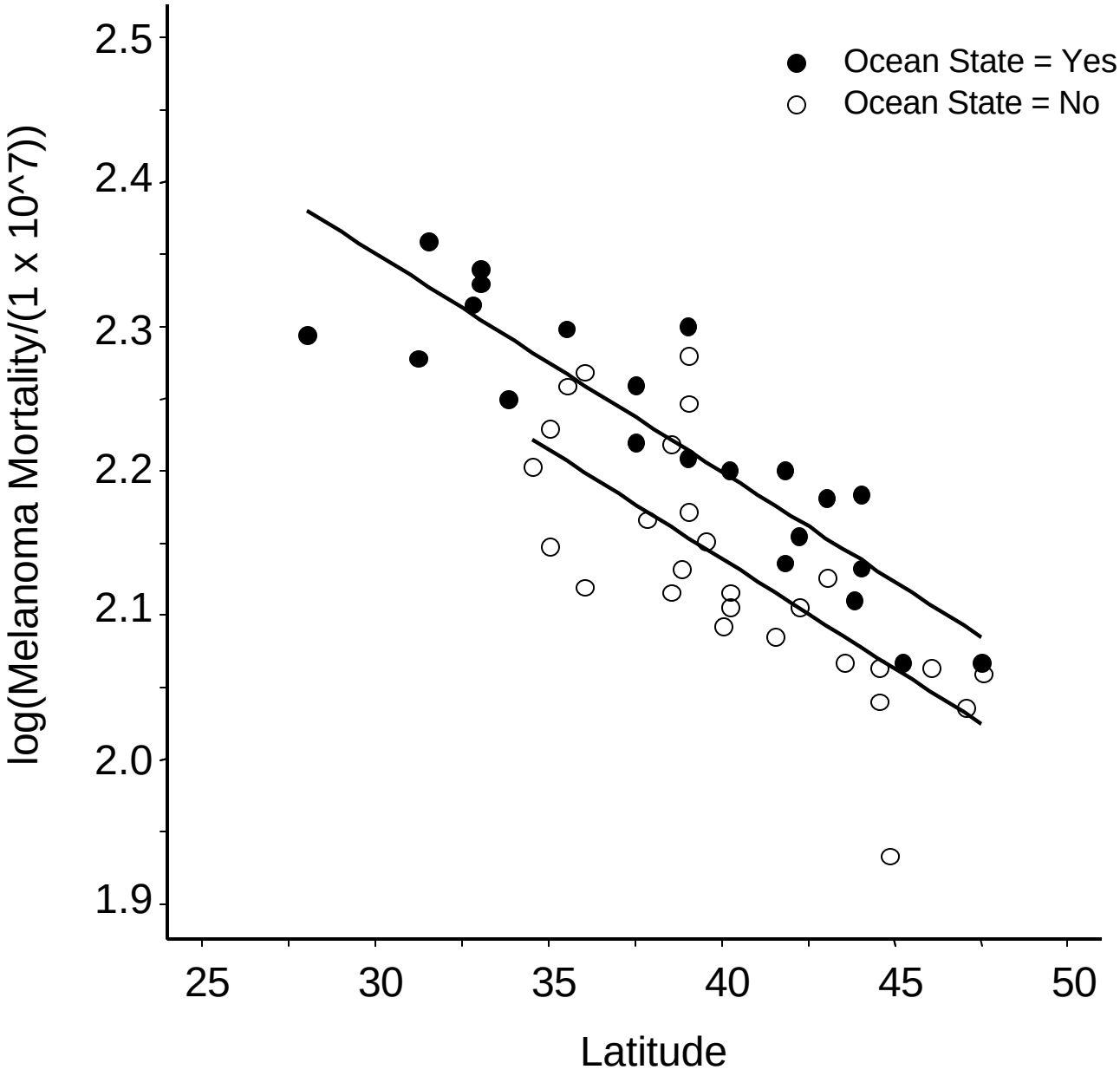
| Parameter | Estimate | SE b | df | $t_{(46,0.975)}$ | Lower 95%CL | Upper 95%CL |
|---|---|---|---|---|---|---|
| Intercept | 2.745 | 0.0630 | | | | |
| Ocean=yes | 0.060 | 0.0143 | 46 | 2.01 | 0.031 | 0.088 |
| Latitude | -0.015 | 0.0014 | 46 | 2.01 | -0.018 | -0.012 |

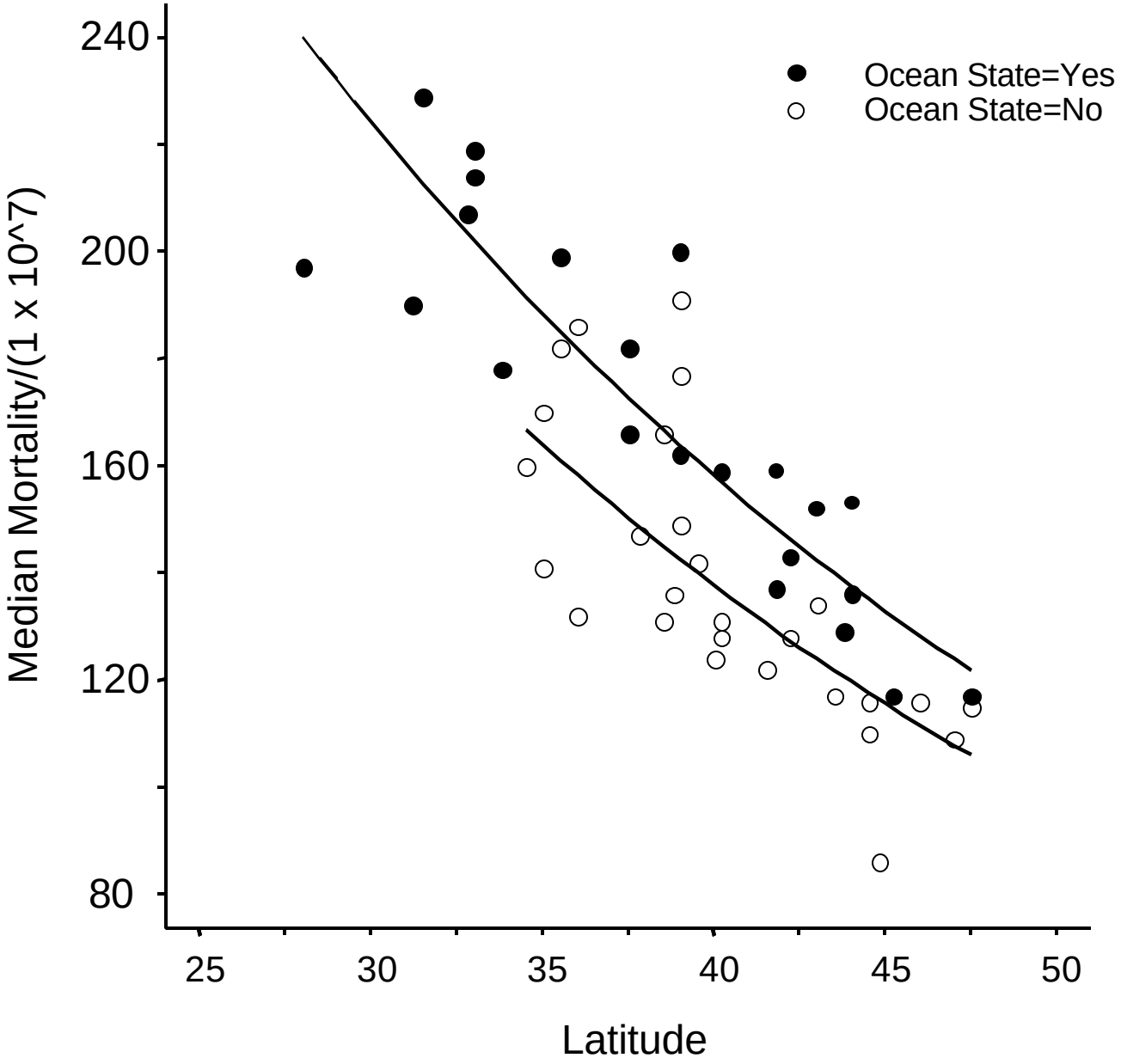
# Residual Diagnostics




Cooks Distance
Value of Cook Distance
0.0
0.10
0.20
9
0
10
20
30
40
50
Residual Index Number
Studentized Residual
Studentized Residuals
-2
0
1
2
27
0
10
20
30
40
50
Residual Index Number
Leverage
Leverage
0.04
0.10
9
0
10
20
30
40
50
Residual Index Number


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


Ocean State=Yes
Ocean State=No
Median Mortality/(1 x 10^7)
240
200
160
120
80
25
30
35
40
45
50
Latitude

Regression Analysis Summary

Melanoma mortality on the log scale was negatively related to latitude ($p<0.001$) and was higher among those states that border an ocean ($p<0.001$). The predicted melanoma mortality increased by 0.15 [95%CL(0.12, 0.18)] units 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 an ocean.

# Simple-Linear Correlation

# Simple Linear Correlation.

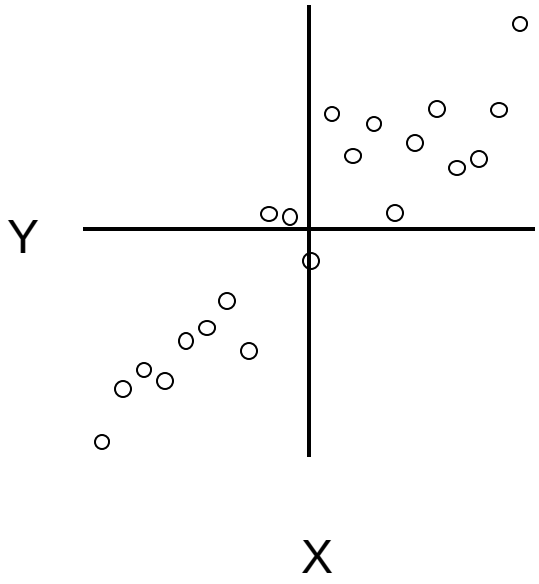
The distinguishing feature between simple-linear regression and simple-linear correlation is that in the simple linear regression setting the outcome variable is clearly defined and the independent variable is considered non-random, whereas in the correlation setting either of the two variables can be thought of as the outcome variable and both variables are assumed to be random.

I) The Pearson Correlation Coefficient *r*.

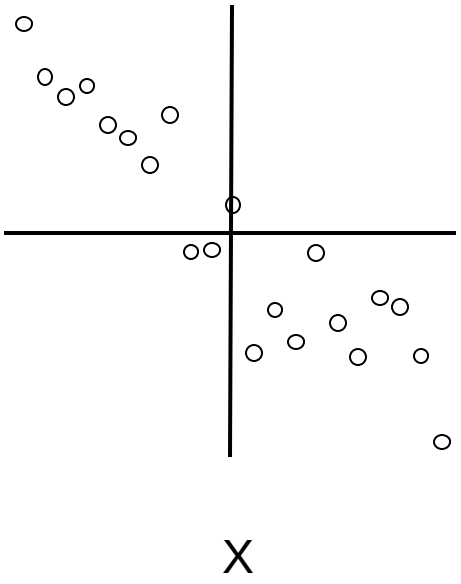
$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (-1 \leq r \geq 1)$$

- If the correlation *r* is greater than 0 then the variables X and Y are said to be positively correlated.
- If the correlation *r* is less than 0 then the variables X and Y are said to be negatively correlated.
- If the correlation *r* is exactly 0 then the variables X and Y are said to be uncorrelated.

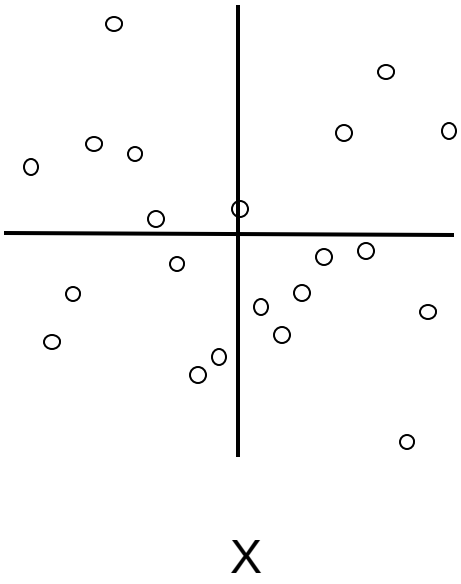
# Correlation




Positive




Negative




No Relationship

## II) Statistical Inference for the Correlation Coefficient.

The sample correlation coefficient $r$ is an estimator for the population correlation coefficient $\rho$.

To test the null hypothesis Ho: $\rho = 0$ versus Ha: $\rho \neq 0$ we use a one-sample *t*-Test.

$$t_{obs} = \frac{r(n-2)^{1/2}}{(1-r^2)^{1/2}}$$

which under Ho follows a *t* distribution with n-2 degrees of freedom.

For a two sided level $\alpha$ test, we reject Ho if $|t_{obs}| \geq t_{(n-2,1-\alpha^*/2)}$.

# Case Study #3

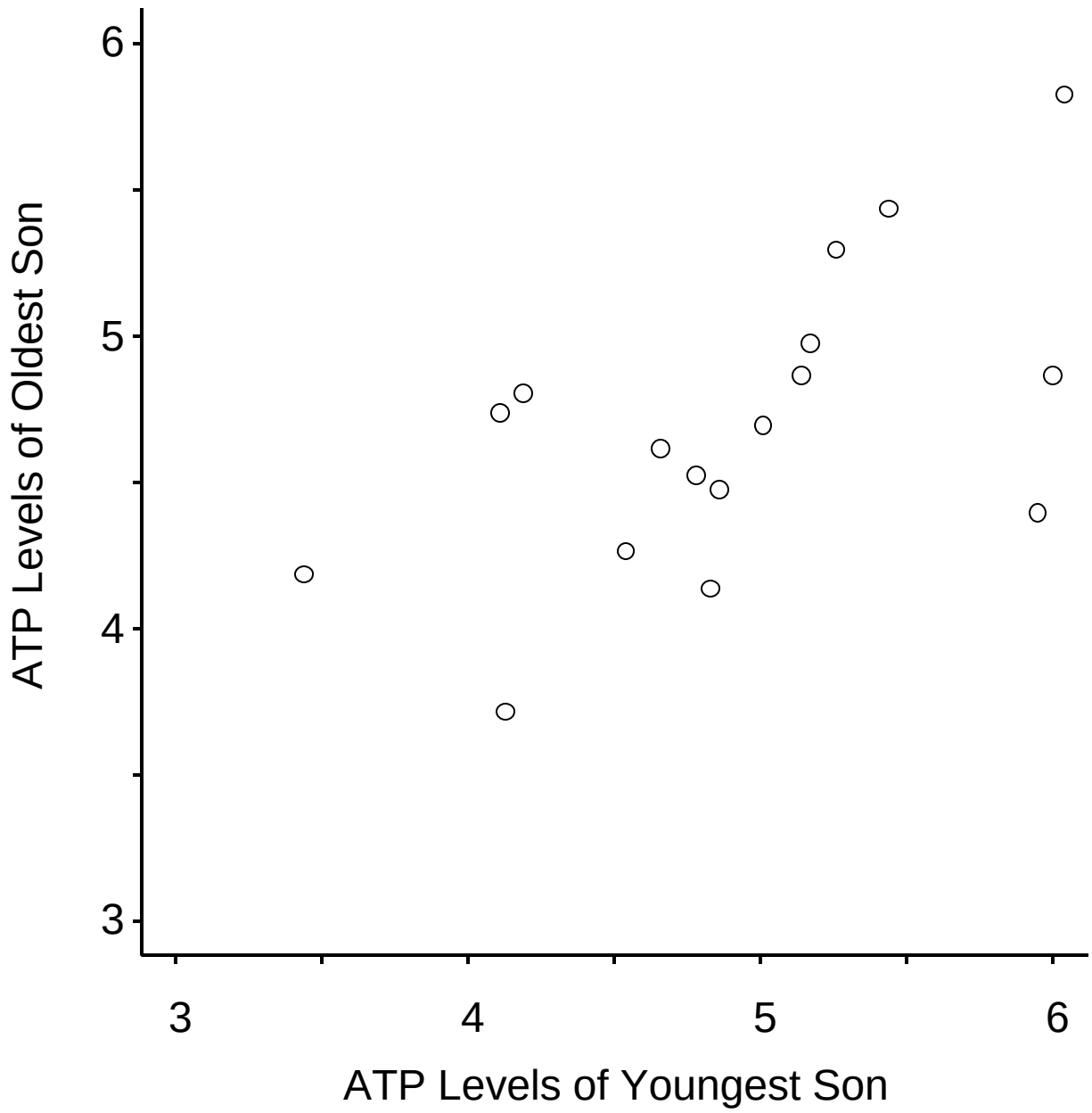
The data relate to the erythorcyte adenisone triphosphate (ATP) levels in the youngest and the oldest sons in 17 families. The investigators had previously found considerable subject to subject variation in red blood cell ATP level and they believed that some of the variation could be attributed to genetic variation. If the investigators' assumption is correct the ATP level between sibs should be positively correlated.

Fisher et. al 1993.

# Example: ATP DATA.

Erythrocyte ATP levels (mmoles/g Hb) from 17 pair of male sibs.

| Family | Age | ATP Youngest Son | Age | ATP Oldest Son |
|---|---|---|---|---|
| 1 | 24 | 4.18 | 41 | 4.81 |
| 2 | 25 | 5.16 | 26 | 4.98 |
| 3 | 19 | 4.85 | 27 | 4.48 |
| 4 | 28 | 3.43 | 32 | 4.19 |
| 5 | 22 | 4.53 | 25 | 4.27 |
| 6 | 7 | 5.13 | 23 | 4.87 |
| 7 | 21 | 4.10 | 24 | 4.74 |
| 8 | 17 | 4.77 | 25 | 4.53 |
| 9 | 25 | 4.12 | 26 | 3.72 |
| 10 | 24 | 4.65 | 25 | 4.62 |
| 11 | 12 | 6.03 | 25 | 5.83 |
| 12 | 16 | 5.94 | 24 | 4.40 |
| 13 | 9 | 5.99 | 22 | 4.87 |
| 14 | 18 | 5.43 | 24 | 5.44 |
| 15 | 14 | 5.00 | 26 | 4.70 |
| 16 | 24 | 4.82 | 26 | 4.14 |
| 17 | 20 | 5.25 | 24 | 5.30 |


ATP Levels of Oldest Son
6
5
4
3
3
4
5
6
ATP Levels of Youngest Son

a) Pearson's Sample Correlation: ATP Data

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} = \frac{3.52}{\sqrt{8.23x4.22}} = 0.60$$

b) One-Sample Tests.

Hypothesis of Ho: $\rho=0$ versus Ha:$\rho\neq 0$

$$t_{obs} = \frac{0.60(17-2)^{1/2}}{(1-0.60^2)^{1/2}} = 2.90$$

$$t_{(15,.975)} = 2.13$$

$t_{obs} > 2.13$ which $\Rightarrow$ we should reject Ho.

III) Interval Estimation for $\rho$.

Fisher's z transformation of $r = z_{obs} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$

Let

$z_o = $ Fisher's z tranformation $p = \frac{1}{2}\ln\left(\frac{1+\tilde{n}}{1-\tilde{n}}\right)$

A two sided 100% x (1- alpha) CI is given $z_0 = (z_1, z_2)$ where

$z_1 = z_{obs} - z_{(1-\acute{a}/2)}/\sqrt{n-3}$

$z_2 = z_{obs} + z_{(1-\acute{a}/2)}/\sqrt{n-3}$

A two - sided 100%(1- alpha) confidence interval for ñ is then give by $(\tilde{n}_1,\tilde{n}_2)$ where

$$\tilde{n}_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$$

$$\tilde{n}_1 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$$

- Computation of 95% CI: ATP Data.

$$z_{obs} = \frac{1}{2}\ln\left(\frac{1+0.60}{1-0.60}\right) = 0.69$$

$$z_1 = 0.69 - 1.96/\sqrt{17-3} = 0.17$$

$$z_2 = 0.69 + 1.96/\sqrt{17-3} = 1.21$$

$$p_1 = \frac{e^{2(0.17)}-1}{e^{2(0.17)}+1} = 0.17$$

$$p_2 = \frac{e^{2(1.21)}-1}{e^{2(1.21)}+1} = 0.84$$

$$95\%\ \text{CI} \quad (0.17 < \tilde{n} \leq 0.84)$$

## Analysis Summary

The red blood cell ATP level of the youngest and the oldest sib was found to be positively correlated ($r$=0.60, $p<0.05$) [95%CL for $\rho$ (0.17,0.84)].

IV) Relationship between the simple linear regression slope parameter estimate (b) and *r*.

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} \qquad b = \frac{L_{xy}}{L_{xx}} \text{ which} \Rightarrow b = r\sqrt{\frac{L_{yy}}{L_{xx}}}$$

Example : ATP Data.

$$b = 0.60\sqrt{\frac{4.22}{8.23}} = 0.42$$

## Suggested Literature Resources.


1) Fisher L.D., van Belle G. Biostatistics: A Methodology for the Health Sciences 1998. John Wiley @ Sons, NY.

2) Neter J., Kutner M.H., Nachtsheim C., Wasserman, W. Applied Linear Statistical Models:Fourth Edition. 1996. IRWIN Publishing Chicago, Ill.

3) Rosner B. Fundamentals of Biostatistics: Fifth Edition. 2000. Duxbury Press. Pacific Grove, CA.


## Web Links

http://hesweb1.med.virginia.edu/biostat/teaching/handouts.html