Regression Analysis

Adopted from:

Lecture Note, Prof. Anthony Tung, School of Computing, NUS

http://sites.stat.psu.edu/~lsimon/stat501/sp03/handouts/index.htm

https://onlinecourses.science.psu.edu/stat501/node/318

Overview

- Simple linear regression
  - Prediction concerning Y
  - Analysis of variance table
  - The general linear test
  - The lack of fit test
  - Transformations

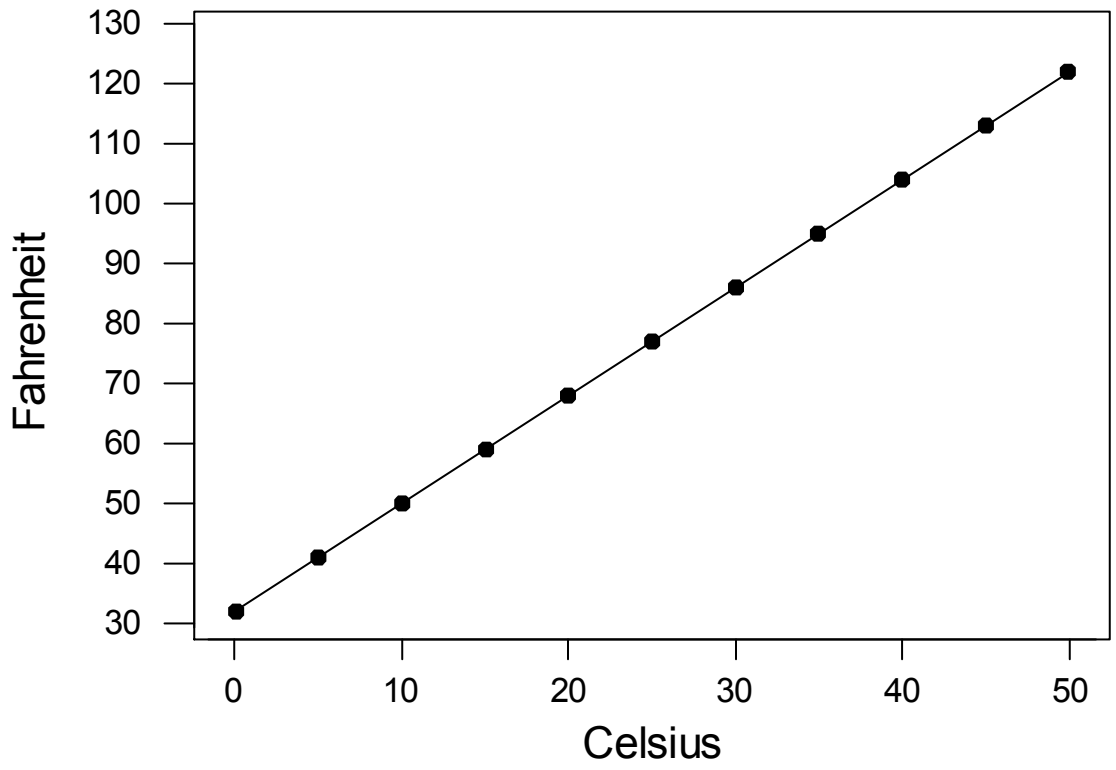
Simple linear regression

Linear regression with one predictor variable

What is simple linear regression?

- A way of evaluating the relationship between two continuous variables.
- One variable is regarded as the **predictor**, **explanatory**, or **independent** variable (x).
- Other variable is regarded as the **response**, **outcome**, or **dependent** variable (y).

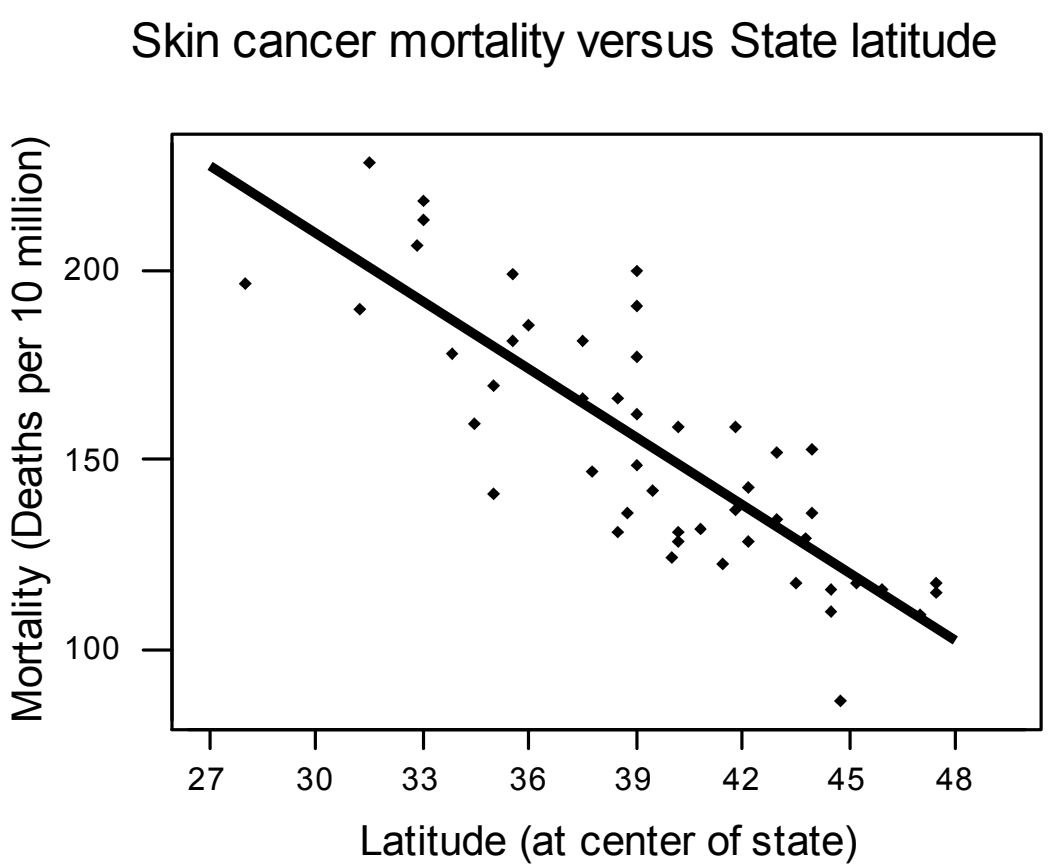
A deterministic (or functional) relationship



Other deterministic relationships

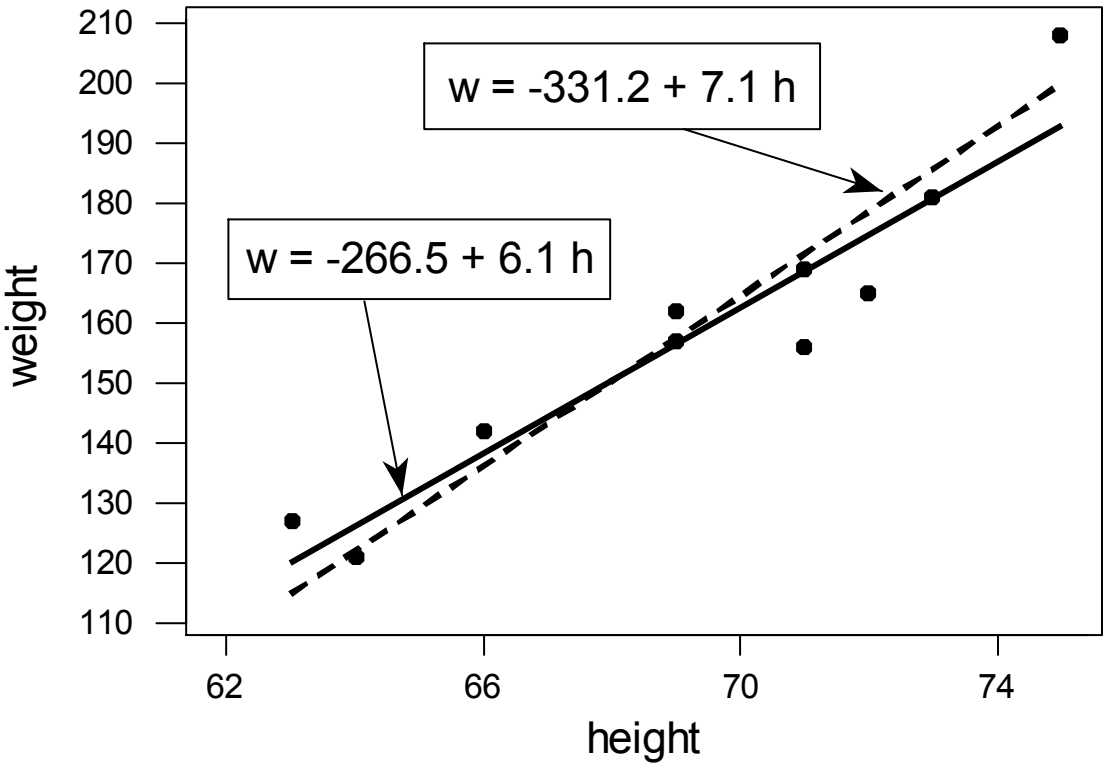
- Circumference = $\pi \times$diameter
- Hooke's Law: $Y = \alpha + \beta X$, where Y = amount of stretch in spring, and X = applied weight.
- Ohm's Law: $I = V/r$, where V = voltage applied, r = resistance, and I = current.
- Boyle's Law: For a constant temperature, $P = \alpha/V$, where P = pressure, α = constant for each gas, and V = volume of gas.

A statistical relationship



A relationship with some "**trend**", but also with some "**scatter**."

Which is the “best fitting line”?



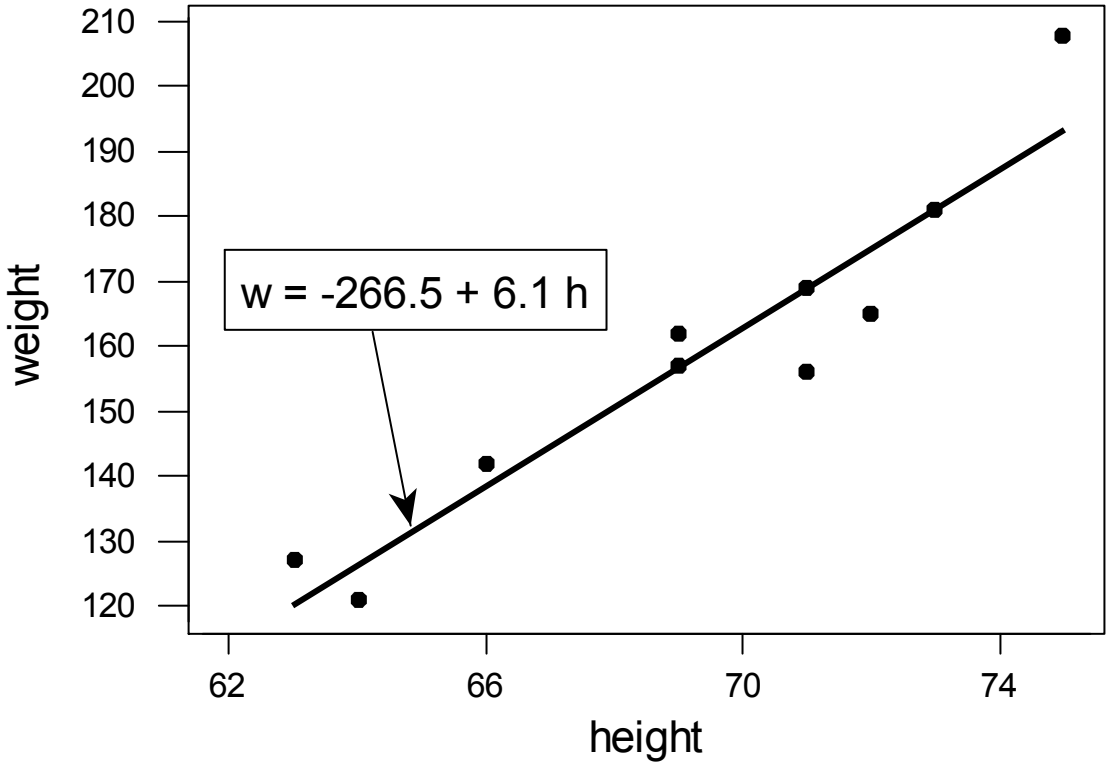
Notation

y_i is the **observed response** for the ith experimental unit.

x_i is the **predictor value** for the ith experimental unit.

$\hat{y}_i$ is the **predicted response** (or **fitted value**) for the ith experimental unit.

Equation of best fitting line: $\hat{y}_i = b_0 + b_1 x_i$



i	x_i	y_i	$\hat{y}_i$
1	64	121	126.3
2	73	181	181.5
3	71	156	169.2
4	69	162	157.0
5	66	142	138.5
6	69	157	157.0
7	75	208	193.8
8	71	169	169.2
9	63	127	120.1
10	72	165	175.4

Prediction error (or residual error)

In using $\hat{y}_i$ to predict the actual response y_i we make a **prediction error** (or a **residual error**) of size $e_i = y_i - \hat{y}_i$

A line that fits the data well will be one for which the n prediction errors are as small as possible in some overall sense.

The "**least squares criterion**"

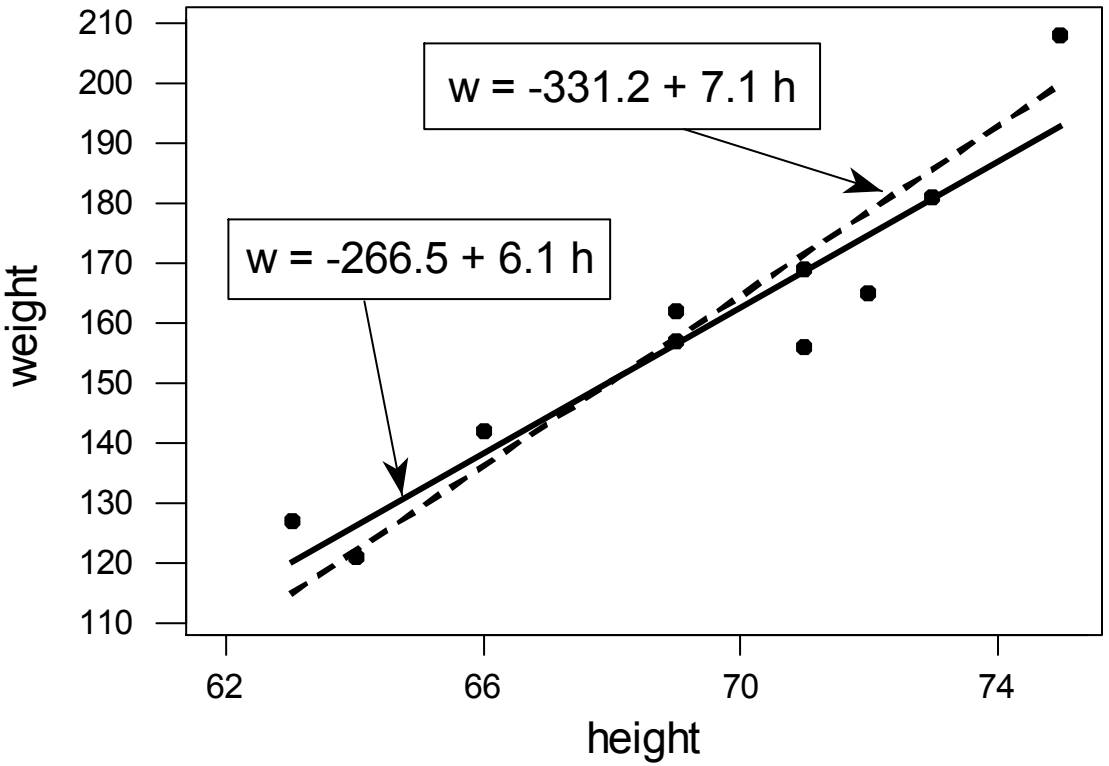
Equation of best fitting line: $\hat{y}_i = b_0 + b_1 x_i$

Choose the values b_0 and b_1 that minimize the sum of the squared prediction errors.

That is, find b_0 and b_1 that minimize:

$$Q = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

Which is the “best fitting line”?



w = -331.2 + 7.1 h

i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	64	121	123.2	-2.2	4.84
2	73	181	187.1	-6.1	37.21
3	71	156	172.9	-16.9	285.61
4	69	162	158.7	3.3	10.89
5	66	142	137.4	4.6	21.16
6	69	157	158.7	-1.7	2.89
7	75	208	201.3	6.7	44.89
8	71	169	172.9	-3.9	15.21
9	63	127	116.1	10.9	118.81
10	72	165	180.0	-15.0	225.00

					766.51

w = -266.5 + 6.1 h

i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	64	121	126.271	-5.3	28.09
2	73	181	181.509	-0.5	0.25
3	71	156	169.234	-13.2	174.24
4	69	162	156.959	5.0	25.00
5	66	142	138.546	3.5	12.25
6	69	157	156.959	0.0	0.00
7	75	208	193.784	14.2	201.64
8	71	169	169.234	-0.2	0.04
9	63	127	120.133	6.9	47.61
10	72	165	175.371	-10.4	108.16

					597.28

The least squares regression line

Using calculus, minimize (take derivative with respect to b_0 and b_1, set to 0, and solve for b_0 and b_1):

$$Q = \sum_{i=1}^{n} \left(y_i - \left(b_0 + b_1 x_i \right) \right)^2$$

and get the least squares estimates b_0 and b_1:

$$b_1 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right)}{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Regression analysis

The regression equation is
weight = - 267 + 6.14 height

Predictor	**Coef**	SE Coef	T	P
Constant	**-266.53**	51.03	-5.22	0.001
height	**6.1376**	0.7353	8.35	0.000

S = 8.641 R-Sq = 89.7% R-Sq(adj) = 88.4%

Analysis of Variance

Source	DF	**SS**	MS	F	P
Regression	1	5202.2	5202.2	69.67	0.000
Residual Error	8	**597.4**	74.7		
Total	9	5799.6			

Prediction of future responses

A common use of the estimated regression line.

$$\hat{\boldsymbol{y}}_{\boldsymbol{i},\boldsymbol{wt}} = -267 + 6.14\boldsymbol{x}_{\boldsymbol{i},\boldsymbol{ht}}$$

Predict mean weight of 66"-inch tall people.

$$\hat{\boldsymbol{y}}_{\boldsymbol{i},\boldsymbol{wt}} = -267 + 6.14(66) = 138.24$$

Predict mean weight of 67"-inch tall people.

$$\hat{\boldsymbol{y}}_{\boldsymbol{i},\boldsymbol{wt}} = -267 + 6.14(67) = 144.38$$

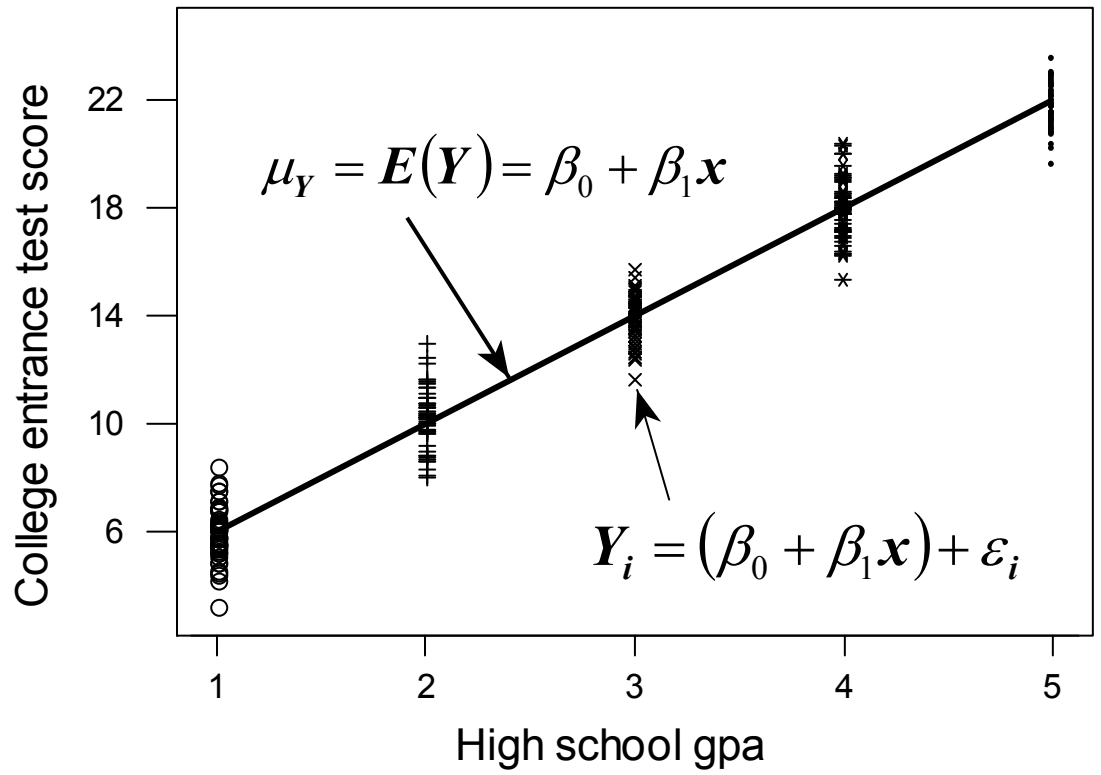
What do the "**estimated regression coefficients**" b_0 and b_1 tell us?

- We can expect the mean response to increase or decrease by b_1 units for every unit increase in x.
- If the "**scope of the model**" includes x = 0, then b_0 is the predicted mean response when x = 0. Otherwise, b_0 is not meaningful.

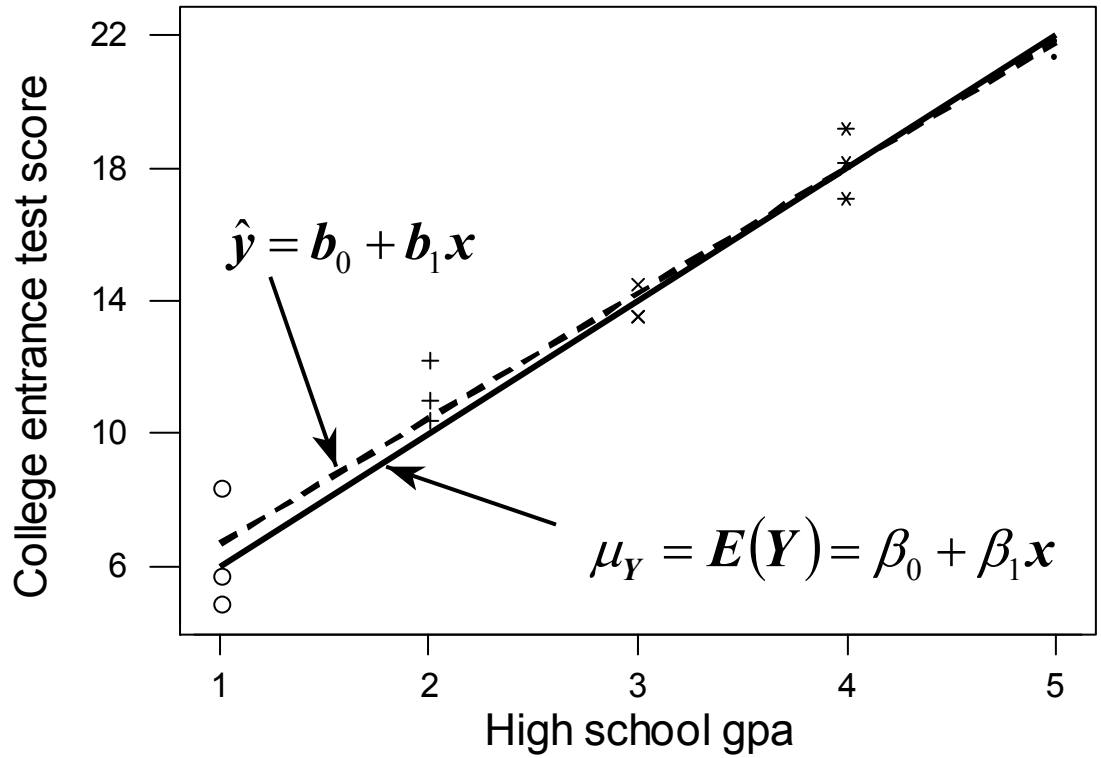
So, the estimated regression coefficients b_0 and b_1 tell us…

- We predict the mean weight to increase by 6.14 pounds for every additional one-inch increase in height.
- It is not meaningful to have a height of 0 inches. That is, the scope of the model does not include x = 0. So, here the intercept b_0 is not meaningful.

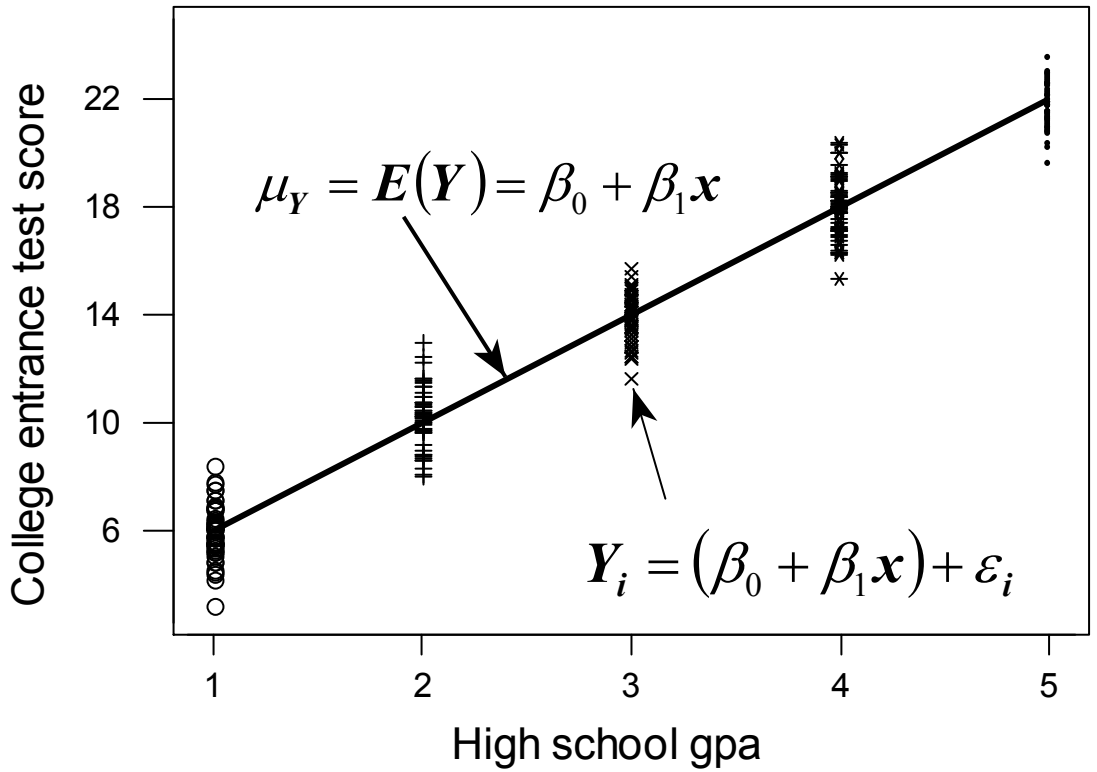
What do b_0 and b_1 estimate?



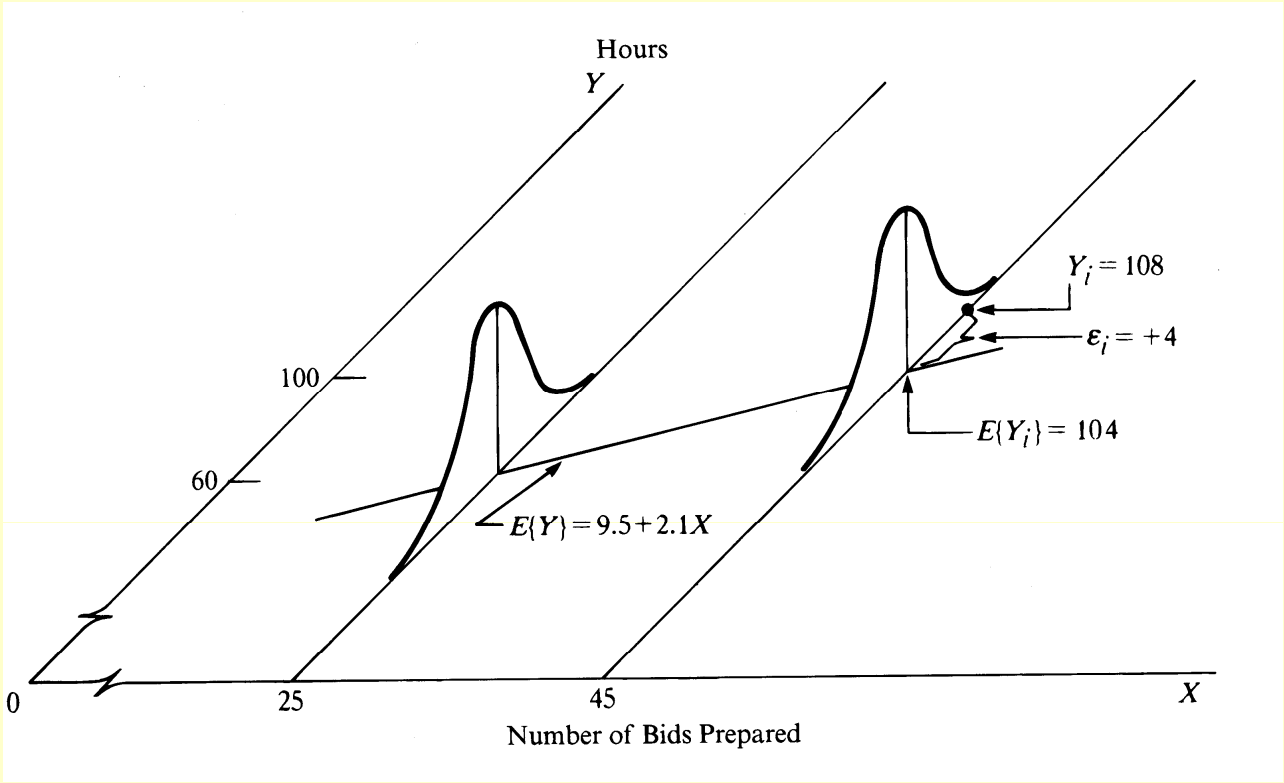
What do b_0 and b_1 estimate?



The simple linear regression model



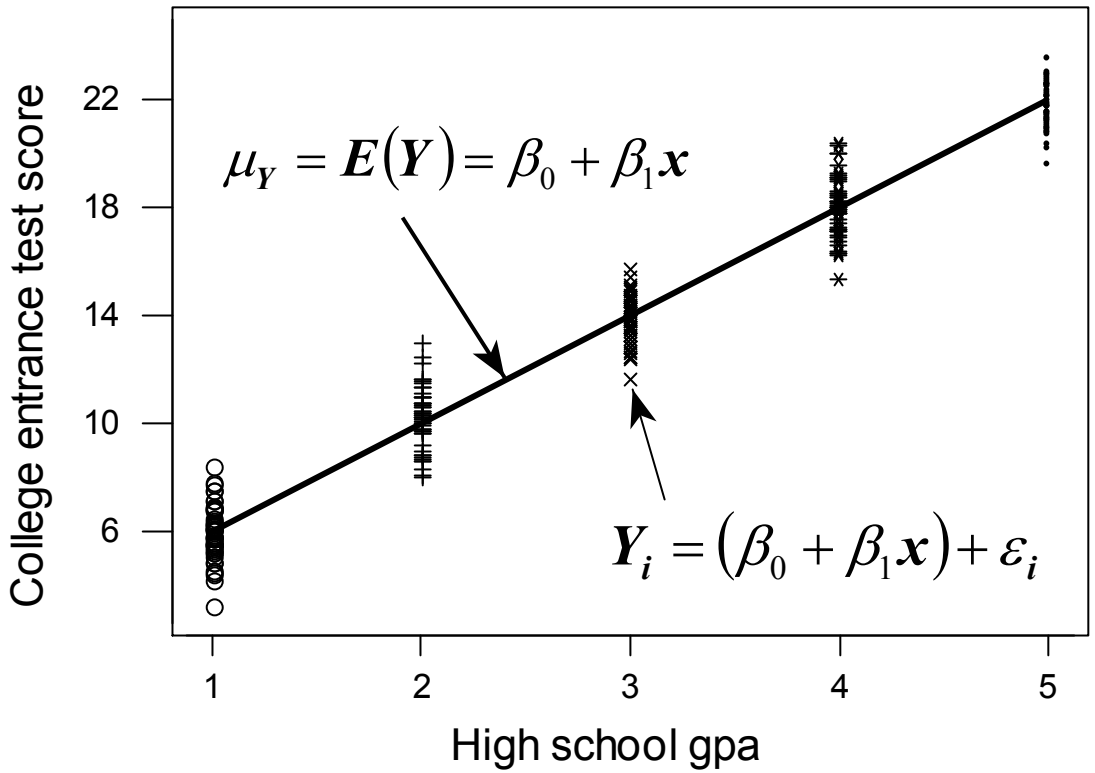
The simple linear regression model



The simple linear regression model

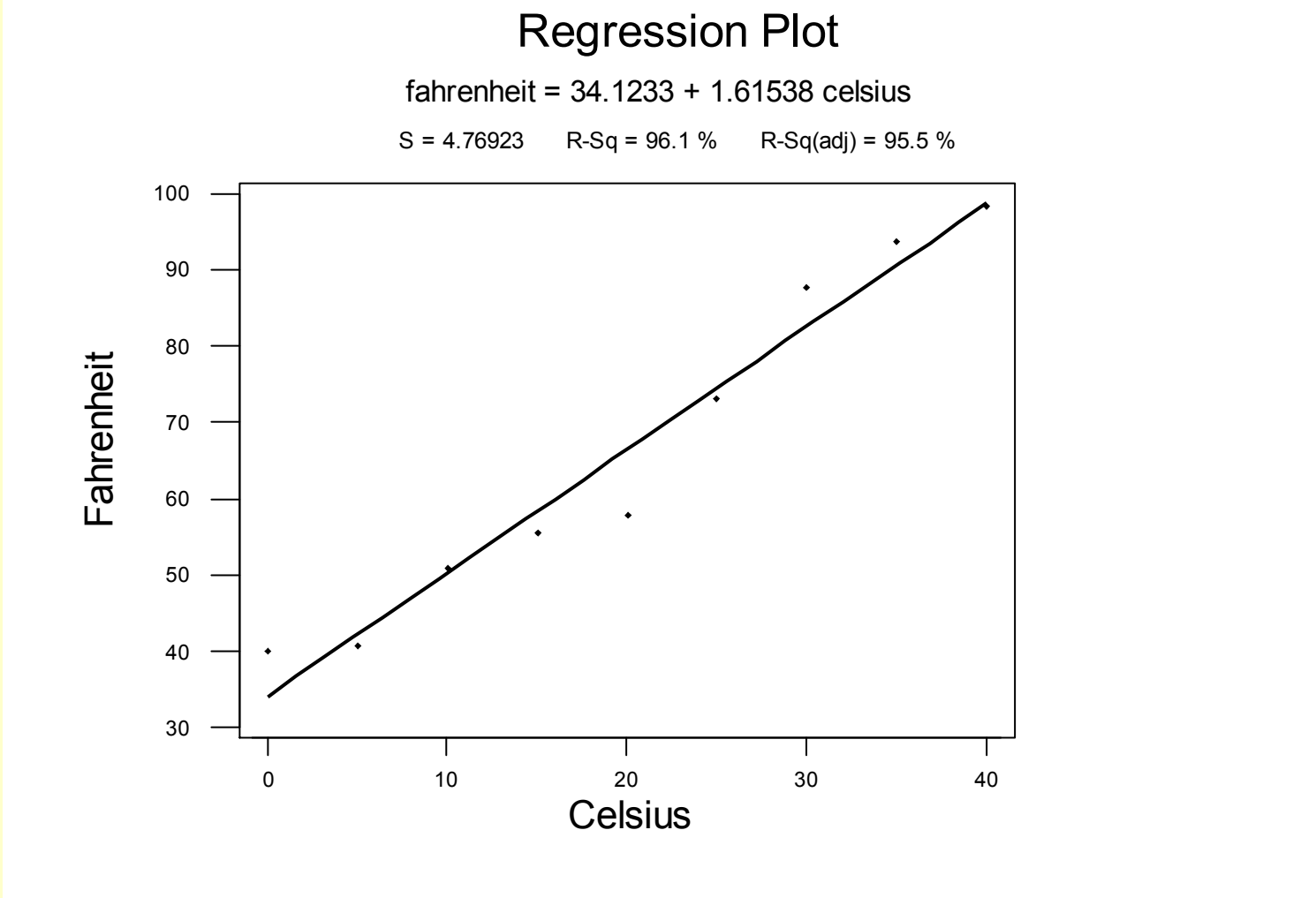
- The mean of the responses, $E(Y_i)$, is a **linear function** of the x_i.
- The errors, ε_i, and hence the responses Y_i, are **independent**.
- The errors, ε_i, and hence the responses Y_i, are **normally distributed**.
- The errors, ε_i, and hence the responses Y_i, have **equal variances** (σ^2) for all x values.

What about (unknown) σ^2?

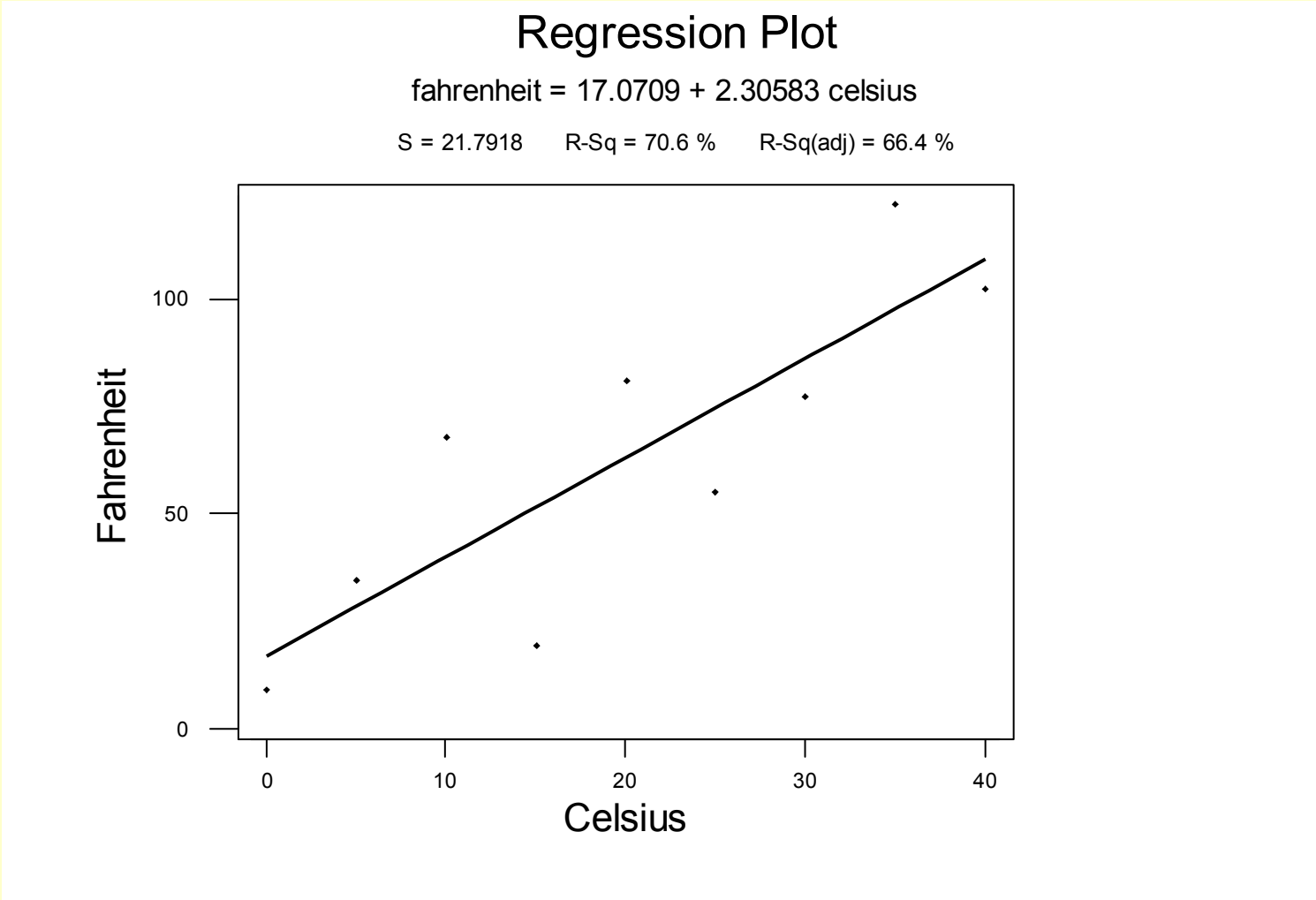


It quantifies how much the responses (y) vary around the (unknown) mean regression line $E(Y) = \beta_0 + \beta_1 x$.

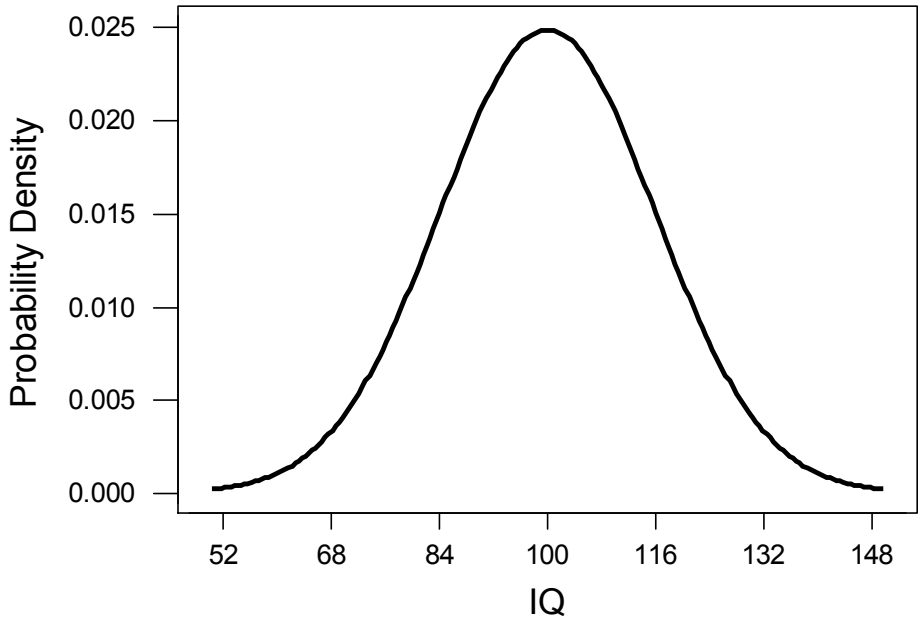
Will this thermometer yield more precise future predictions …?



… or this one?


Regression Plot
fahrenheit = 17.0709 + 2.30583 celsius
S = 21.7918 R-Sq = 70.6 % R-Sq(adj) = 66.4 %
Fahrenheit
100
50
0
0
10
20
30
40
Celsius

Recall the "sample variance"

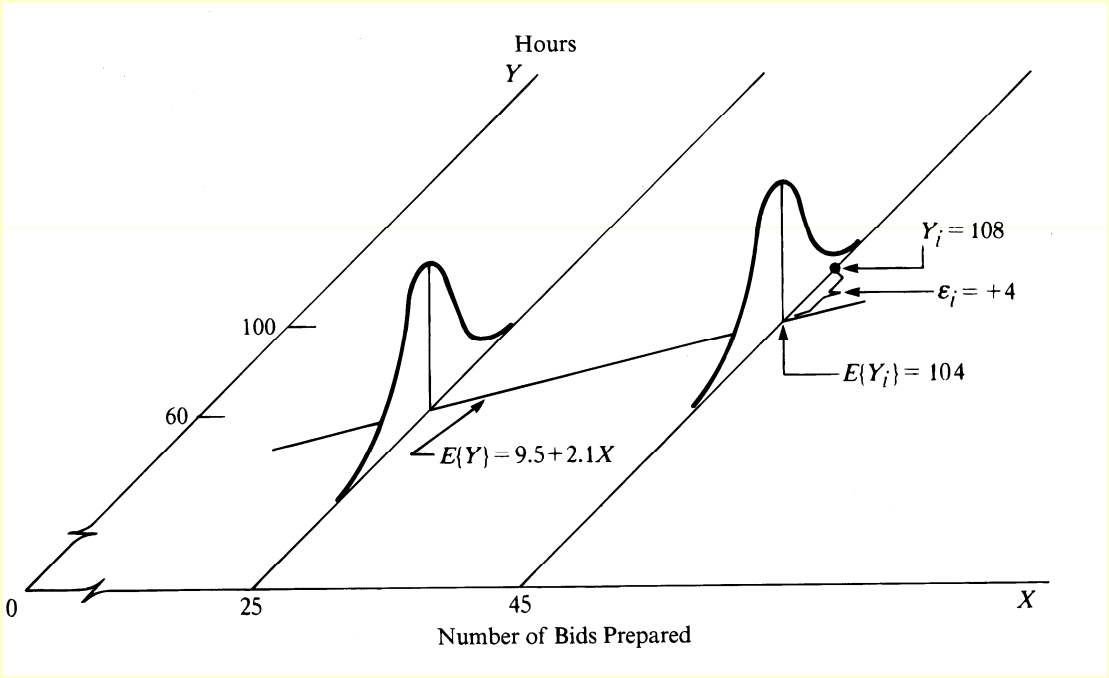


The **sample variance**

$$s^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{n-1}$$

estimates σ^2, the variance of the <u>one</u> population.

Estimating σ^2 in regression setting

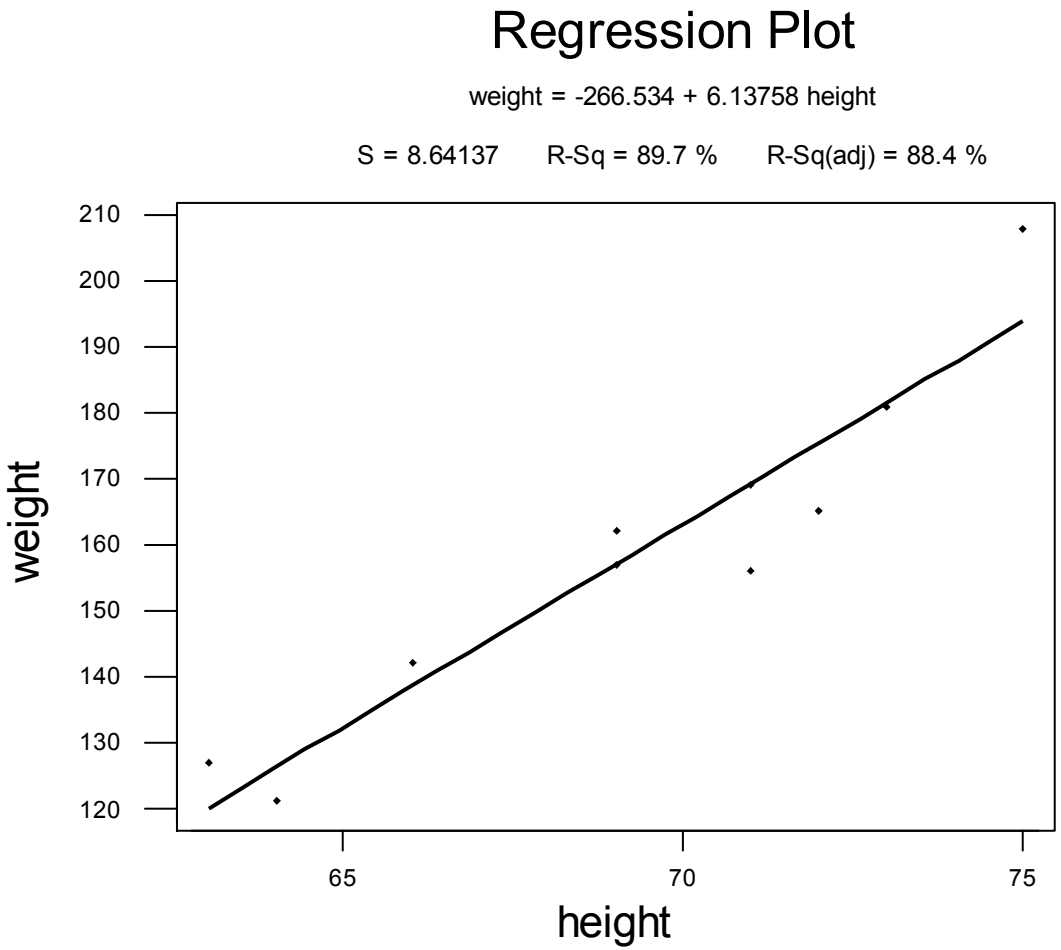


The **mean square error**

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

estimates σ^2, the common variance of the <u>many</u> populations.

Estimating σ^2 from fitted line plot



Estimating σ^2 from regression analysis

The regression equation is
weight = - 267 + 6.14 height

Predictor	Coef	SE Coef	T	P
Constant	-266.53	51.03	-5.22	0.001
height	6.1376	0.7353	8.35	0.000

S = **8.641** R-Sq = 89.7% R-Sq(adj) = 88.4%

Analysis of Variance

Source	DF	SS	**MS**	F	P
Regression	1	5202.2	5202.2	69.67	0.000
Residual Error	8	597.4	**74.7**		
Total	9	5799.6			

Inference for (or drawing conclusions about) β_0 and β_1

Confidence intervals and hypothesis tests

Relationship between state latitude and skin cancer mortality?

#	State	LAT	MORT
1	Alabama	33.0	219
2	Arizona	34.5	160
3	Arkansas	35.0	170
4	California	37.5	182
5	Colorado	39.0	149
!	□	□	□
49	Wyoming	43.0	134

- Mortality rate of white males due to malignant skin melanoma from 1950-1959.
- LAT = degrees (north) latitude of center of state
- MORT = mortality rate due to malignant skin melanoma per 10 million people

(1-α)100% t-interval for slope parameter β_1

Formula in words:

Sample estimate ± (t-multiplier × standard error)

Formula in notation:

$$b_1 \pm t_{\left(1-\alpha/2, n-2\right)} \times \left(\frac{\sqrt{MSE}}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}} \right)$$

Hypothesis test for slope parameter β_1

Null hypothesis H_0: β_1 = some number β
Alternative hypothesis H_A: $\beta_1 \neq$ some number β

Test statistic

$$t^* = \frac{b_1 - \beta}{\left(\sqrt{MSE} \Big/ \sqrt{\sum (x_i - \bar{x})^2} \right)} = \frac{b_1 - \beta}{se(b_1)}$$

P-value = How likely is it that we'd get a test statistic t* as extreme as we did if the null hypothesis is true?

The P-value is determined by referring to a **t-distribution** with **n-2** degrees of freedom.

Inference for slope parameter β_1

The regression equation is Mort = 389 - 5.98 Lat

Predictor	**Coef**	**SE Coef**	**T**	**P**
Constant	389.19	23.81	16.34	0.000
Lat	**-5.9776**	**0.5984**	**-9.99**	**0.000**

S = 19.12 R-Sq = 68.0% R-Sq(adj) = 67.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

(1-α)100% t-interval for intercept parameter β_0

Formula in words:

Sample estimate ± (t-multiplier × standard error)

Formula in notation:

$$b_0 \pm t_{\left(1-\alpha/2,\, n-2\right)} \times \sqrt{MSE} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum \left(x_i - \bar{x}\right)^2}}$$

Hypothesis test for intercept parameter β_0

Null hypothesis H_0: β_0 = some number β
Alternative hypothesis H_A: $\beta_0 \neq$ some number β

Test statistic

$$t^* = \frac{b_0 - \beta}{\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}} = \frac{b_0 - \beta}{se(b_0)}$$

P-value = How likely is it that we'd get a test statistic t* as extreme as we did if the null hypothesis is true?

The P-value is determined by referring to a **t-distribution** with **n-2** degrees of freedom.

Inference for intercept parameter β_0

The regression equation is Mort = 389 - 5.98 Lat

Predictor	**Coef**	**SE Coef**	**T**	**P**
Constant	**389.19**	**23.81**	**16.34**	**0.000**
Lat	-5.9776	0.5984	-9.99	0.000

S = 19.12 R-Sq = 68.0% R-Sq(adj) = 67.3%

Analysis of Variance

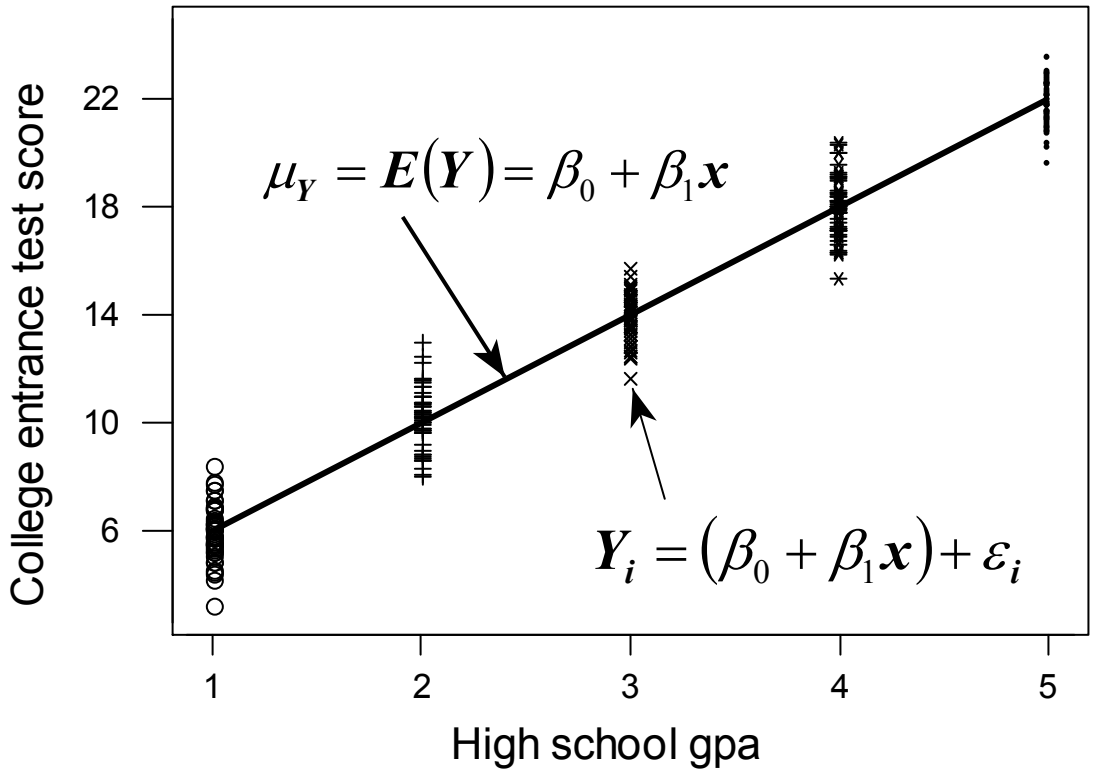
Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

What assumptions?

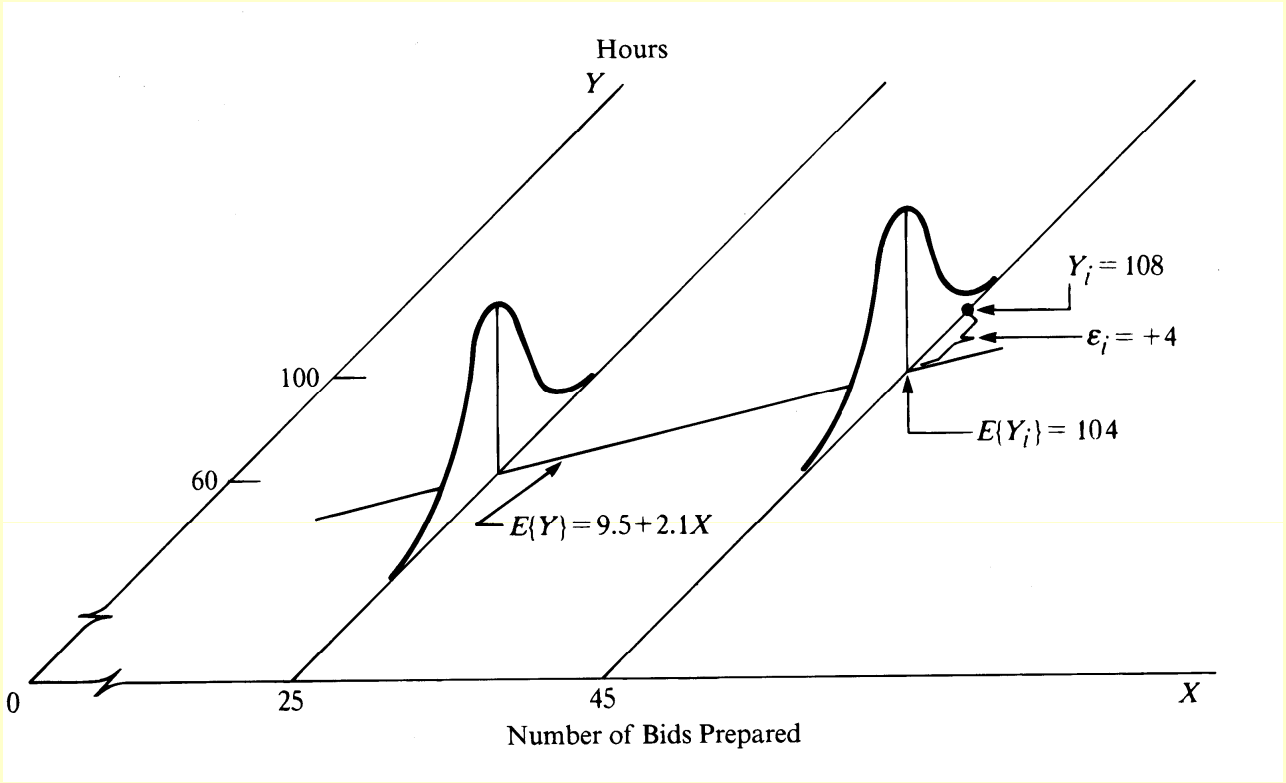
- The intervals and tests depend on the assumption that the error terms (and thus responses) follow a normal distribution.
- Not a big deal if the error terms (and thus responses) are only approximately normal.
- If have a large sample, then the error terms can even deviate far from normality.

Prediction concerning the response Y

Simple linear regression model



Simple linear regression model



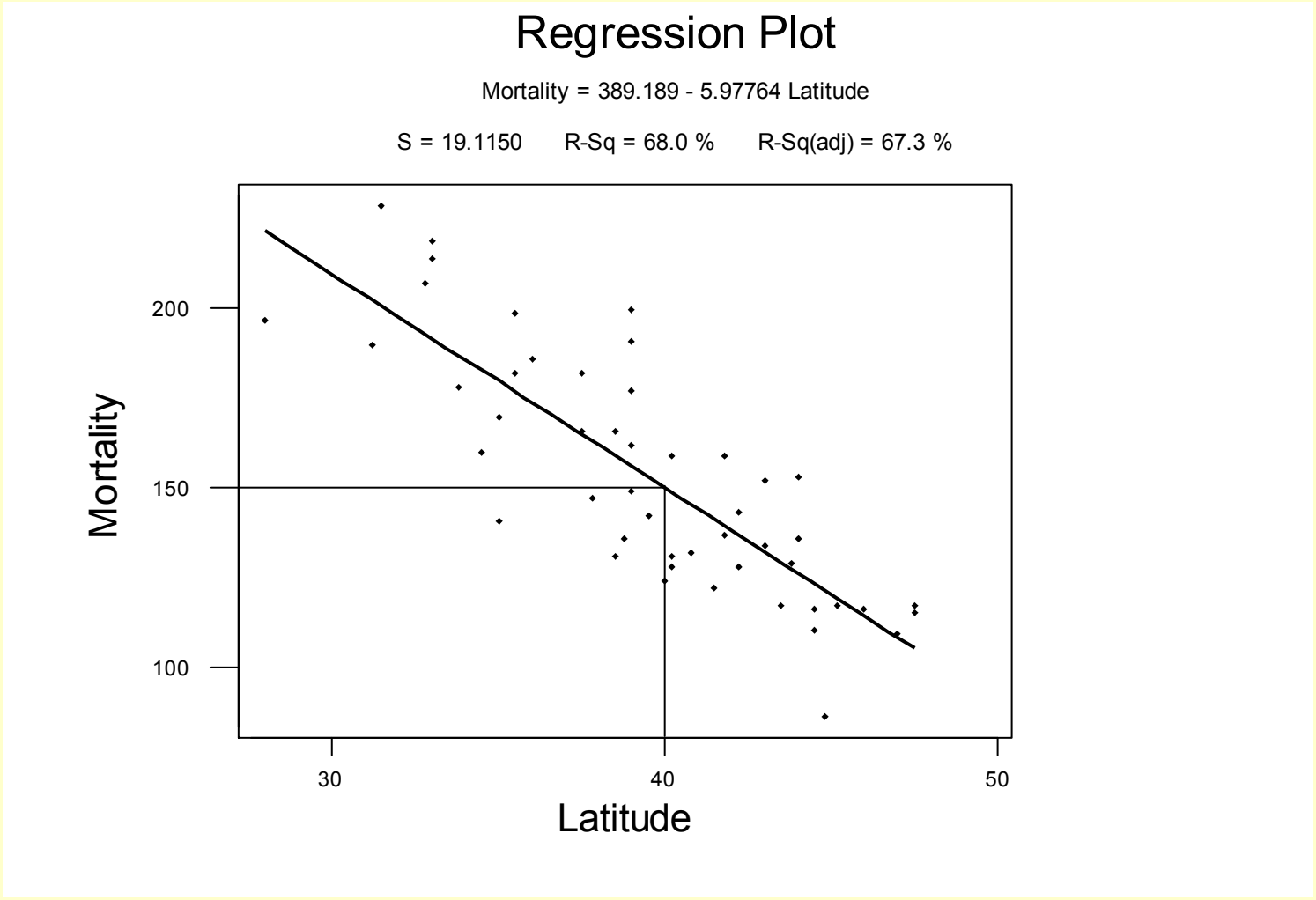
Three different research questions

- What is the **mean response**, $E(Y_h)$, for a given value, x_h, of the predictor variable?
- What would one **predict a new observation** , $Y_{h(new)}$, to be for a given value, x_h, of the predictor variable?
- What would one **predict the mean of m new observations**, $\bar{Y}_{h(new)}$, to be for a given value, x_h , of the predictor variable?

Example:
Skin cancer mortality and latitude

- What is the expected (mean) mortality rate for all locations at 40^{o} N latitude?
- What is the predicted mortality rate for 1 new randomly selected location at 40^{o} N?
- What is the predicted mortality rate for 10 new randomly selected locations at 40^{o} N?

Example:
Skin cancer mortality and latitude



"Point estimators"

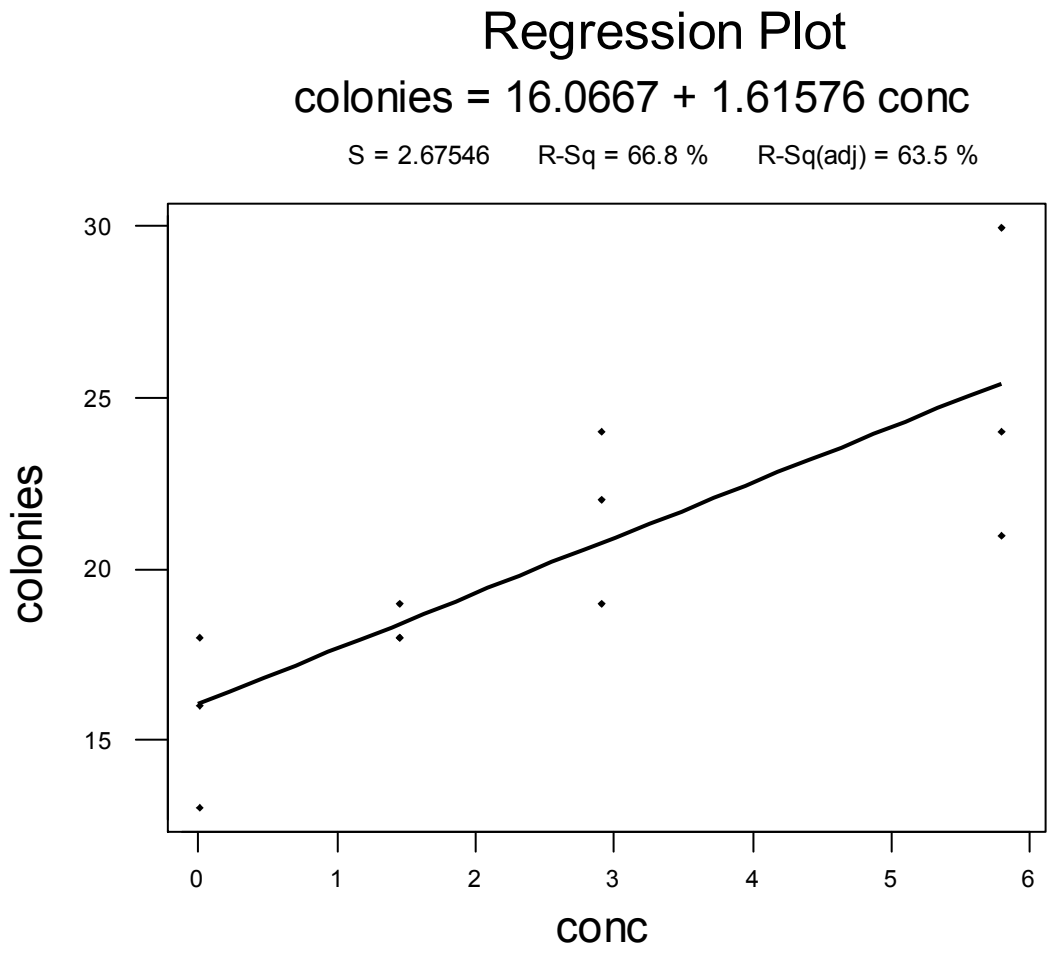
$\hat{Y}_h = b_0 + b_1 x_h$ is the best point estimator in each case.

That is, it is:

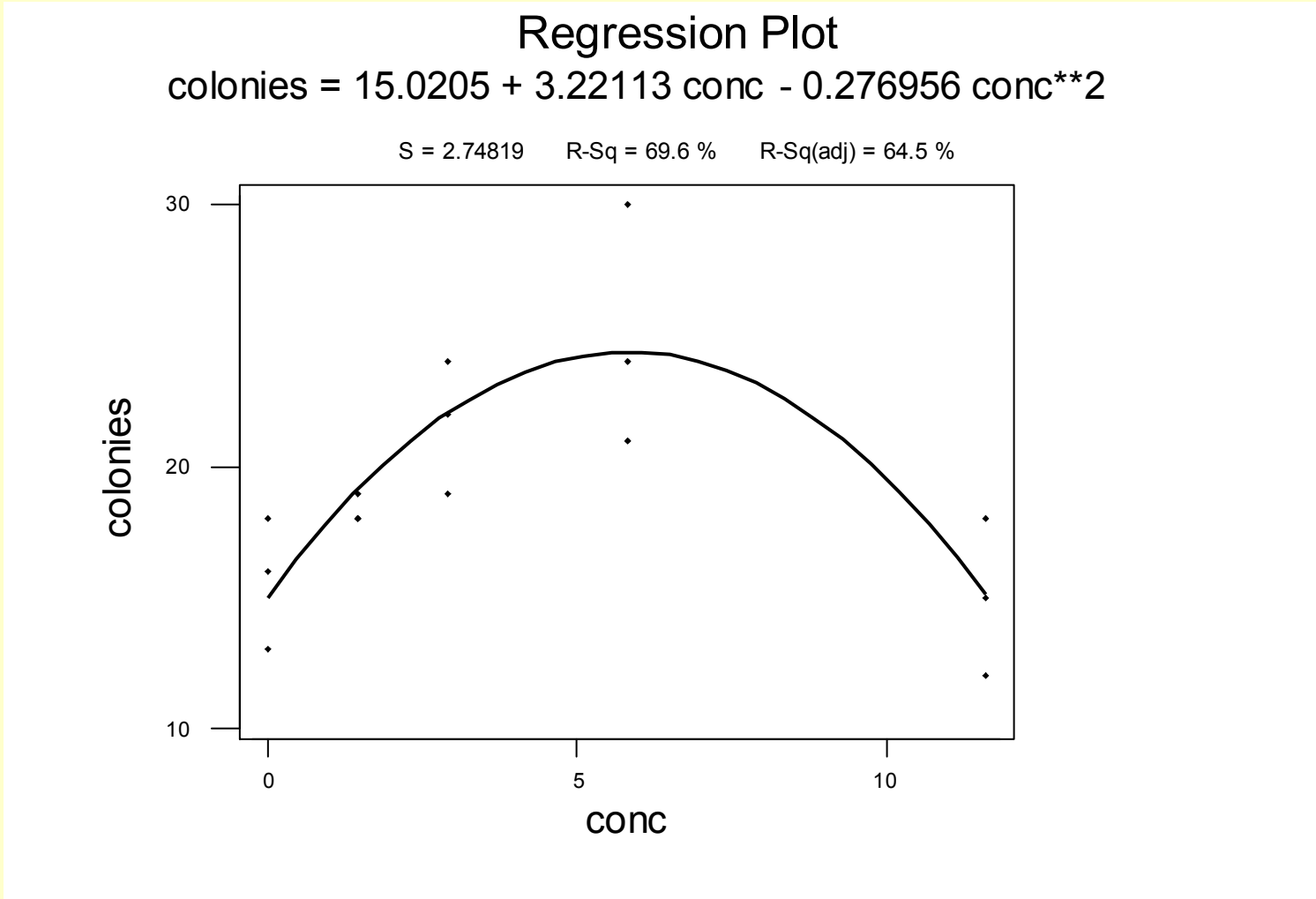
- the best guess of the mean response at x_h
- the best guess of a new observation at x_h
- the best guess of a mean of m new observations at x_h

But, as always, to be confident in the answer to our research question, we should put an interval around our best guess.

It is dangerous to "**extrapolate**" beyond scope of model.

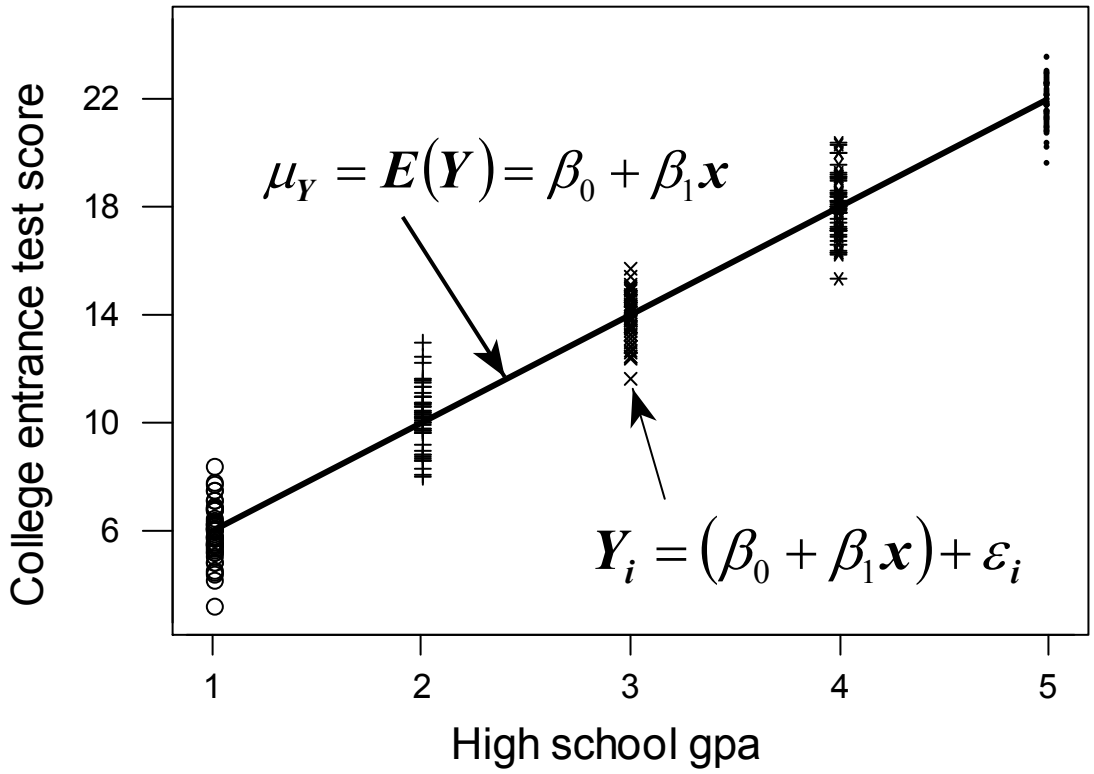


It is dangerous to "**extrapolate**" beyond scope of model.



Confidence interval for the population mean response $E(Y_h)$

Again, what are we estimating?



(1-α)100% t-interval for mean response $E(Y_h)$

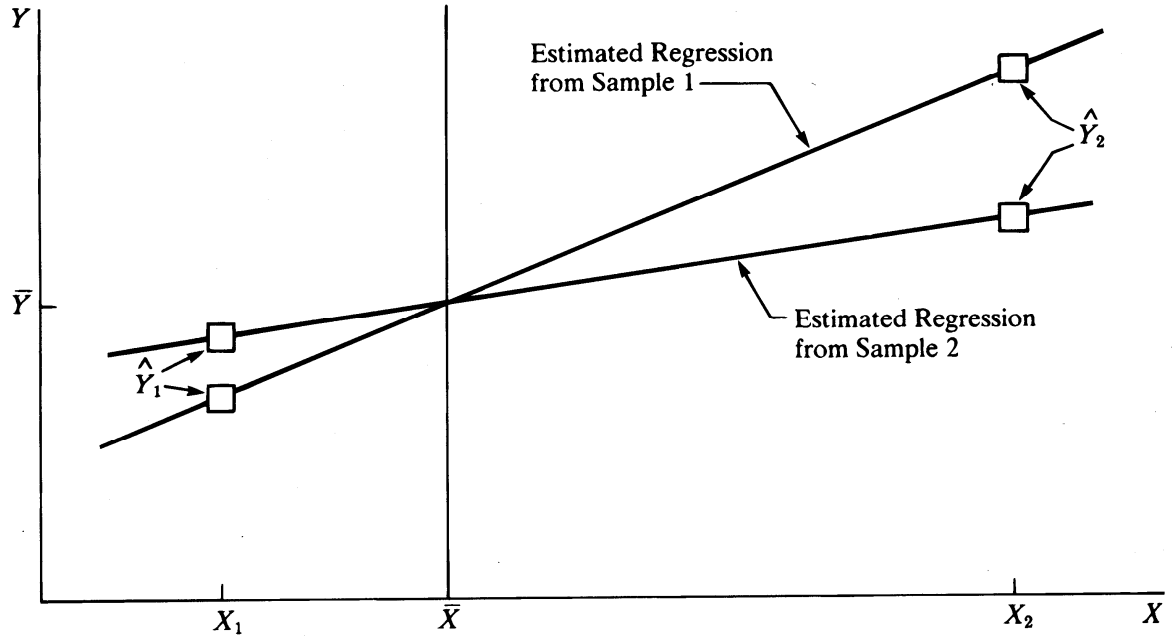
Formula in words:

Sample estimate ± (t-multiplier × standard error)

Formula in notation:

$$\hat{y}_h \pm t_{\left(1-\alpha/2, n-2\right)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{\left(x_h - \bar{x}\right)^2}{\sum \left(x_i - \bar{x}\right)^2} \right)}$$

FIGURE 2.3 **Effect on $\hat{Y}_h$ of Variation in b_1 from Sample to Sample in Two Samples with Same Means $\bar{Y}$ and $\bar{X}$.**



Implications on precision

- The greater the spread in the x_i values, the narrower the confidence interval, the more precise the prediction of $E(Y_h)$.
- Given the same set of x_i values, the further x_h is from the (sample) mean of the x_i, the wider the confidence interval, the less precise the prediction of $E(Y_h)$.

Predicted Values for New Observations

New	**Fit**	**SE Fit**	**95.0% CI**	95.0% PI
1	**150.08**	**2.75**	**(144.6,155.6)**	(111.2,188.93)
2	**221.82**	**7.42**	**(206.9,236.8)**	(180.6,263.07)X

X denotes a row with X values away from the center

Values of Predictors for New Observations

New Obs	Latitude
1	40.0
2	28.0

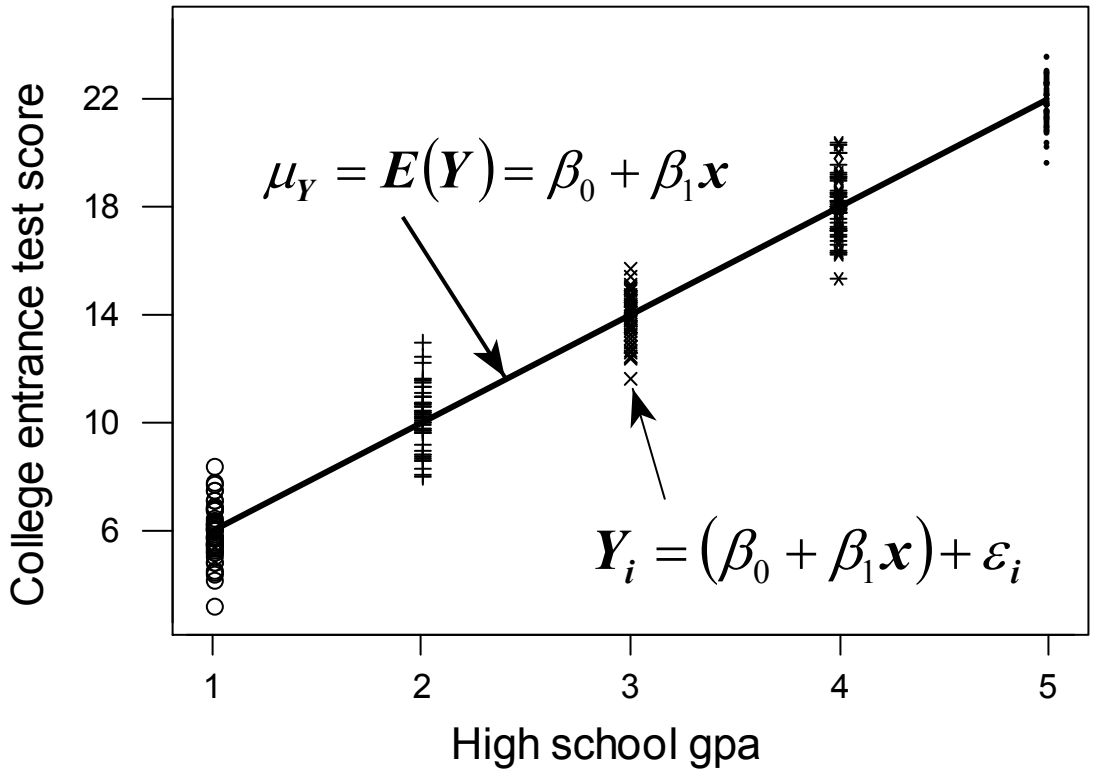
Mean of Lat = 39.533

Comments on assumptions

- x_h is a value within scope of model, but it is not necessary that it is one of the x values in the data set.
- The confidence interval formula for $E(Y_h)$ works okay even if the error terms are only approximately normally distributed.
- If you have a large sample, the error terms can even deviate substantially from normality without greatly affecting appropriateness of the confidence interval.

Prediction interval for a new response $Y_{h(new)}$

Again, what are we predicting?



(1-α)100% prediction interval for new response $Y_{h(new)}$

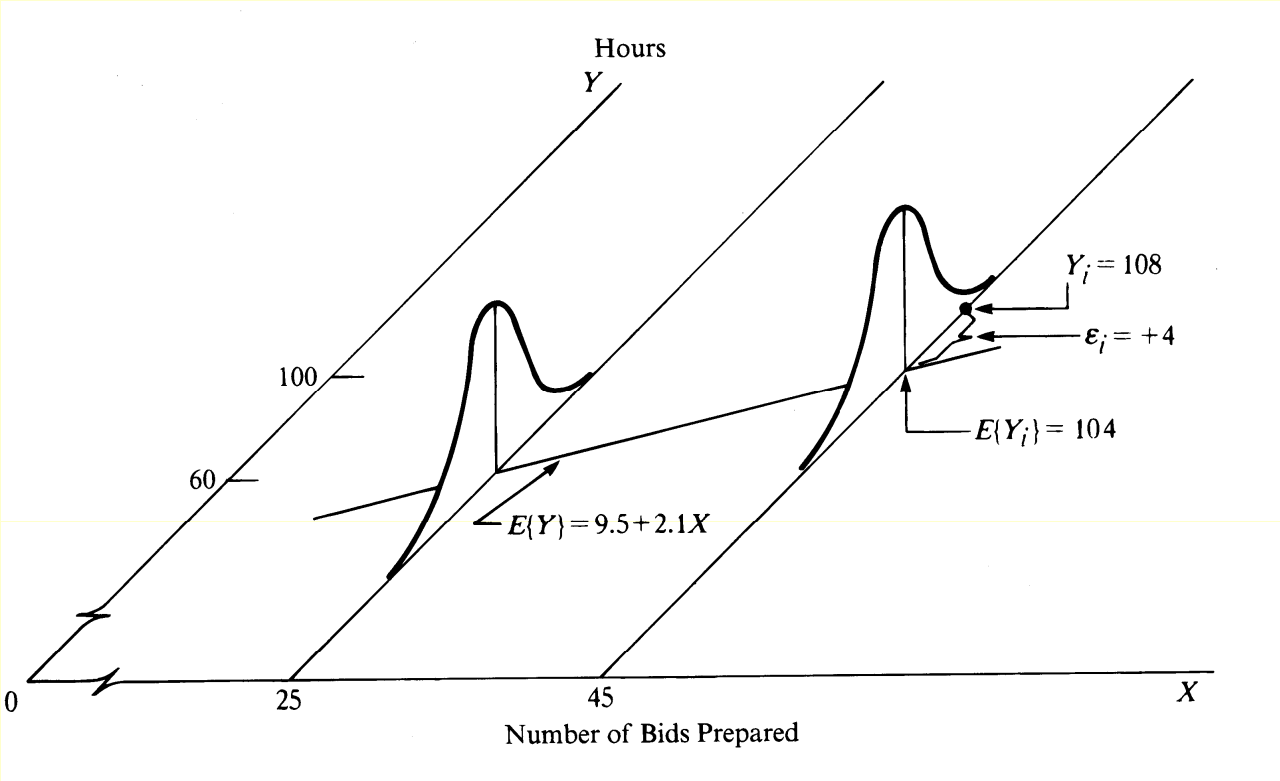
Formula in words:

Sample prediction ± (t-multiplier × standard error)

Formula in notation:

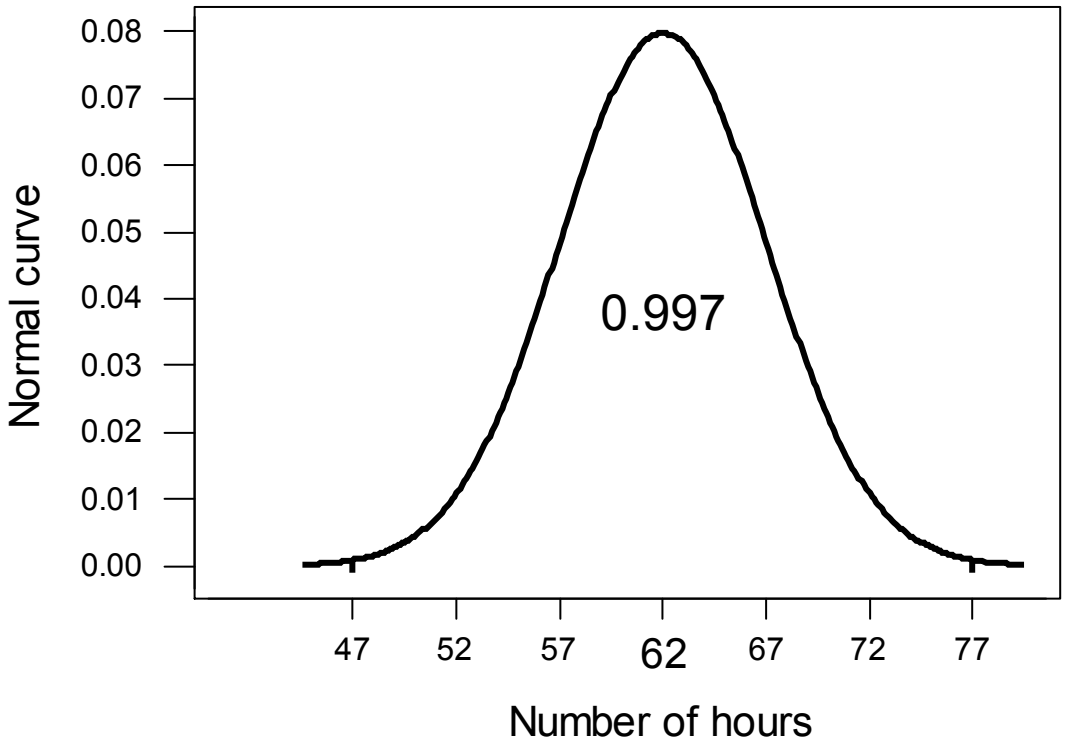
$$\hat{y}_h \pm t_{\left(1-\alpha/2, n-2\right)} \times \sqrt{MSE\left(1+\frac{1}{n}+\frac{\left(x_h-\bar{x}\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right)}$$

Prediction of $Y_{h(new)}$ if mean $E(Y)$ is known



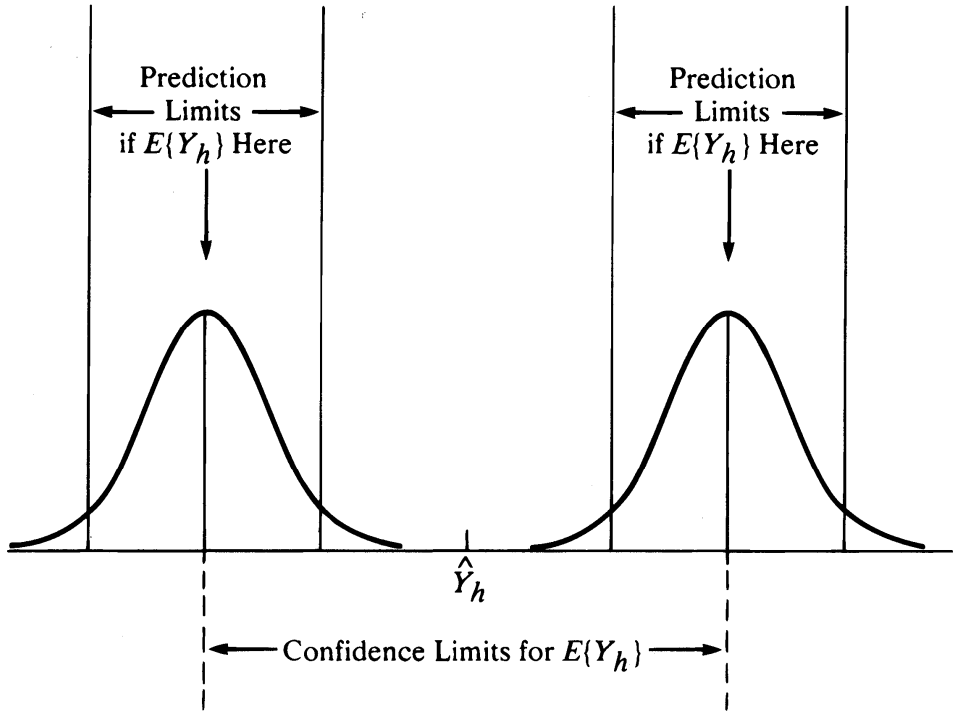
Assume $\sigma^2 = 25$ so $\sigma = 5$

Prediction of $Y_{h(new)}$ if mean $E(Y)$ is known



Prediction of $Y_{h(new)}$ if mean $E(Y)$ is not known

FIGURE 2.5 Prediction of $Y_{h(new)}$ when Parameters Unknown.



Summary of prediction issues

- We cannot be certain of the mean of the distribution of Y.
- Prediction limits for $Y_{h(new)}$ must take into account:
  - variation in the possible mean of the distribution of Y
  - variation in the responses Y within the probability distribution

Variation of the prediction

The variation in the prediction of a new response depends on two components:

1. the variation due to estimating the mean $E(Y_h)$ with $\hat{y}_h$
2. the variation in Y within the probability distribution

$$\sigma^2 + \sigma^2(\hat{Y}_h)$$

which is estimated by:

$$MSE + MSE\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = MSE\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

(1-α)100% prediction interval for new response $Y_{h(new)}$

Formula in words:

Sample prediction ± (t-multiplier × standard error)

Formula in notation:

$$\hat{y}_h \pm t_{\left(1-\alpha/2, n-2\right)} \times \sqrt{MSE\left(1+\frac{1}{n}+\frac{\left(x_h-\bar{x}\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right)}$$

Confidence intervals and prediction intervals for response

- **Stat** >> **Regression** >> **Regression …**
- Specify response and predictor(s).
- Select Options…
  - In "**Prediction intervals for new observations**" box, specify either the X value or a column name containing multiple X values.
  - Specify confidence level (default is 95%).
- Click on OK. Click on OK.
- Results appear in session window.

S = 19.12 R-Sq = 68.0% R-Sq(adj)= 67.3%

Predicted Values for New Observations

New	**Fit**	**SE Fit**	95.0% CI	**95.0% PI**	
1	**150.08**	**2.75**	(144.6,155.6)	**(111.2,188.93)**	
2	**221.82**	**7.42**	(206.9,236.8)	**(180.6,263.07)**	X

X denotes a row with X values away from the center

Values of Predictors for New Observations

New Obs	Latitude
1	40.0
2	28.0

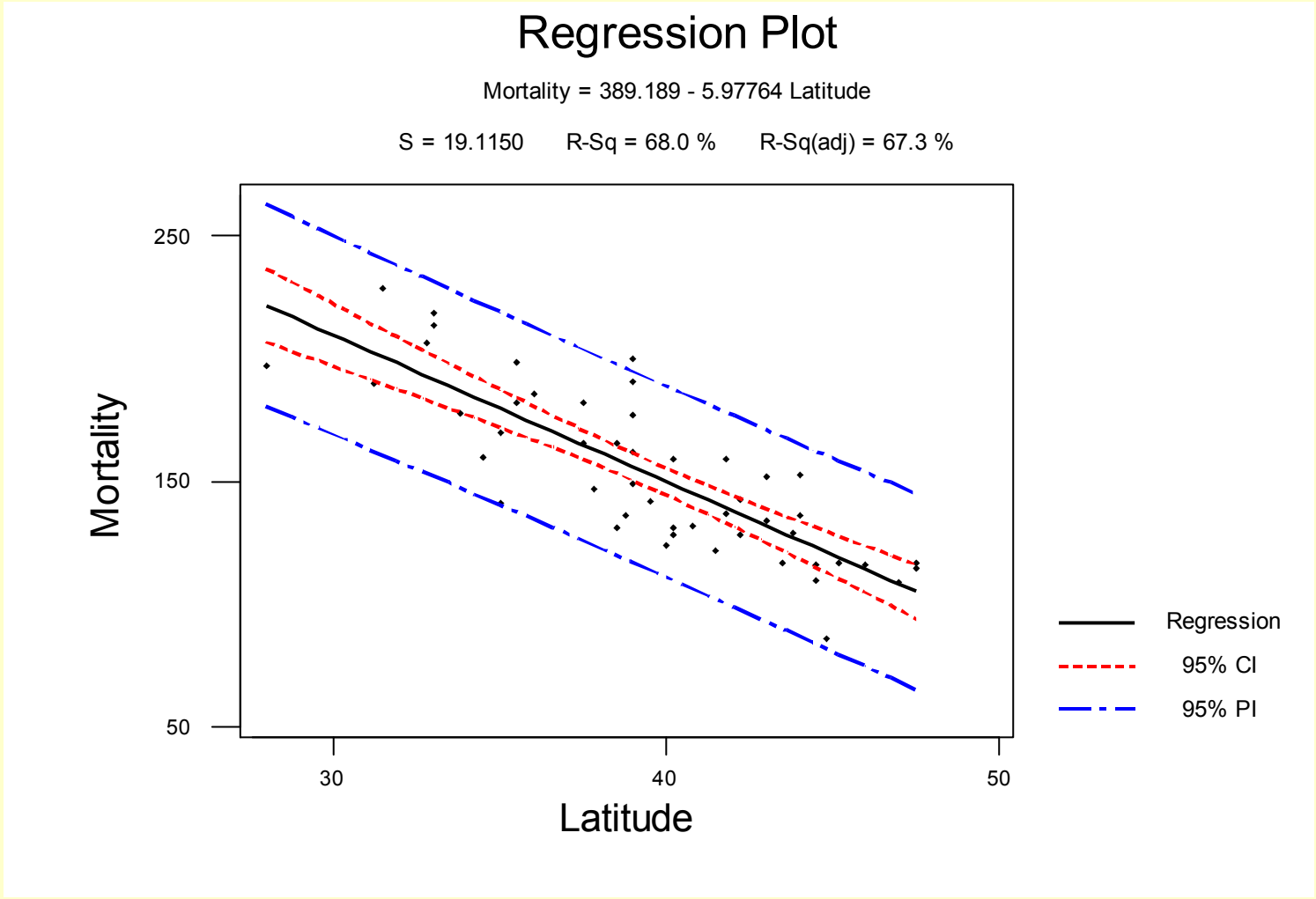
Mean of Lat = 39.533

Comments on assumptions

- x_h is a value within scope of model, but it is not necessary that it is one of the x values in the data set.
- The formula for the prediction interval depends strongly on the assumption that the error terms are normally distributed.

A plot of the confidence interval and prediction interval

- **Stat** >> **Regression** >> **Fitted line plot** …
- Specify predictor and response.
- Under Options …
  - Select **Display confidence bands**.
  - Select **Display prediction bands**.
  - Specify desired confidence level (95% default)
- Select OK. Select OK.


Regression Plot
Mortality = 389.189 - 5.97764 Latitude
S = 19.1150 R-Sq = 68.0 % R-Sq(adj) = 67.3 %
Mortality
250
150
50
30
40
50
Latitude
Regression
95% CI
95% PI

Analysis of variance approach to regression analysis

… an (alternative) approach to testing for a linear association

Example: Mortality and Latitude

The regression equation is Mort = 389 - 5.98 Lat

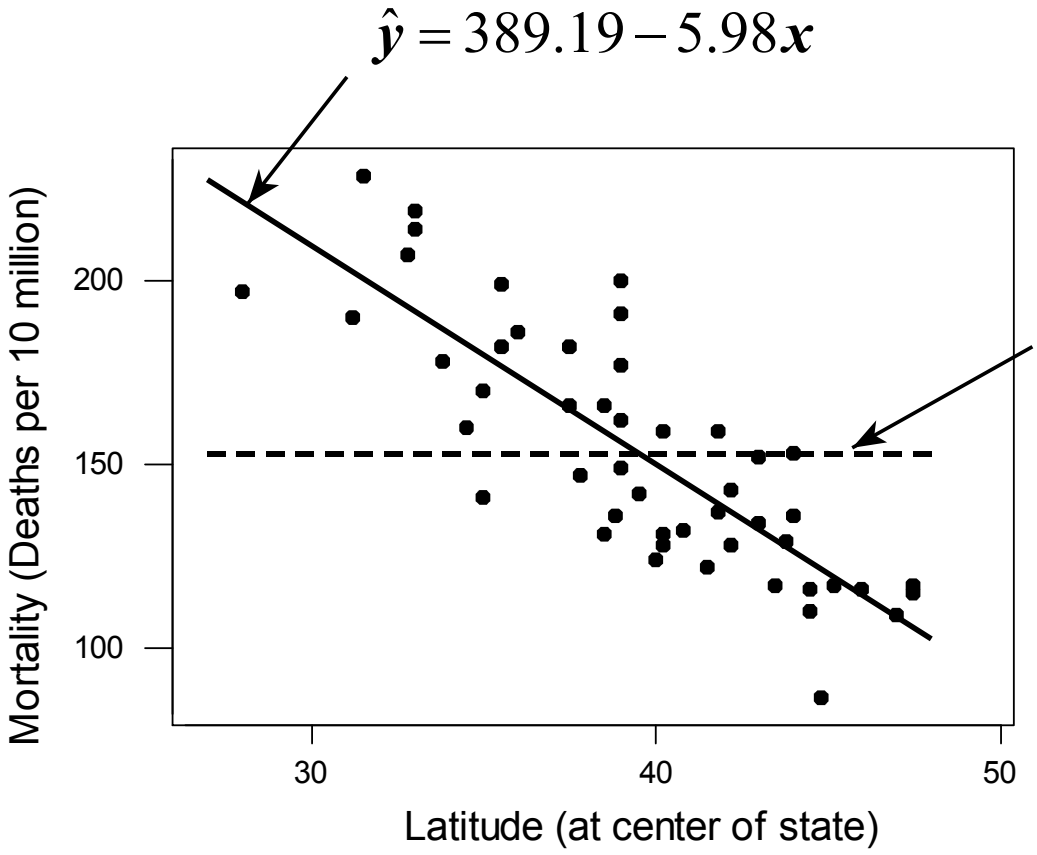
Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

S = 19.12 R-Sq = 68.0% R-Sq(adj) = 67.3%

Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	**1**	**36464**	**36464**	**99.80**	**0.000**
Residual Error	**47**	**17173**	**365**		
Total	**48**	**53637**			

Example: Mortality and Latitude



$$\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 = 36464$$

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = 17173$$

$$\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = 53637$$

Example: Height and GPA

The regression equation is gpa = 3.41 - 0.0066 height

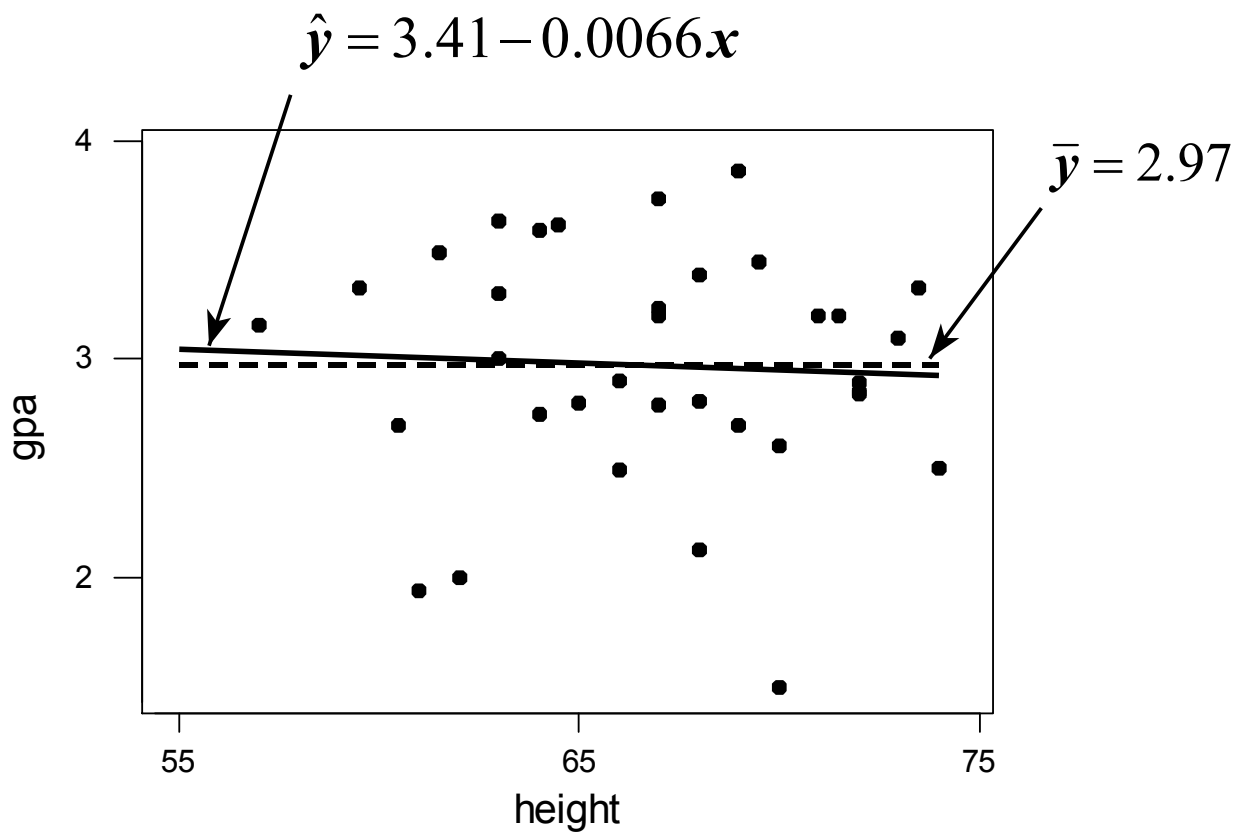
Predictor	Coef	SE Coef	T	P
Constant	3.410	1.435	2.38	0.023
height	-0.00656	0.02143	-0.31	0.761

S = 0.5423 R-Sq = 0.3% R-Sq(adj) = 0.0%

Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	**1**	**0.0276**	**0.0276**	**0.09**	**0.761**
Residual Error	**33**	**9.7055**	**0.2941**		
Total	**34**	**9.7331**			

Example: Height and GPA



$$\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 = 0.0276$$

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = 9.7055$$

$$\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = 9.7331$$

The basic idea

- Break down the variation in Y ("**total sum of squares**") into two components:
  - a component that is "due to" the change in X ("**regression sum of squares**")
  - a component that is just due to random error ("**error sum of squares**")
- If the regression sum of squares is a large component of the total sum of squares, it suggests that there is a linear association.

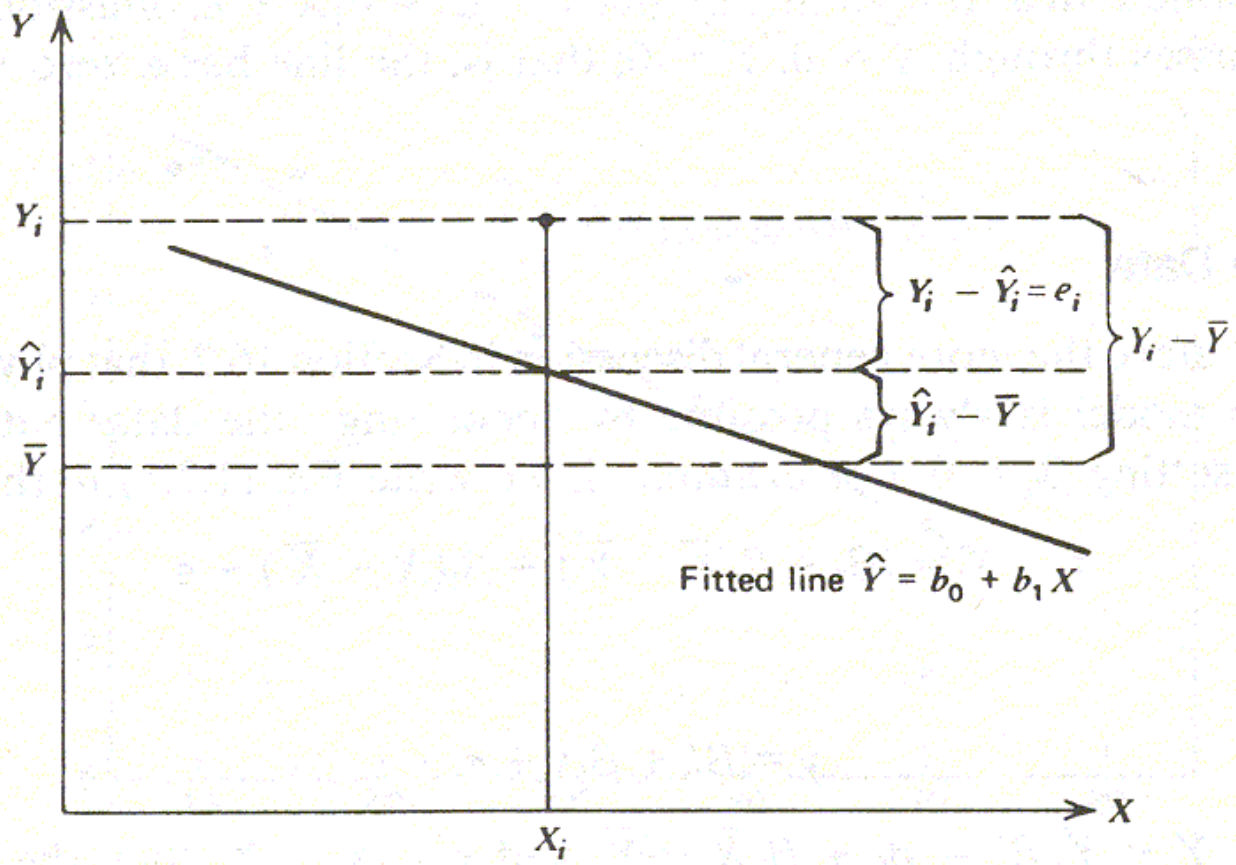


Figure 1.7. Geometrical meaning of the identity (1.3.1).

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

The above decomposition holds for the sum of the squared deviations, too:

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

Total sum of squares (SSTO)

Regression sum of squares (SSR)

Error sum of squares (SSE)

$$SSTO = SSR + SSE$$

Breakdown of degrees of freedom

Degrees of freedom associated with SSTO

$$(n-1) = (1) + (n-2)$$

Degrees of freedom associated with SSR

Degrees of freedom associated with SSE

Example: Mortality and Latitude

The regression equation is Mort = 389 - 5.98 Lat

Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

S = 19.12 R-Sq = 68.0% R-Sq(adj) = 67.3%

Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	**1**	**36464**	**36464**	**99.80**	**0.000**
Residual Error	**47**	**17173**	**365**		
Total	**48**	**53637**			

Definitions of Mean Squares

We already know the **mean square error** (**MSE**) is defined as:

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

Similarly, the **regression mean square** (**MSR**) is defined as:

$$MSR = \frac{\sum (\hat{y}_i - \bar{y})^2}{1} = \frac{SSR}{1}$$

Analysis of Variance (ANOVA) Table

TABLE 2.2 ANOVA Table for Simple Linear Regression.

Source of Variation	SS	df	MS	$E\{MS\}$
Regression	$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \Sigma(X_i - \bar{X})^2$
Error	$SSE = \Sigma(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \Sigma(Y_i - \bar{Y})^2$	$n-1$		

Expected Mean Squares

$$E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

$$E(MSE) = \sigma^2$$

- If $\beta_1 = 0$, we'd expect the ratio MSR/MSE to be …
- If $\beta_1 \neq 0$, we'd expect the ratio MSR/MSE to be …
- Use ratio, MSR/MSE, to reject whether or not $\beta_1 = 0$.

The formal F-test for slope parameter β_1

Null hypothesis $H_0: \beta_1 = 0$
Alternative hypothesis $H_A: \beta_1 \neq 0$

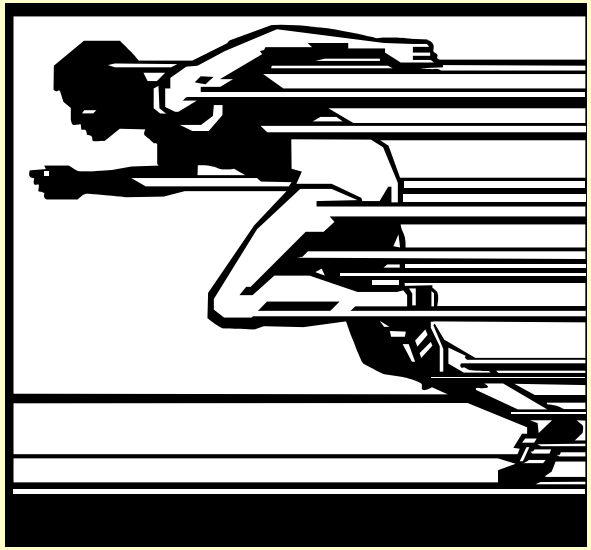
Test statistic $F^* = \frac{MSR}{MSE}$

P-value = What is the probability that we'd get an F* statistic as large as we did, if the null hypothesis is true? (One-tailed test!)

The P-value is determined by comparing F* to an **F distribution** with *1* **numerator degree of freedom** and *n-2* **denominator degrees of freedom**.

Winning times (in seconds) in Men's 200 meter Olympic sprints, 1900-1996.

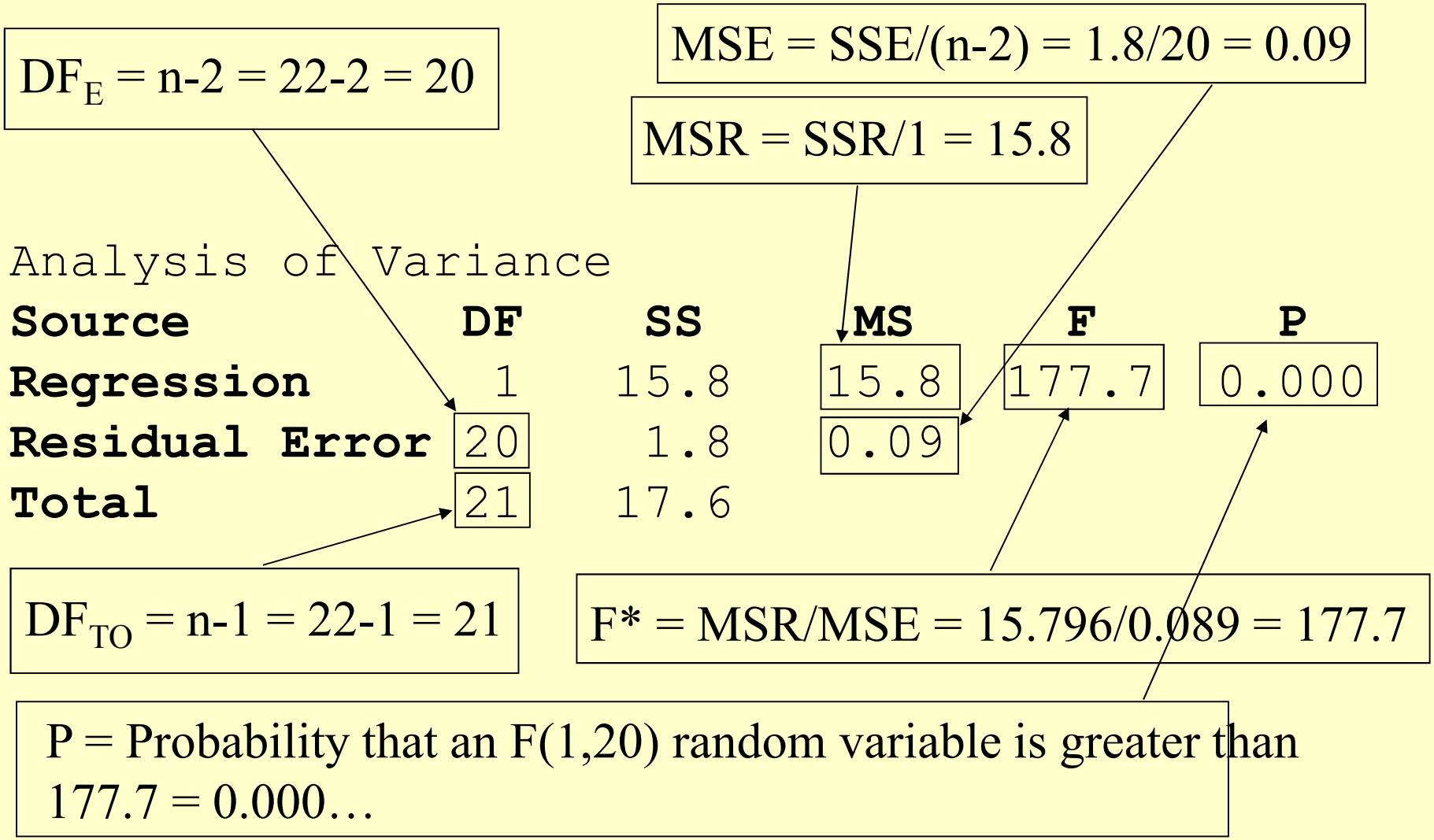
Are men getting faster?



Row	Year	Men200m
1	1900	22.20
2	1904	21.60
3	1908	22.60
4	1912	21.70
5	1920	22.00
6	1924	21.60
7	1928	21.80
8	1932	21.20
9	1936	20.70
10	1948	21.10
11	1952	20.70
12	1956	20.60
13	1960	20.50
14	1964	20.30
15	1968	19.83
16	1972	20.00
17	1976	20.23
18	1980	20.19
19	1984	19.80
20	1988	19.75
21	1992	20.01
22	1996	19.32


Regression Plot
Men200m = 76.1534 - 0.0283833 Year
S = 0.298134 R-Sq = 89.9 % R-Sq(adj) = 89.4 %
Men200m
22.5
21.5
20.5
19.5
1900
1950
2000
Year

Analysis of Variance Table



For simple linear regression model, the F-test and t-test are equivalent.

Predictor	Coef	SE Coef	T	P
Constant	76.153	4.152	18.34	0.000
Year	-0.0284	0.00213	**-13.33**	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15.796	15.796	**177.7**	0.000
Residual Error	20	1.778	0.089		
Total	21	17.574			

$$(-13.33)^2 = 177.7$$

$$\left(t^*_{(n-2)}\right)^2 = F^*_{(1,n-2)}$$

Equivalence of F-test to t-test

- For a given α level, the F-test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is algebraically equivalent to the two-tailed t-test.
- Will get exactly same P-values, so…
  - If one test rejects H_0, then so will the other.
  - If one test does not reject H_0, then so will the other.

Should I use the F-test or the t-test?

- The F-test is only appropriate for testing that the slope differs from 0 ($\beta_1 \neq 0$).
- Use the t-test to test that the slope is positive ($\beta_1 > 0$) or negative ($\beta_1 < 0$).
- F-test is more useful for multiple regression model when we want to test that more than one slope parameter is 0.

Getting ANOVA table

- The Analysis of Variance (ANOVA) Table is default output for either command
  - **Stat** >> **Regression** >> **Regression …**
  - **Stat** >> **Regression** >> **Fitted line plot …**

The **general linear test** approach to regression analysis

Three basic steps

- Define a (larger) **full model**.
- Define a (smaller) **reduced model**.
- Use an **F statistic** to decide whether or not to reject the smaller reduced model in favor of the larger full model.

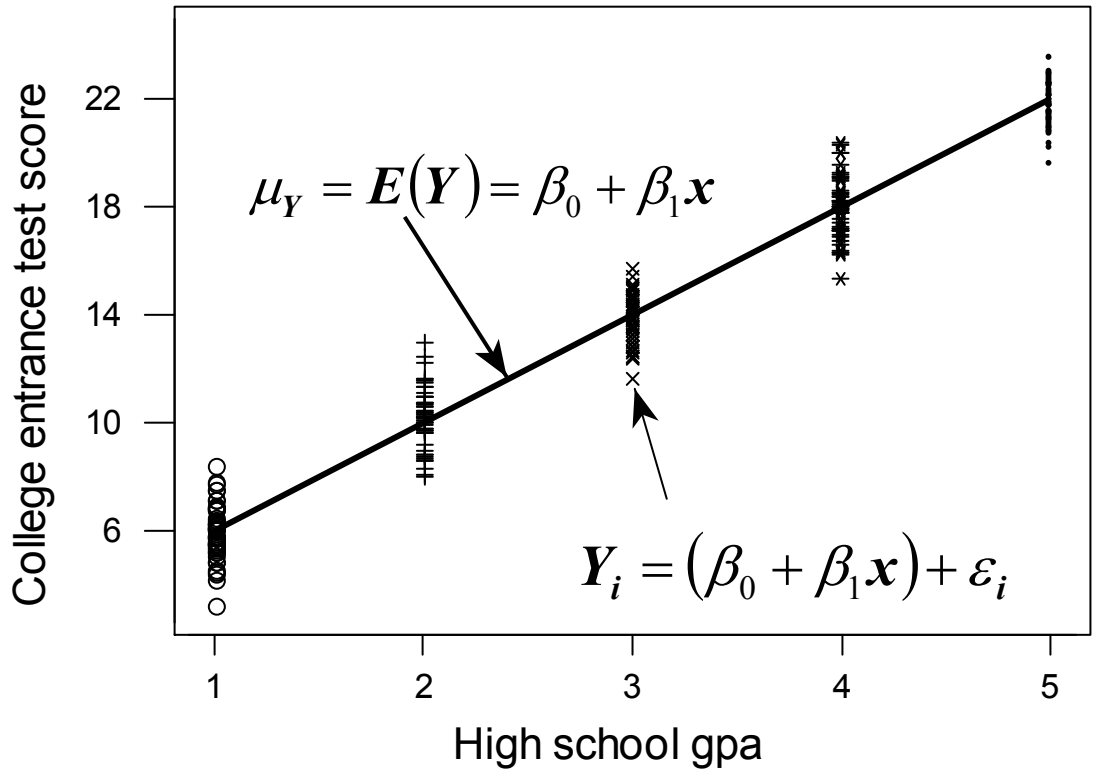
The full model

The **full model** (or **unrestricted model**) is the model thought to be most appropriate for the data.

For simple linear regression, the full model is:

$$\boldsymbol{Y}_i = \beta_0 + \beta_1 \boldsymbol{X}_i + \varepsilon_i$$

The full model



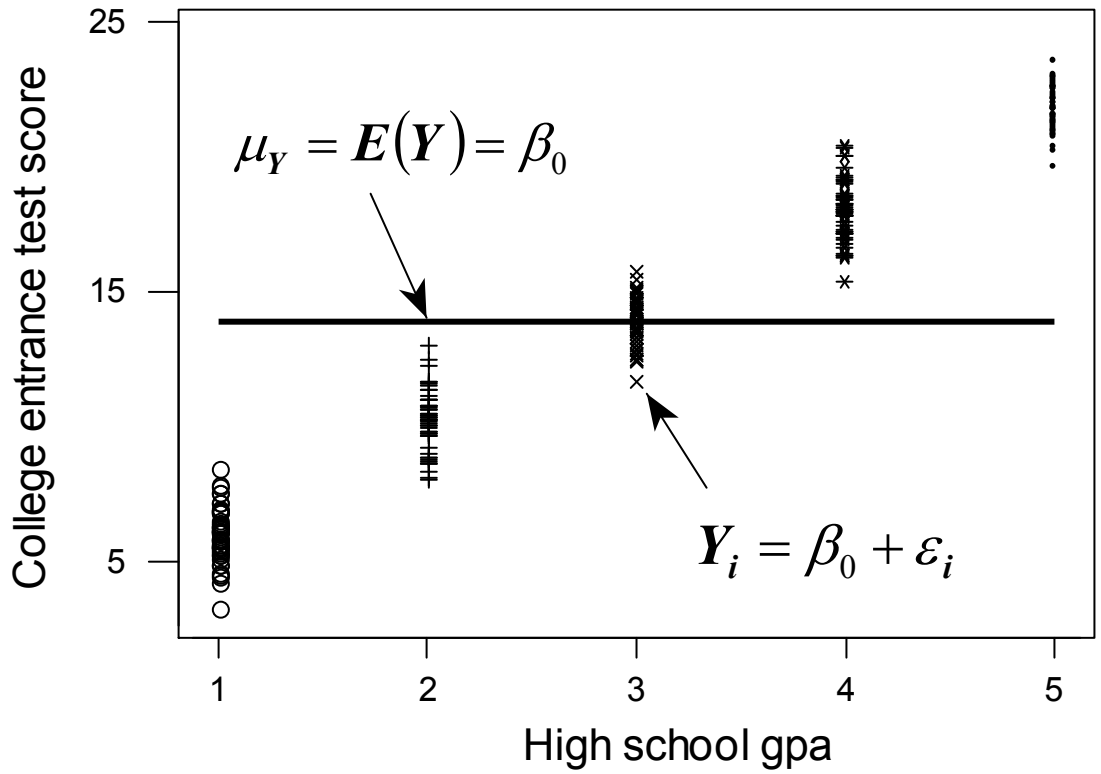
The reduced model

The **reduced model** (or **restricted model**) is the model described by the null hypothesis H_0.

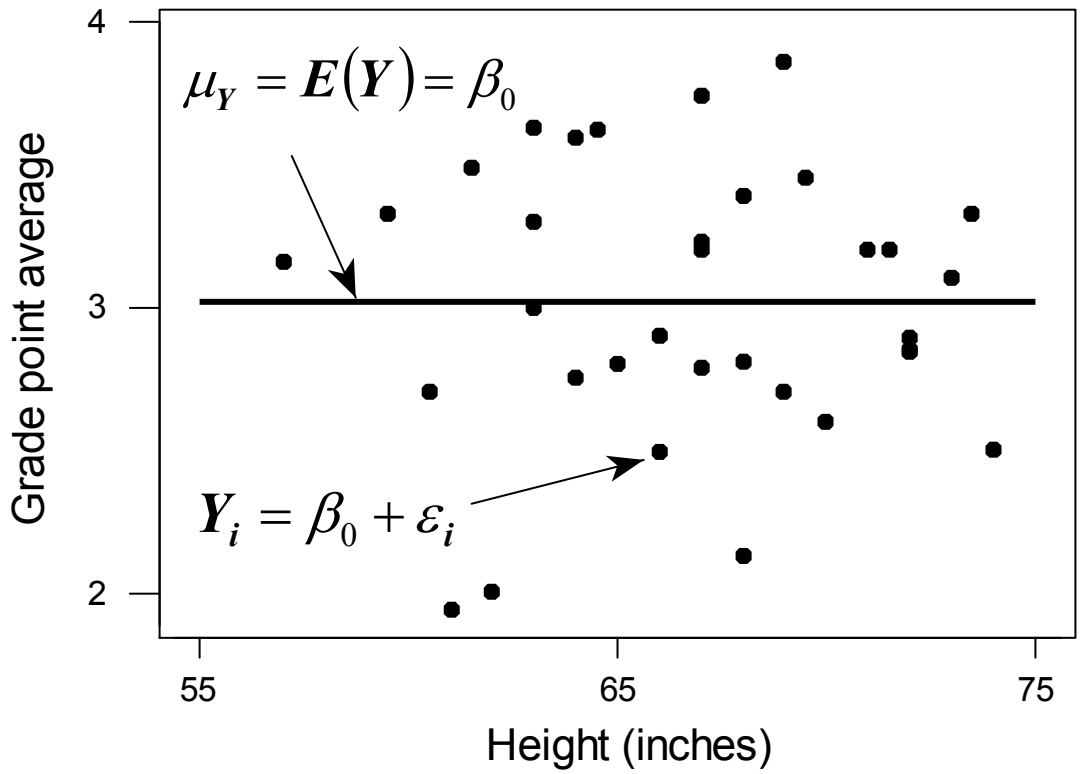
For simple linear regression, the null hypothesis is H_0: $\beta_1 = 0$. Therefore, the reduced model is:

$$\boldsymbol{Y_i} = \beta_0 + \boldsymbol{\varepsilon_i}$$

The reduced model



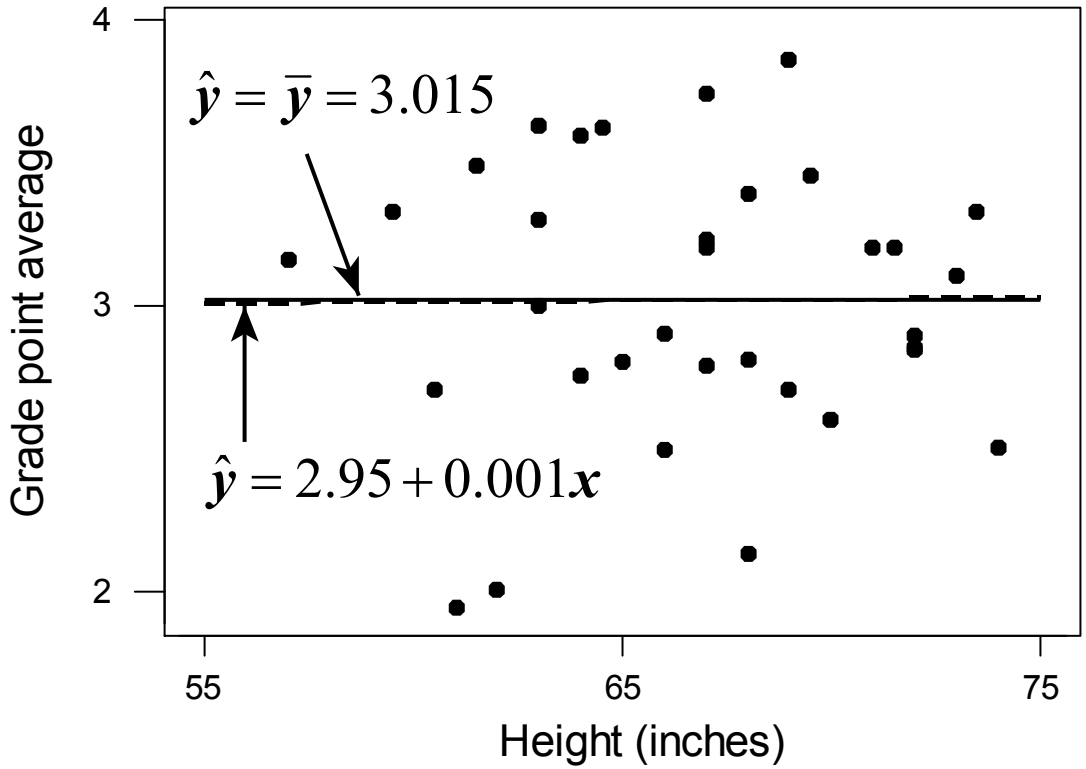
The reduced model



The general linear test approach

- "Fit the full model" to the data.
  - Obtain least squares estimates of β_0 and β_1.
  - Determine error sum of squares – "**SSE(F)**."
- "Fit the reduced model" to the data.
  - Obtain least squares estimate of β_0.
  - Determine error sum of squares – "**SSE(R)**."

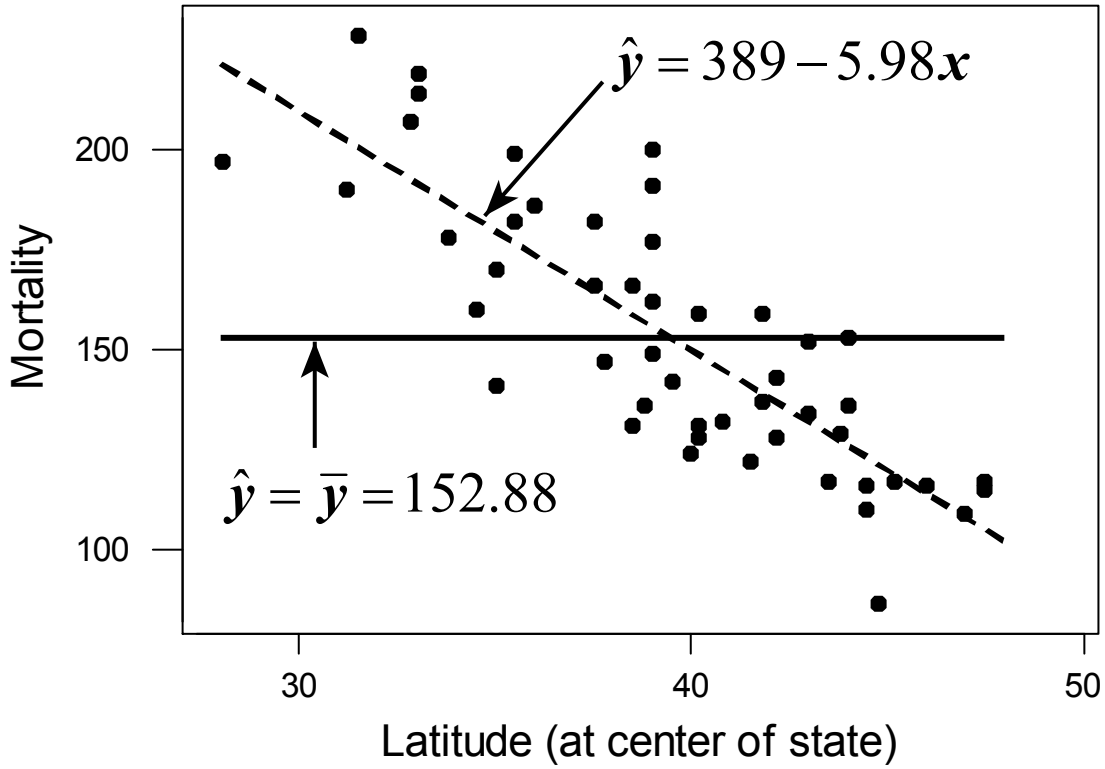
The general linear test approach (cont'd)



$$SSE(F) = 7.5028$$

$$SSE(R) = 7.5035$$

The general linear test approach (cont'd)



$$SSE(F) = 17173$$

$$SSE(R) = 53637$$

The general linear test approach (cont'd)

- Compare SSE(R) and SSE(F).
- SSE(R) is always larger than (or same as) SSE(F).
  - If SSE(F) is close to SSE(R), then variation around fitted full model regression function is almost as large as variation around fitted reduced model regression function.
  - If SSE(F) and SSE(R) differ greatly, then the additional parameter(s) in the full model substantially reduce the variation around the fitted regression function.

How close is close?

The test statistic is a function of SSE(R)-SSE(F):

$$F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F} \right) \div \left(\frac{SSE(F)}{df_F} \right)$$

The degrees of freedom (df_R and df_F) are those associated with the reduced and full model error sum of squares, respectively.

Reject H_0 if F* is large (or if P-value is small).

But for simple linear regression, it's just the same F test as before …

$$F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F} \right) \div \left(\frac{SSE(F)}{df_F} \right)$$

$$df_R = n - 1 \qquad SSE(R) = SSTO$$

$$df_F = n - 2 \qquad SSE(F) = SSE$$

$$F^* = \left(\frac{SSTO - SSE}{(n-1) - (n-2)} \right) \div \left(\frac{SSE}{(n-2)} \right) = \frac{MSR}{MSE}$$

The formal F-test for slope parameter β_1

Null hypothesis $H_0: \beta_1 = 0$
Alternative hypothesis $H_A: \beta_1 \neq 0$

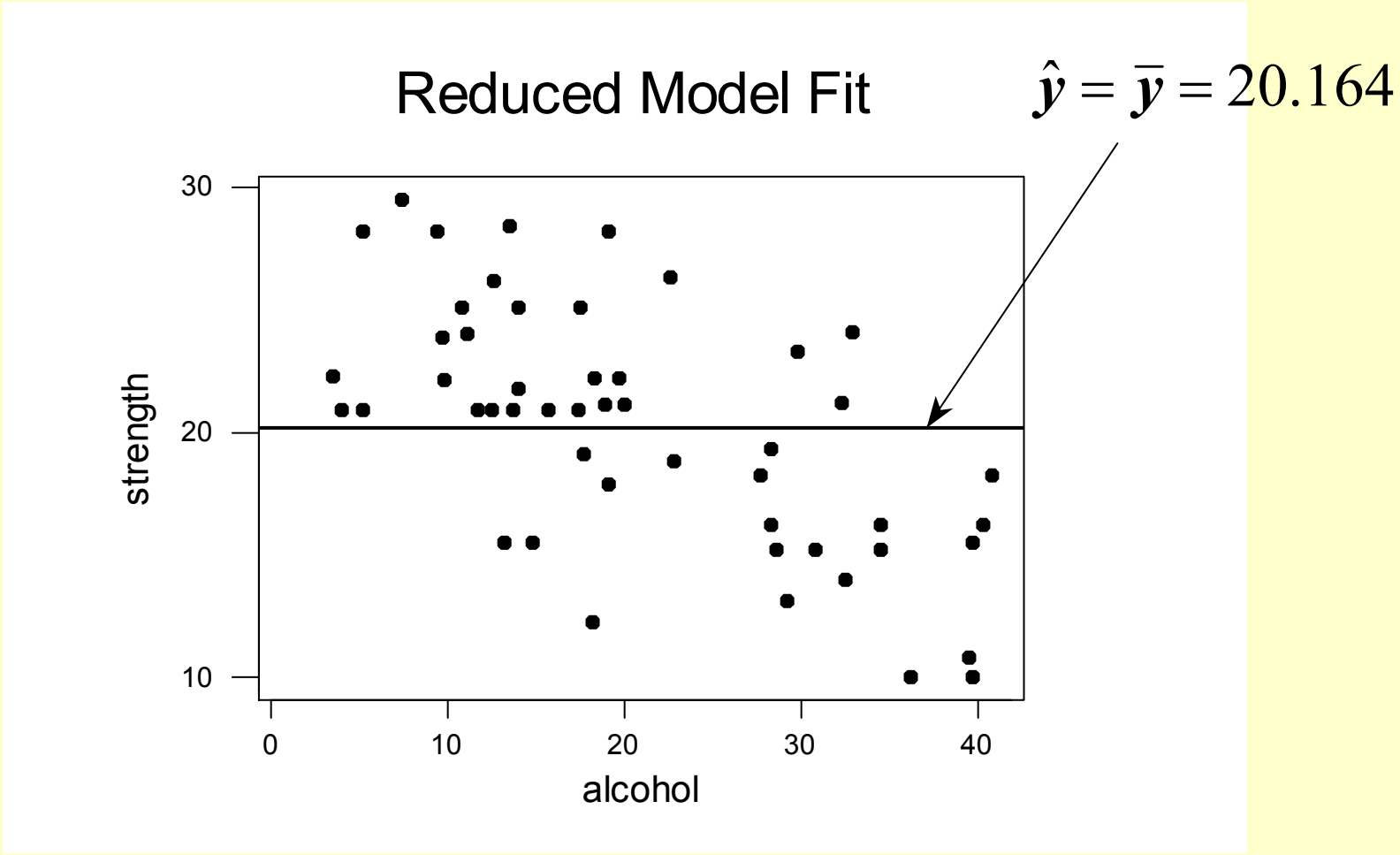
Test statistic $F^* = \frac{MSR}{MSE}$

P-value = What is the probability that we'd get an F* statistic as large as we did, if the null hypothesis is true? (One-tailed test!)

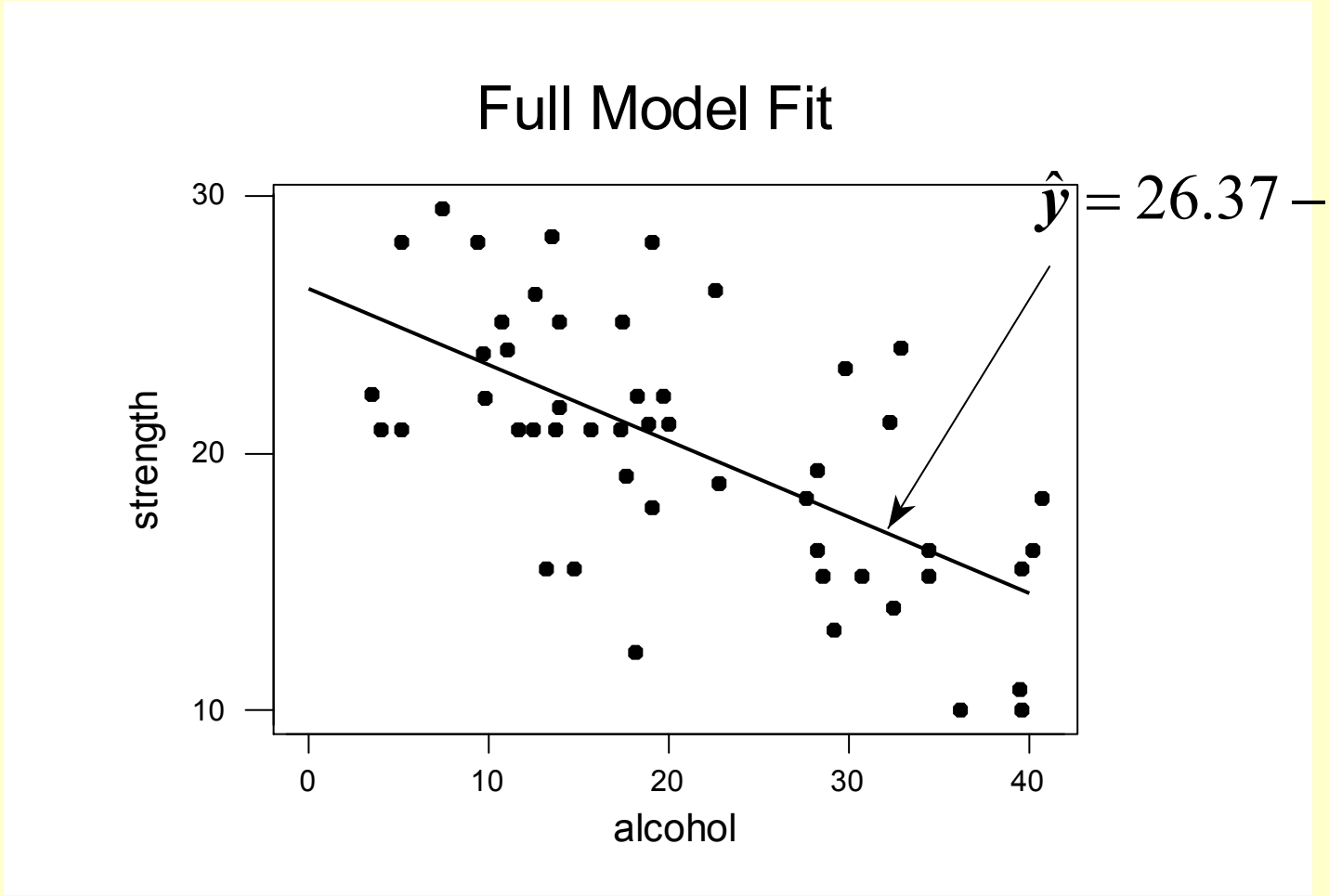
The P-value is determined by comparing F* to an **F distribution** with *1* **numerator degree of freedom** and *n-2* **denominator degrees of freedom**.

Example: Alcoholism and Muscle strength?

- Report on strength tests for a sample of 50 alcoholic men
  - X = total lifetime dose of alcohol (kg per kg of body weight)
  - Y = strength of deltoid muscle in man's non-dominant arm



$$SSE(R) = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 = 1224.32$$



$$SSE(F) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = 720.27$$

The ANOVA table

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	504.04	504.040	33.5899	0.000
Error	48	720.27	15.006		
Total	49	1224.32			

SSE(R)=SSTO

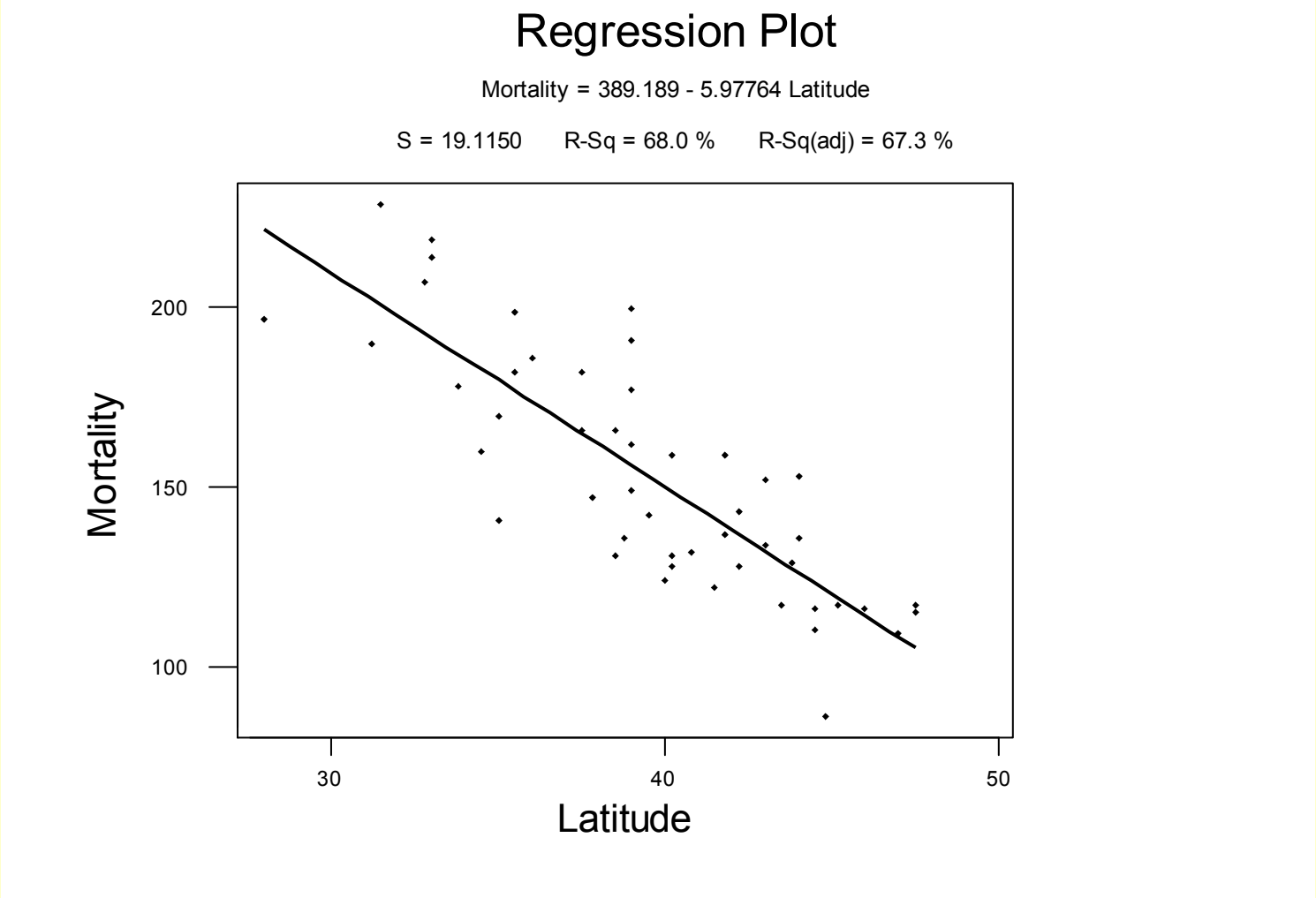
SSE(F)=SSE

There is a statistically significant linear association between alcoholism and arm strength.

Lack of Fit (LOF) Test

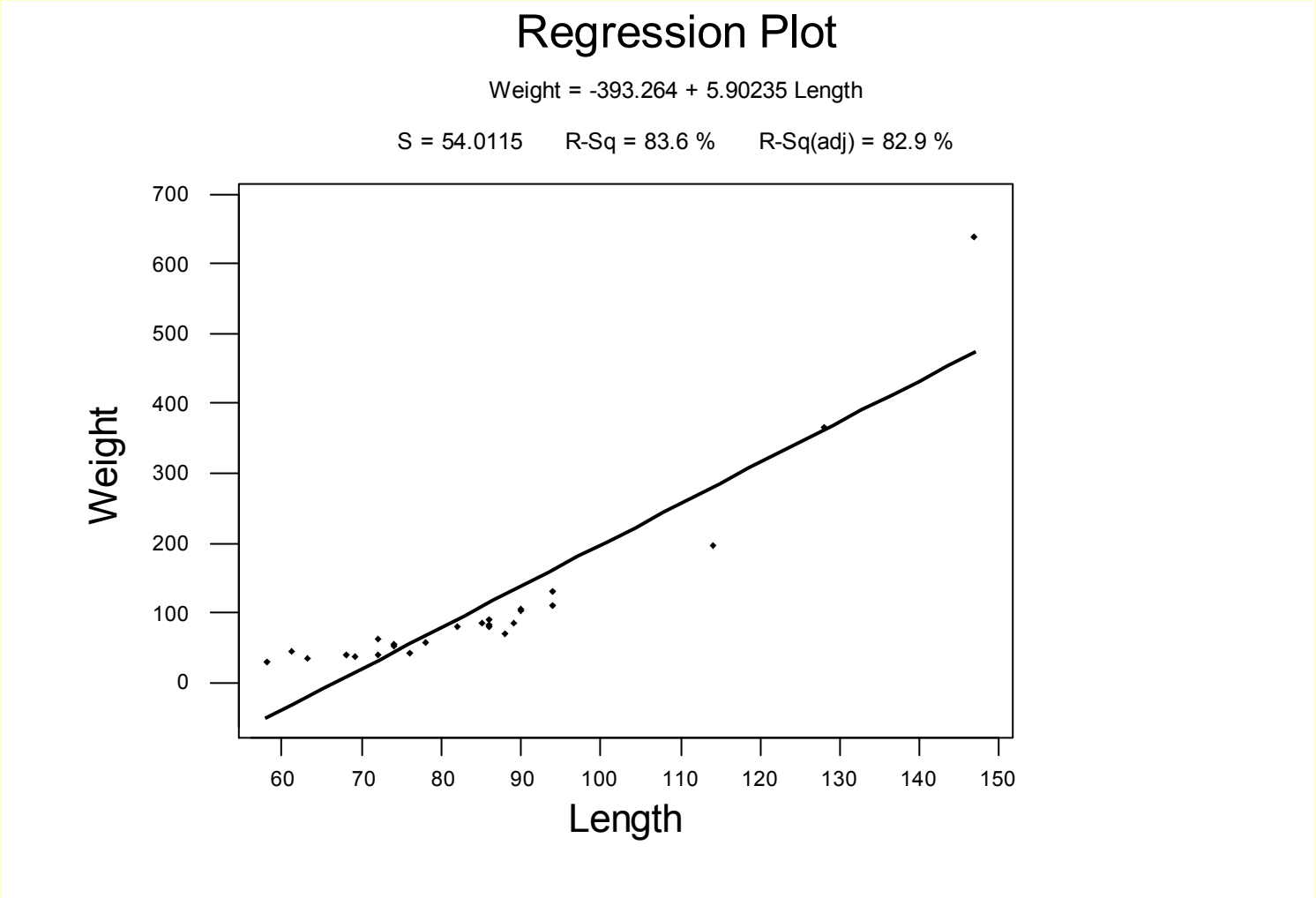
An F test for checking whether a specific type of regression function adequately fits the data

Example 1



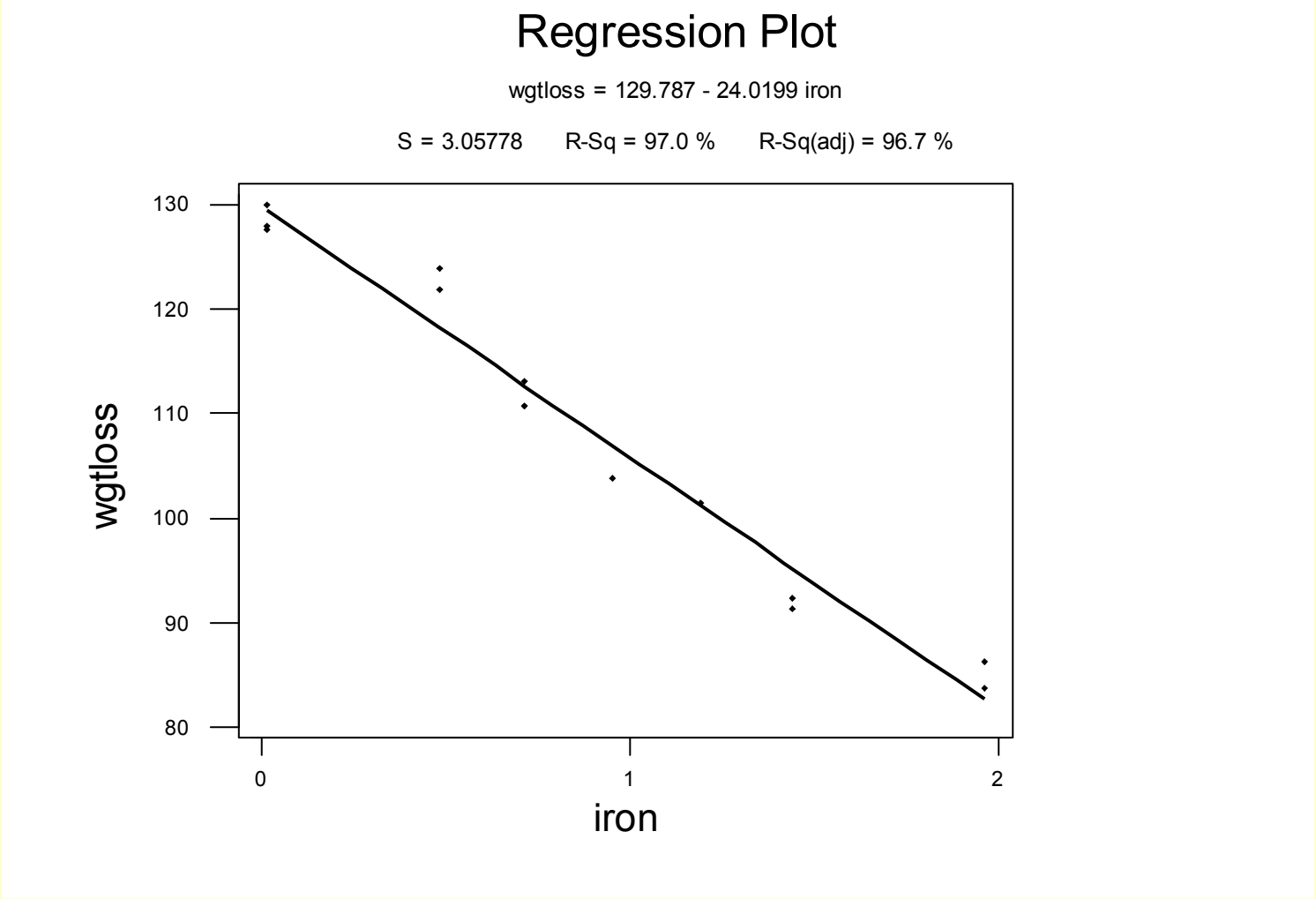
Do the data suggest that a linear function is not adequate in describing the relationship between skin cancer mortality and latitude?

Example 2



Do the data suggest that a linear function is not adequate in describing the relationship between the length and weight of an alligator?

Example 3



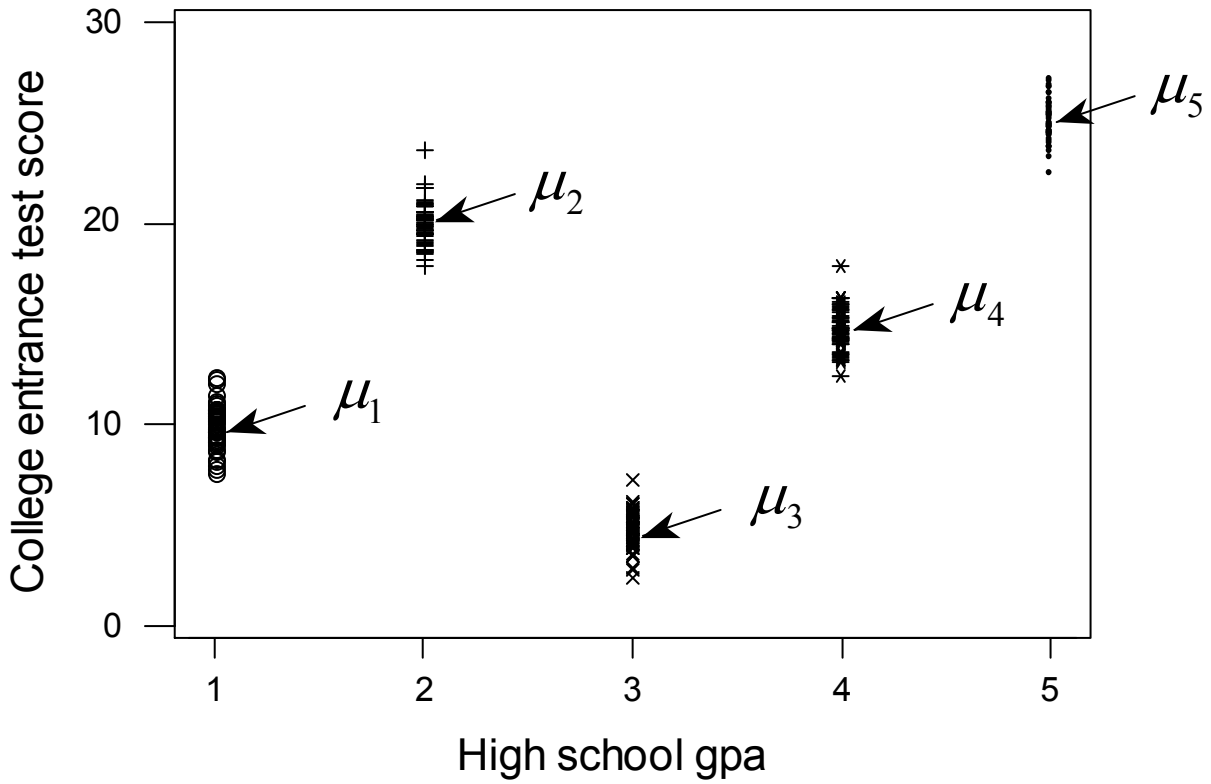
Do the data suggest that a linear function is not adequate in describing the relationship between iron content and weight loss due to corrosion?

Lack of fit test for a linear function

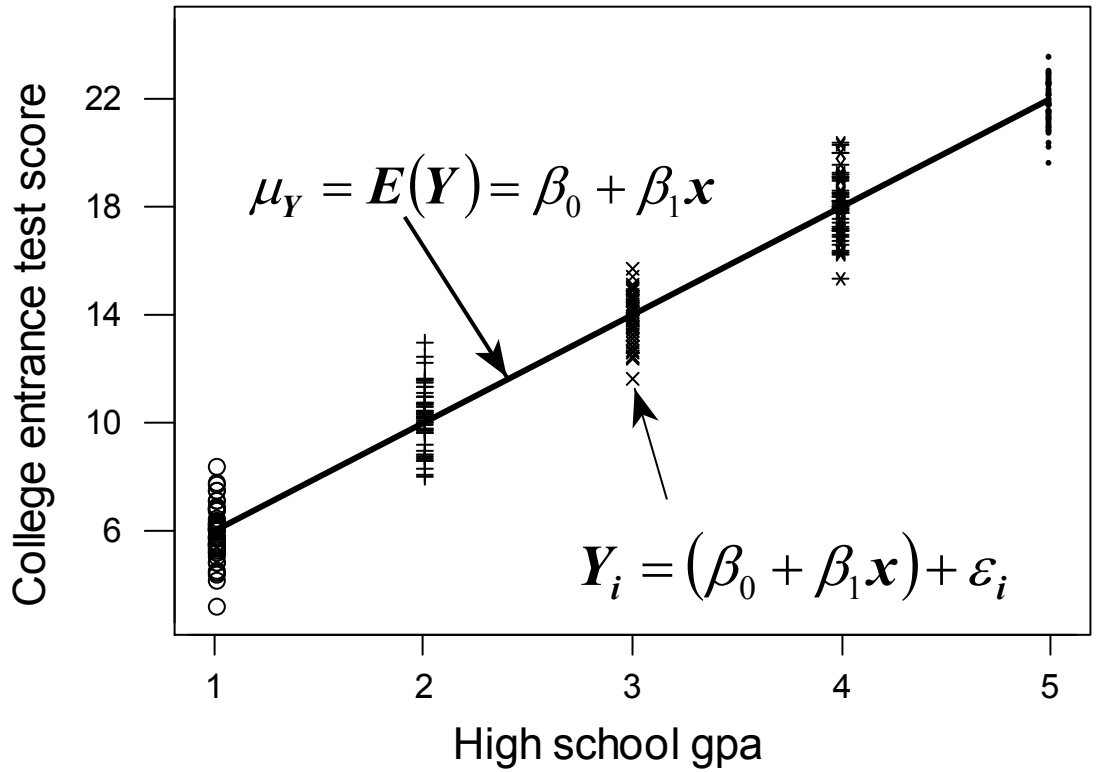
- Use general linear test approach.
- Full model is most general model with no restrictions on the means μ_j at each X_j level.
- Reduced model assumes that the μ_j are a linear function of the X_j, *i.e.*, $\mu_j = \beta_0 + \beta_1 X_j$.

The full model

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$



The reduced model



Assumptions and requirements

- The errors, ε_i, and hence the responses Y_i, are **independent**.
- The errors, ε_i, and hence the responses Y_i, are **normally distributed**.
- The errors, ε_i, and hence the responses Y_i, have **equal variances** (σ^2) for all x values.
- The LOF test requires repeat observations, called **replicates**, for at least one of the X values.

Notation

iron	wgtloss
0.01	127.6
0.01	130.1
0.01	128.0
0.48	124.0
0.48	122.0
0.71	110.8
0.71	113.1
0.95	103.9
1.19	101.5
1.44	92.3
1.44	91.4
1.96	83.7
1.96	86.2

- **c different levels** of X (c=7 with X_1=0.01, X_2=0.48, …, X_7=1.96)
- **n_j = number of replicates** for j^{th} level of X (X_j) (n_1=3, n_2=2, …, n_7=2) for a total of $\mathbf{n = n_1 + \ldots + n_c}$ observations.
- **Y_{ij} = observed value of the response variable** for the i^{th} replicate of X_j (Y_{11}=127.6, Y_{21}=130.1, …, Y_{27}=86.2)

The Full Model

Assume nothing about (or "put no structure on") the means of the responses, μ_j, at the j^{th} level of X:

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

Least squares estimates of μ_j are sample means, $\hat{\mu}_j = \bar{Y}_j$,

of responses at X_j level.

"**Pure error sum of squares**"

$$SSE(F) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_j\right)^2 = SSPE$$

The Reduced Model

Assume the means of the responses, μ_j, are a linear function of the j^{th} level of X:

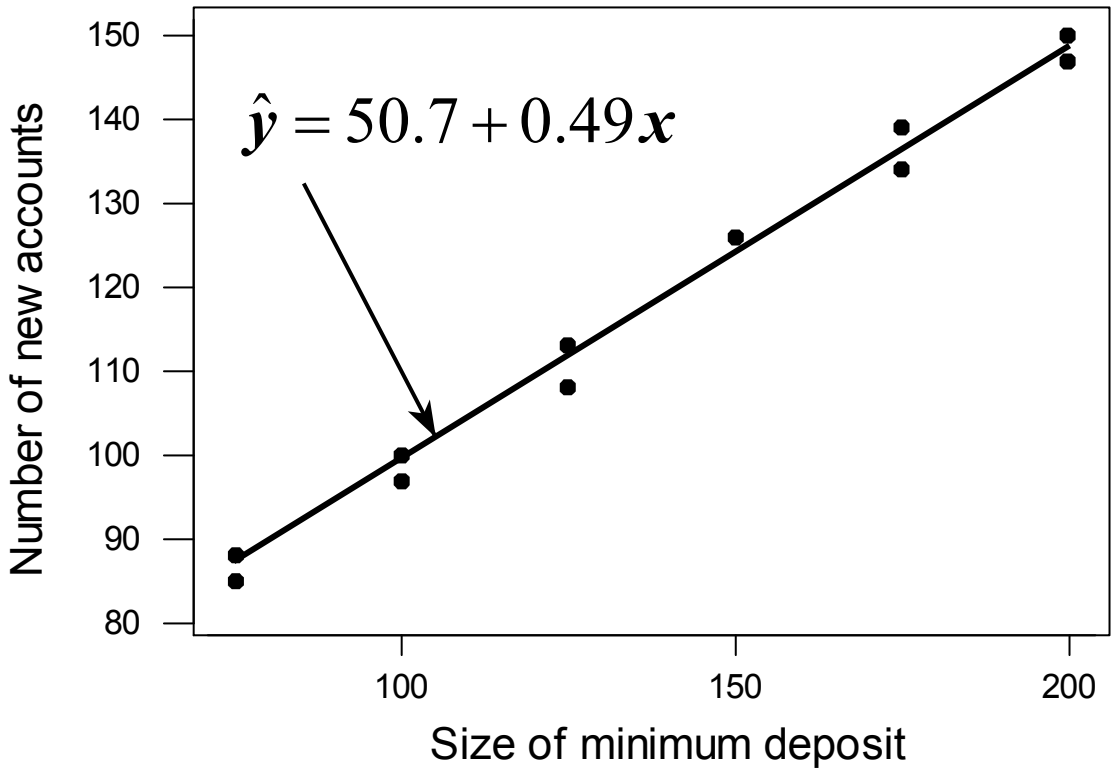
$$\boldsymbol{Y_{ij} = \beta_0 + \beta_1 X_j + \varepsilon_{ij}}$$

Least squares estimates of μ_j are as usual: $\boldsymbol{\hat{Y}_{ij} = b_0 + b_1 X_j}$

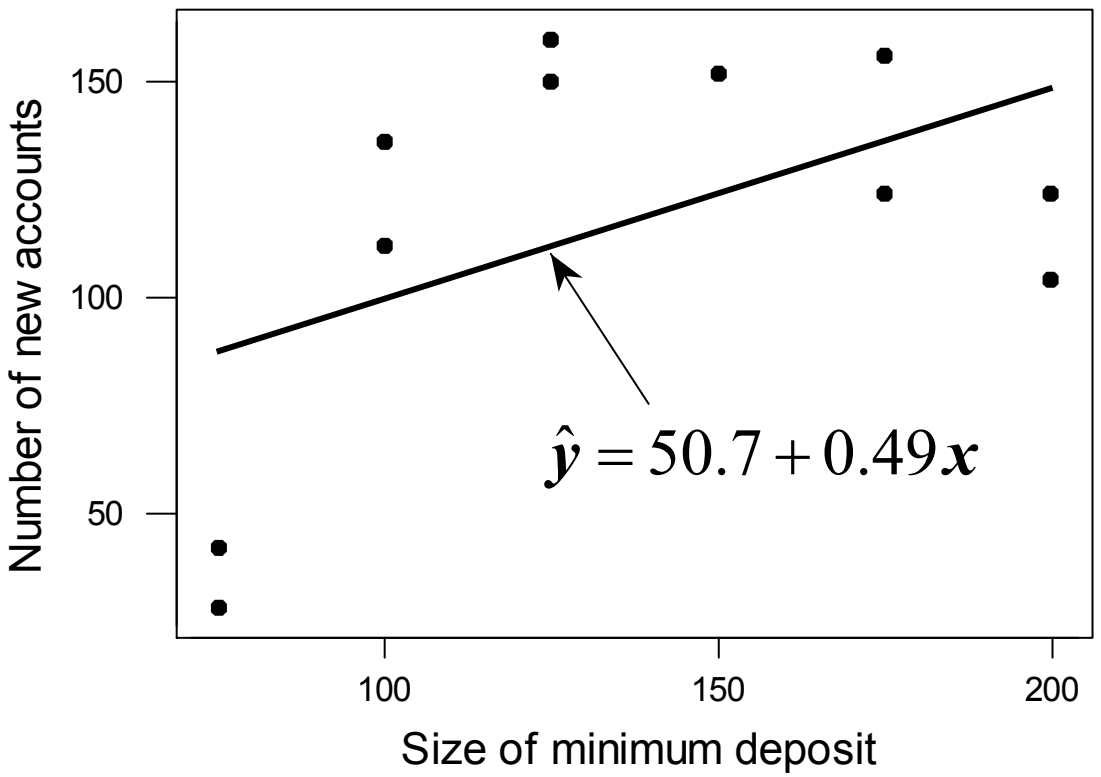
"**Error sum of squares**"

$$\boldsymbol{SSE(R) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(Y_{ij} - \hat{Y}_{ij} \right)^2 = SSE}$$

Decomposing the error



Decomposing the error



Error sum of squares decomposition

$$\left(Y_{ij} - \hat{Y}_{ij}\right) = \left(Y_{ij} - \overline{Y}_{j}\right) + \left(\overline{Y}_{j} - \hat{Y}_{ij}\right)$$

error deviation **pure error deviation** **lack of fit deviation**

$$\sum_{j}\sum_{i}\left(Y_{ij} - \hat{Y}_{ij}\right)^2 = \sum_{j}\sum_{i}\left(Y_{ij} - \overline{Y}_{j}\right)^2 + \sum_{j}\sum_{i}\left(\overline{Y}_{j} - \hat{Y}_{ij}\right)^2$$

$$SSE = SSPE + SSLF$$

The general linear test

$$F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F} \right) \div \left(\frac{SSE(F)}{df_F} \right)$$

$$df_R = n - 2 \qquad SSE(R) = SSE$$

$$df_F = n - c \qquad SSE(F) = SSPE$$

$$F^* = \left(\frac{SSE - SSPE}{(n-2) - (n-c)} \right) \div \left(\frac{SSPE}{(n-2)} \right) = \left(\frac{SSLF}{c-2} \right) \div \left(\frac{SSPE}{(n-2)} \right) = \frac{MSLF}{MSPE}$$

The test (intuitively)

- If the largest portion of the error sum of squares is due to lack of fit, the F test should be large.
- A large F* statistic leads to a small P-value (determined by $F(c-2, n-2)$ distribution).
- If the P-value is small, reject the null and conclude significant lack of (linear) fit.

The formal LOF test

Null hypothesis H_0: (Reduced) $Y_{ij} = \beta_0 + \beta_1 x_j + \varepsilon_{ij}$

Alternative hypothesis H_A: (Full) $Y_{ij} = \mu_j + \varepsilon_{ij}$

Test statistic $F^* = \dfrac{MSLF}{MSPE}$

P-value = What is the probability that we'd get an F* statistic as large as we did, if the null hypothesis is true? (One-tailed test!)

The P-value is determined by comparing F* to an **F distribution** with *c-2* **numerator degrees of freedom** and *n-2* **denominator degrees of freedom**.

LOF Test summarized in an ANOVA Table

TABLE 3.6 **General ANOVA Table for Testing Lack of Fit of Simple Linear Regression Function and ANOVA Table for Bank Example.**

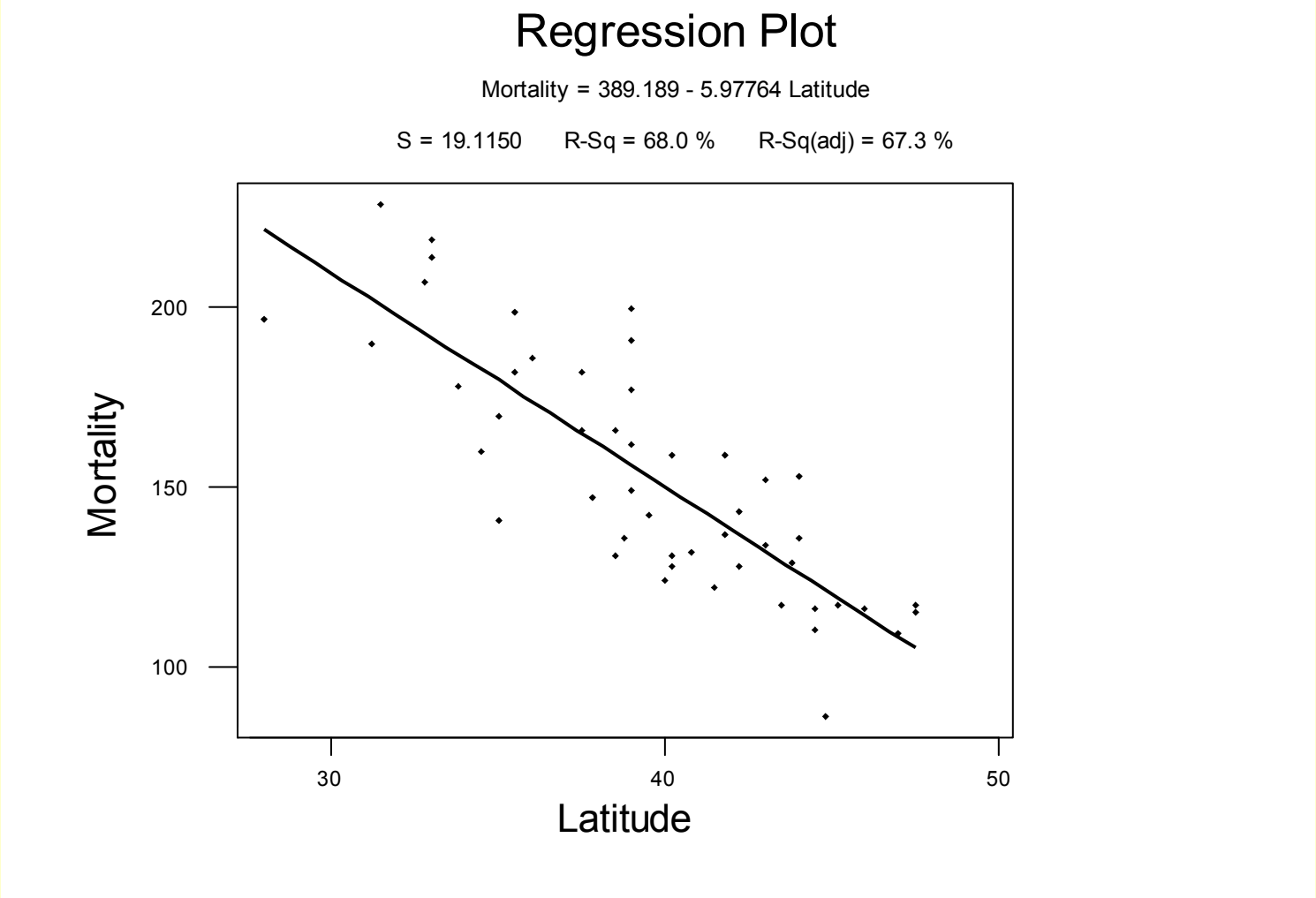
(a) General

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = \Sigma\Sigma(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \Sigma\Sigma(Y_{ij} - \hat{Y}_{ij})^2$	$n - 2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \Sigma\Sigma(\bar{Y}_j - \hat{Y}_{ij})^2$	$c - 2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \Sigma\Sigma(Y_{ij} - \bar{Y}_j)^2$	$n - c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \Sigma\Sigma(Y_{ij} - \bar{Y})^2$	$n - 1$	

LOF Test

- **Stat** >> **Regression** >> **Regression …**
- Specify predictor and response.
- Under Options…
  - under **Lack of Fit Tests**, select the box labeled **Pure error**.
- Select OK. Select OK.

Example 1



Do the data suggest that a linear function is not adequate in describing the relationship between skin cancer mortality and latitude?

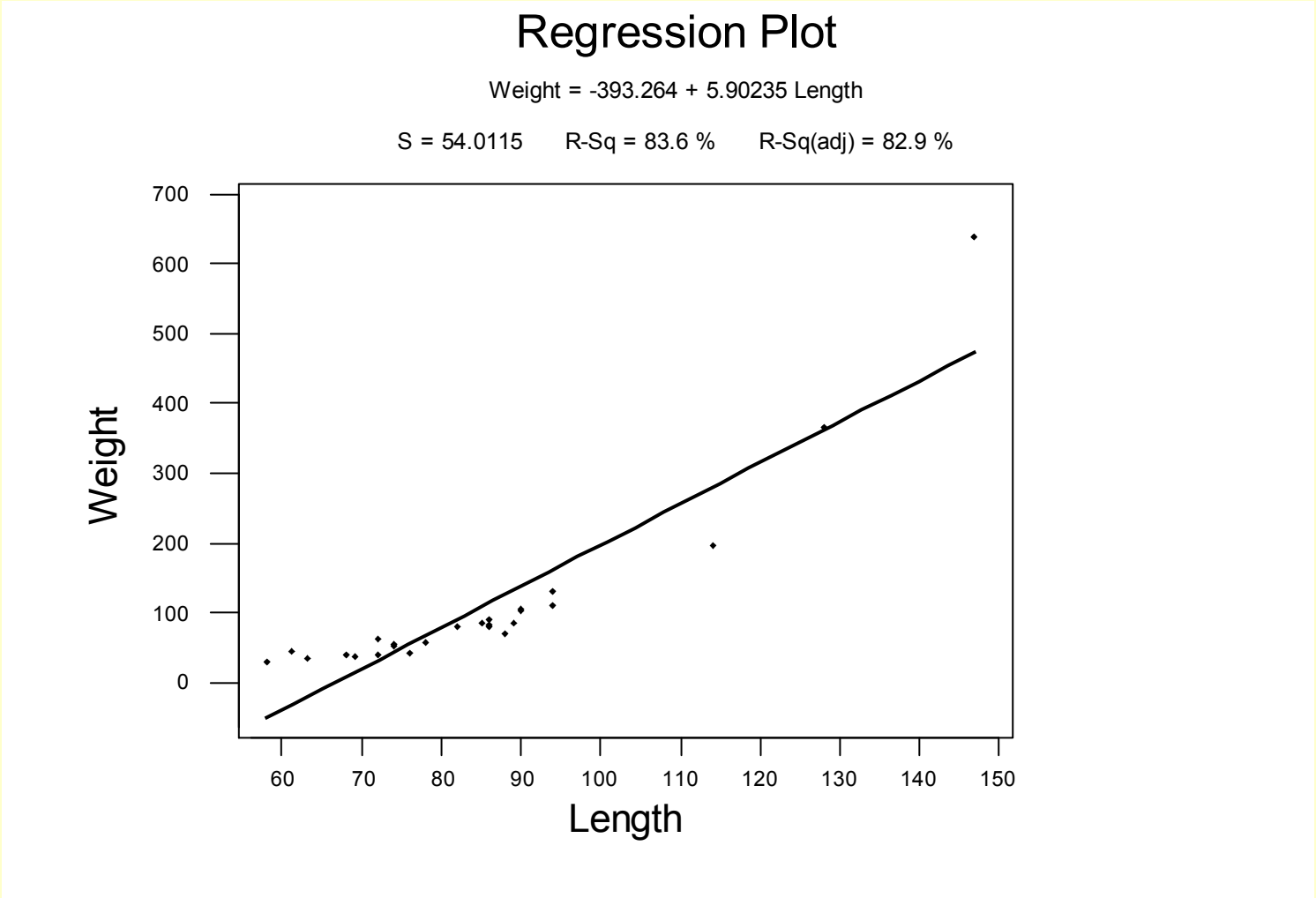
Example 1: Mortality and Latitude

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	**1**	**36464**	**36464**	**99.80**	**0.000**
Residual Error	**47**	**17173**	**365**		
Lack of Fit	**30**	**12863**	**429**	**1.69**	**0.128**
Pure Error	**17**	**4310**	**254**		
Total	**48**	**53637**			

19 rows with no replicates

Example 2



Do the data suggest that a linear function is not adequate in describing the relationship between the length and weight of an alligator?

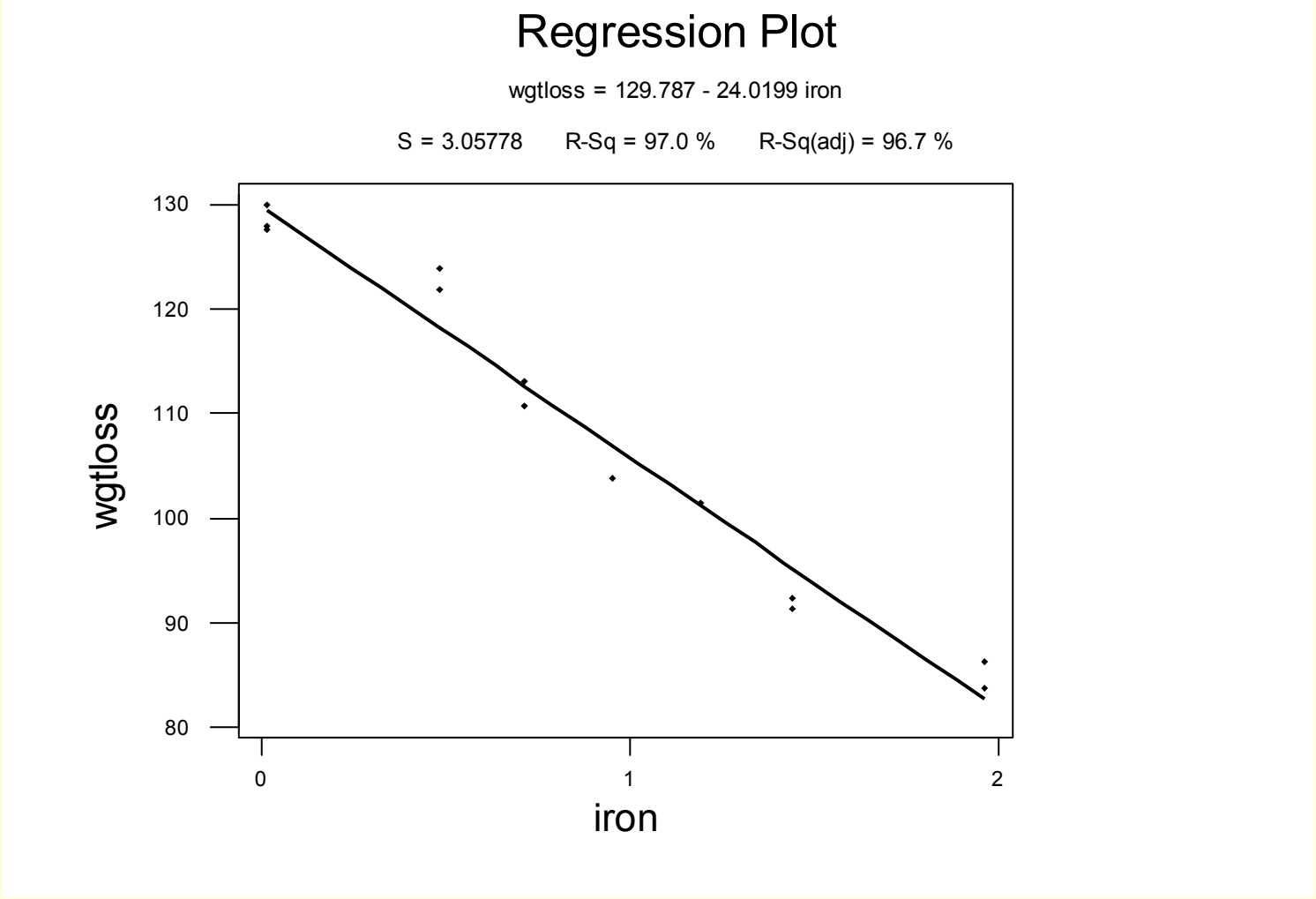
Example 2: Alligator length and weight

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	**1**	**342350**	**342350**	**117.35**	**0.000**
Residual Error	**23**	**67096**	**2917**		
Lack of Fit	**17**	**66567**	**3916**	**44.36**	**0.000**
Pure Error	**6**	**530**	**88**		
Total	**24**	**409446**			

14 rows with no replicates

Example 3



Do the data suggest that a linear function is not adequate in describing the relationship between iron content and weight loss due to corrosion?

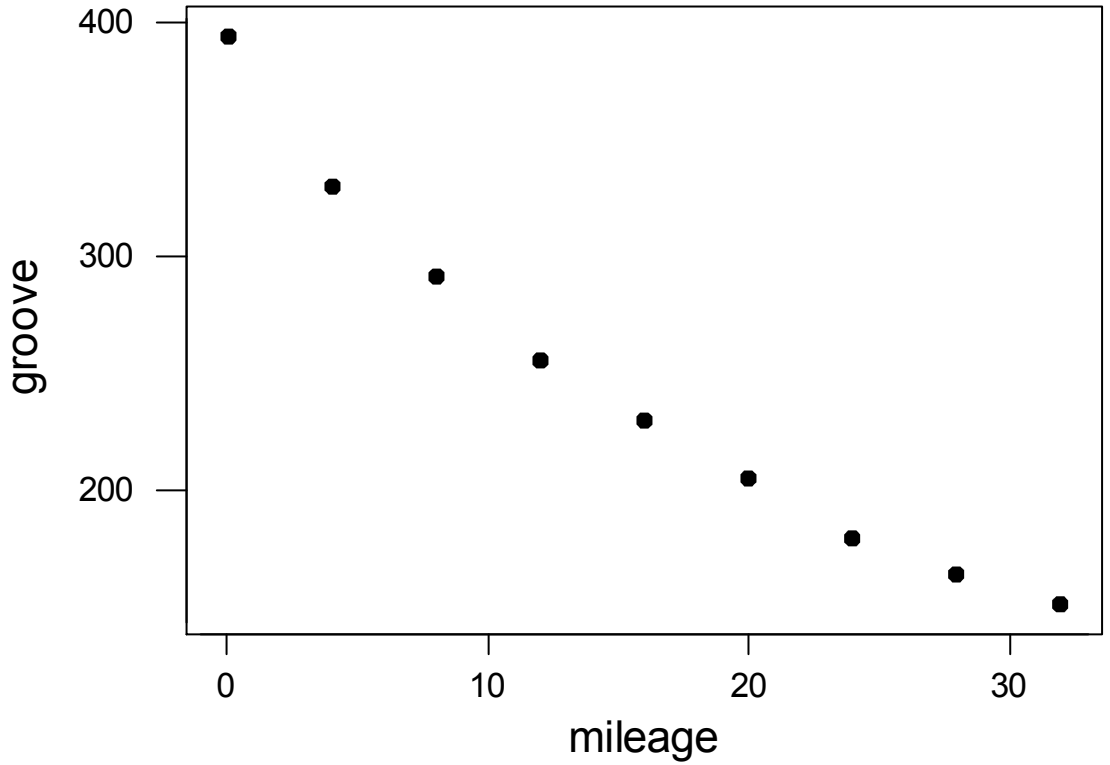
Example 3: Iron and corrosion

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	**1**	**3293.8**	**3293.8**	**352.27**	**0.000**
Residual Error	**11**	**102.9**	**9.4**		
Lack of Fit	**5**	**91.1**	**18.2**	**9.28**	**0.009**
Pure Error	**6**	**11.8**	**2.0**		
Total	**12**	**3396.6**			

2 rows with no replicates

Example 4



Do the data suggest that a linear function is not adequate in describing the relationship between mileage and groove depth?

Example 4: Tread wear

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	**1**	**50887**	**50887**	**140.71**	**0.000**
Residual Error	**7**	**2532**	**362**		
Total	**8**	**53419**			

No replicates. Cannot do pure error test.

Model Checking

Using residuals to check the validity of the linear regression model assumptions

The simple linear regression model

- The mean of the responses, $E(Y_i)$, is a **linear function** of the x_i.
- The errors, ε_i, and hence the responses Y_i, are **independent**.
- The errors, ε_i, and hence the responses Y_i, are **normally distributed**.
- The errors, ε_i, and hence the responses Y_i, have **equal variances** (σ^2) for all x values.

The simple linear regression model

Assume (**!!**) response is **linear** function of trend and error:

$$Y_i = \left(\beta_0 + \beta_1 X_i\right) + \varepsilon_i$$

with the **independent** error terms ε_i following a **normal** distribution with mean 0 and **equal variance** σ^2.

Why do we have to check our model?

- All estimates, intervals, and hypothesis tests have been developed assuming that the model is correct.
- If the model is incorrect, then the formulas and methods we use are at risk of being incorrect.

When should we worry most?

- All tests and intervals are very sensitive to
  - departures from independence.
  - moderate departures from equal variance.
- Tests and intervals for β_0 and β_1 are fairly **robust** against departures from normality.
- Prediction intervals are quite sensitive to departures from normality.

What can go wrong with the model?

- Regression function is **not linear**.
- Error terms are **not independent**.
- Error terms are **not normal**.
- Error terms do **not** have **equal variance**.
- The model fits all but one or a few outlier observations.
- An important predictor variable has been left out of the model.

The basic idea of residual analysis

The observed **residuals**:

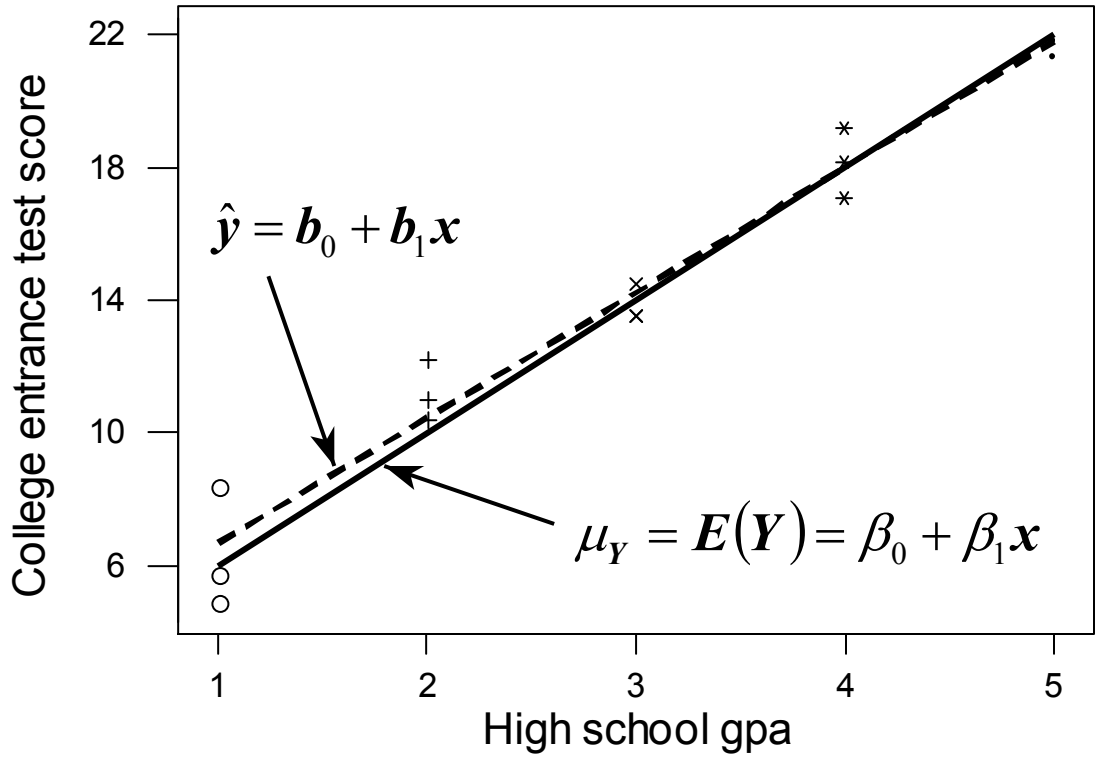
$$e_i = y_i - \hat{y}_i$$

should reflect the properties assumed for the unknown true error terms:

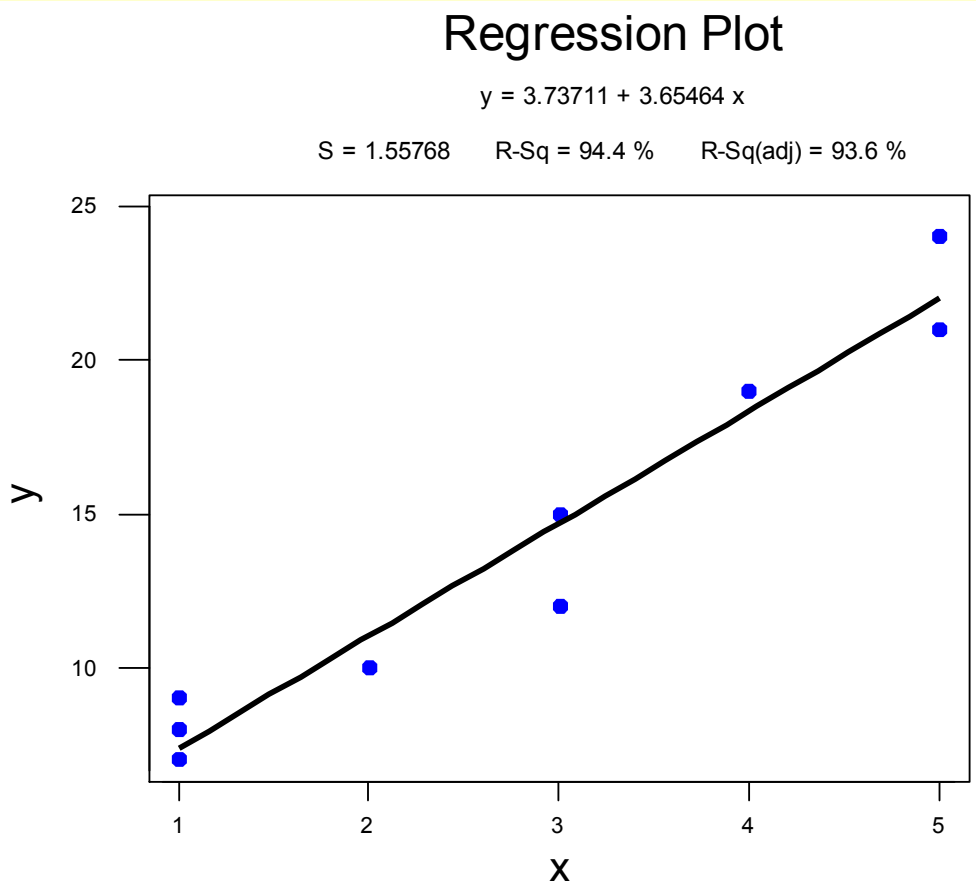
$$\varepsilon_i = Y_i - E(Y_i)$$

So, investigate the observed residuals to see if they behave "properly."

Distinction between true errors ε_i and residuals e_i



The sample mean of the residuals e_i is always 0.

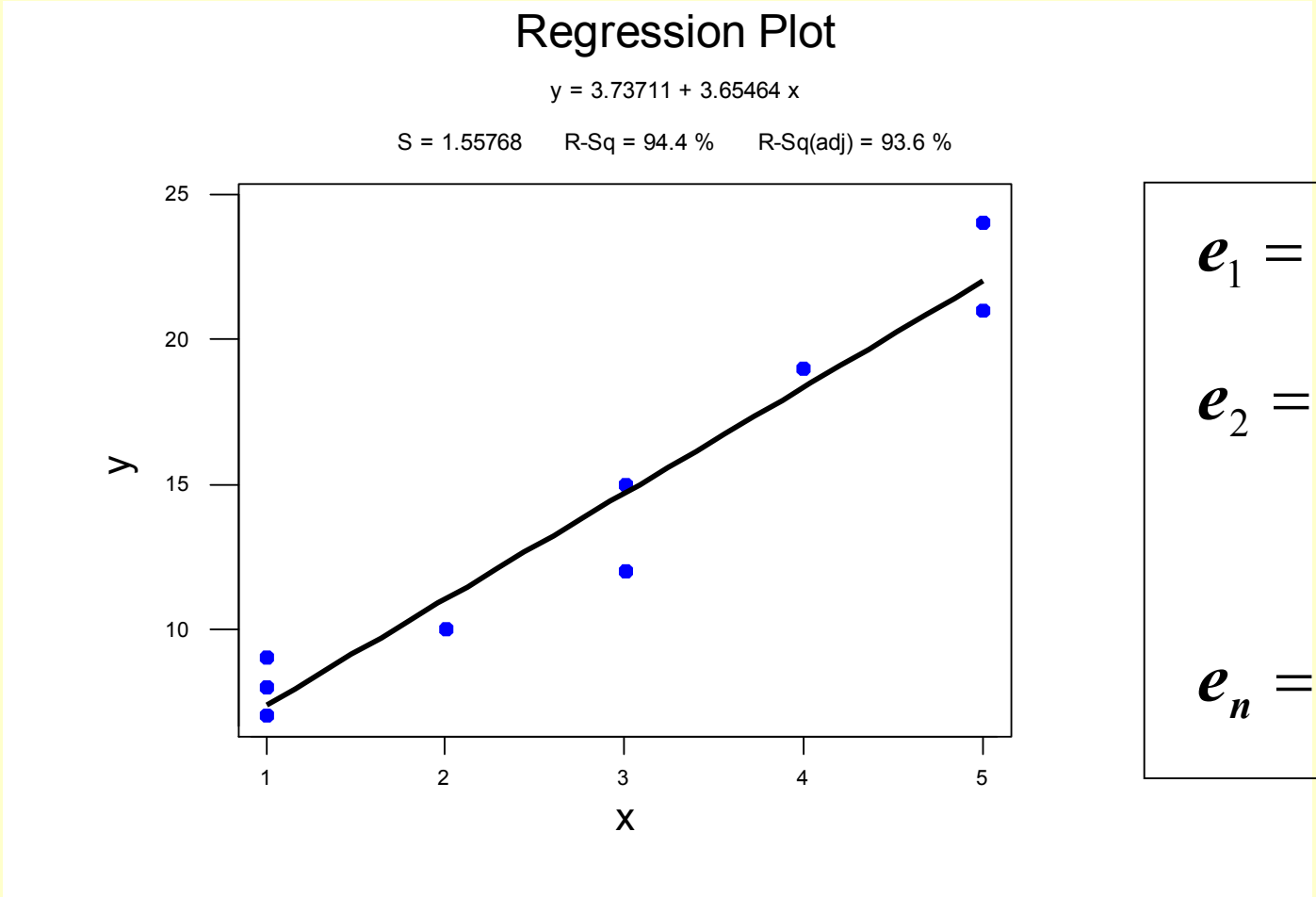


x	y	RESIDUAL
1	9	1.60825
1	7	-0.39175
1	8	0.60825
2	10	-1.04639
3	15	0.29897
3	12	-2.70103
4	19	0.64433
5	24	1.98969
5	21	-1.01031

		0.00001

(round-off error)

The residuals are not independent.



$$e_1 = y_1 - (b_0 + b_1 x)$$

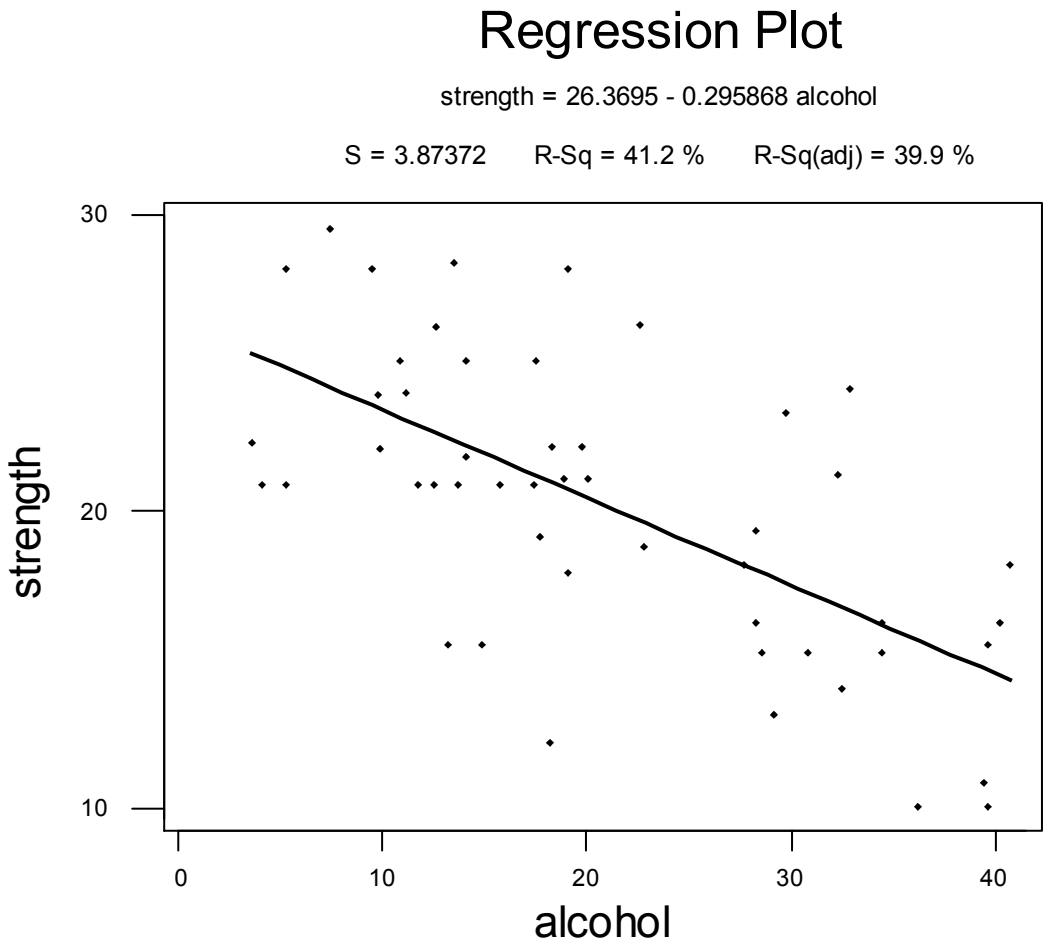
$$e_2 = y_2 - (b_0 + b_1 x)$$

$$e_n = y_n - (b_0 + b_1 x)$$

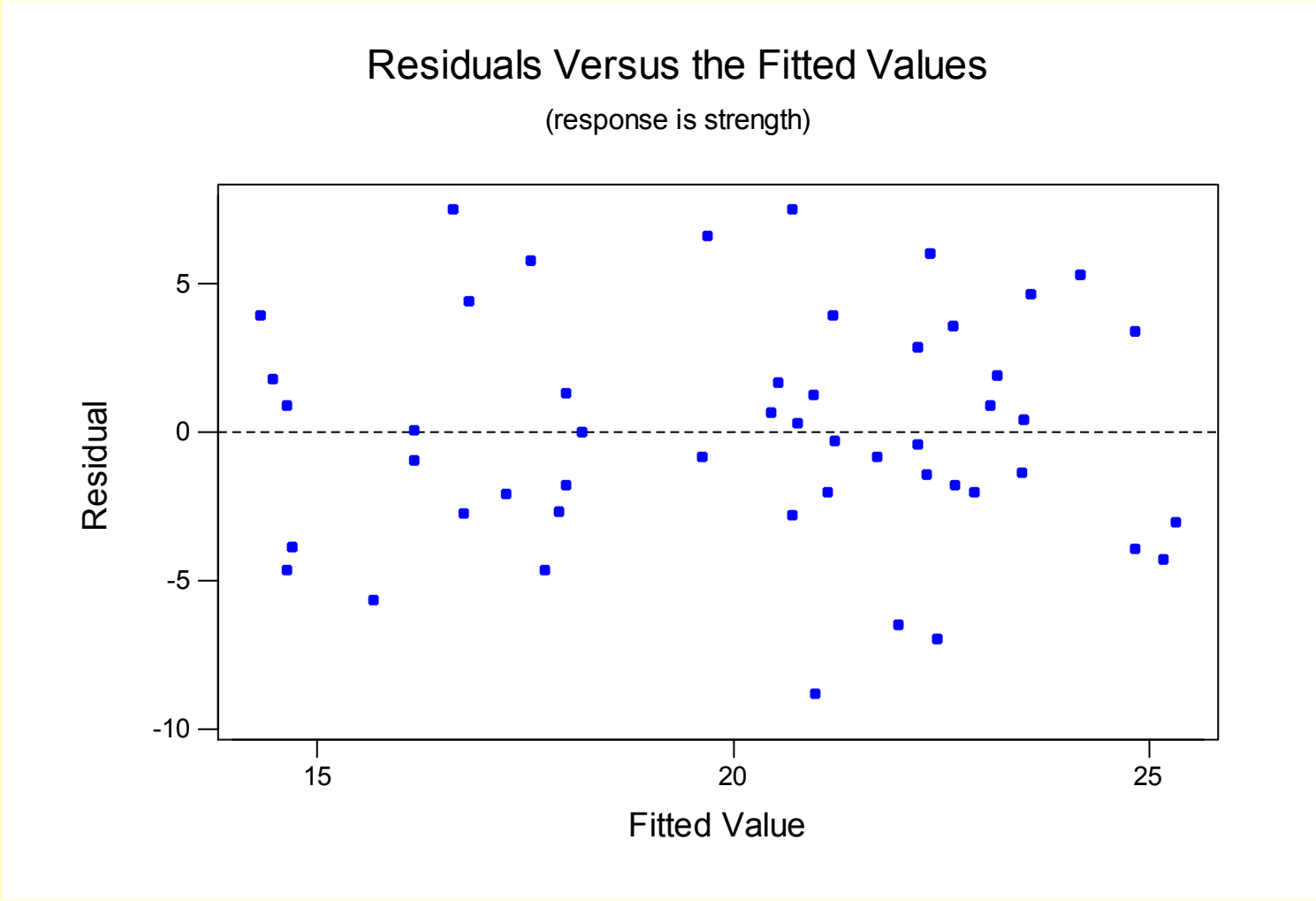
A **residuals vs. fits** plot

- A scatter plot with **residuals** on the *y* axis and **fitted values** on the *x* axis.
- Helps to identify non-linearity, outliers, and non-constant variance.

Example:
Alcoholism and muscle strength?



A well-behaved residuals vs. fits plot



Characteristics of a **well-behaved** residual vs. fits plot

- The residuals "bounce randomly" around the 0 line. (Linear is reasonable).
- No one residual "stands out" from the basic random pattern of residuals. (No outliers).
- The residuals roughly form a "horizontal band" around 0 line. (Constant variance).

Durbin-Watson Procedure

- 1. Used to Detect Autocorrelation
  - Residuals in one time period are related to residuals in another period
  - Violation of independence assumption
- 2. Durbin-Watson Test Statistic

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Durbin-Watson Rules

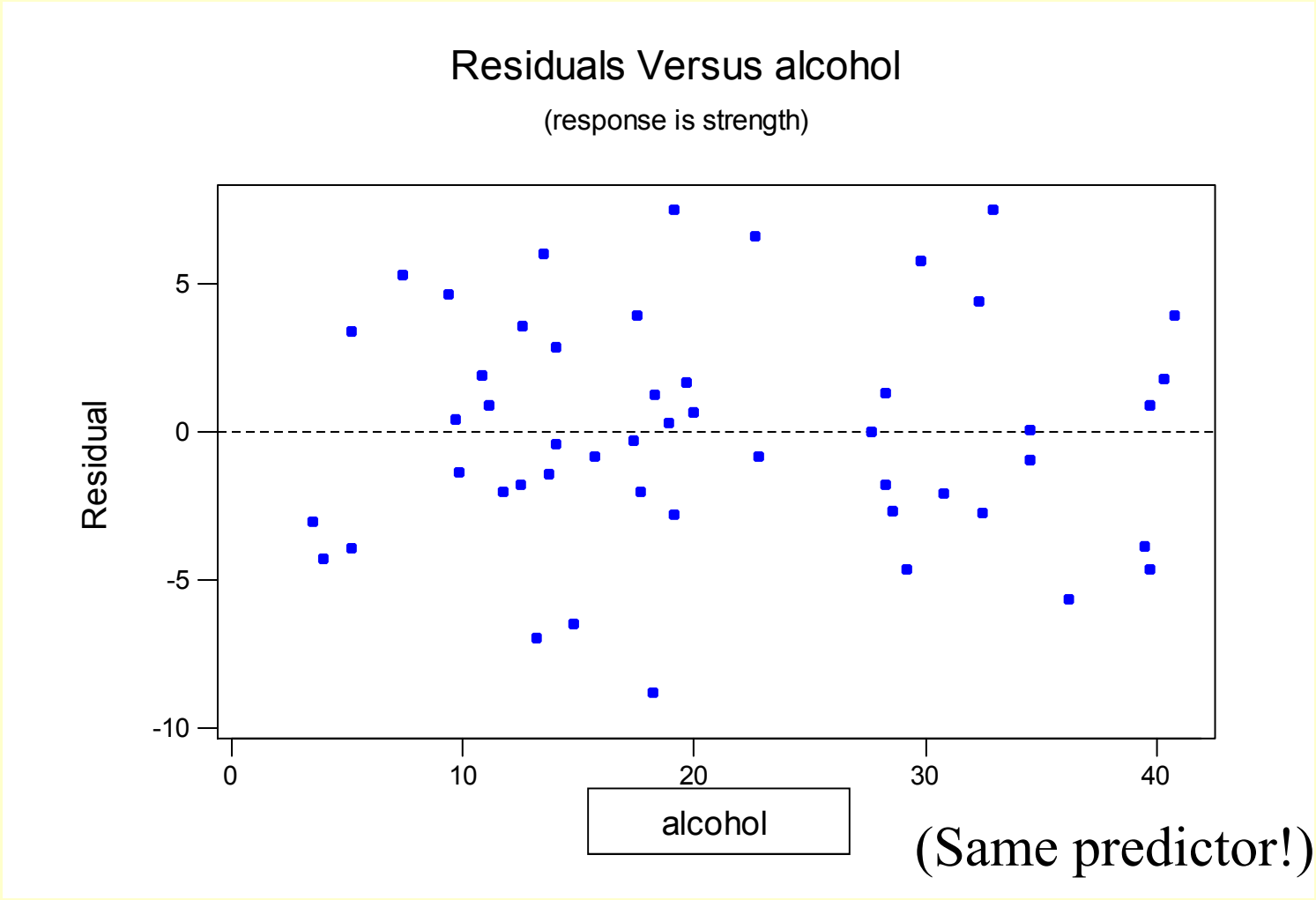
- For given α, n, & p:
- If $D < d_L$, then auto-correlation exists
- If $D > d_U$, then no auto-correlation exists
- If $d_L < D < d_U$, then no definite conclusion

		X variables, excluding the intercept									
Observations		1		2		3		4		5	
N	Prob.	D-L	D-U	D-L	D-U	D-L	D-U	D-L	D-U	D-L	D-U
15	0.05	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
	0.01	0.81	1.07	0.7	1.25	0.59	1.46	0.49	1.70	0.39	1.96
20	0.05	1.20	1.71	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
	0.01	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
25	0.05	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
	0.01	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
30	0.05	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
	0.01	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
40	0.05	1.44	1.54	1.39	1.60	1.34	1.66	1.39	1.72	1.23	1.79
	0.01	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
50	0.05	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
	0.01	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
60	0.05	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
	0.01	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
80	0.05	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
	0.01	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
100	0.05	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78
	0.01	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

A **residuals vs. predictor** plot

- A scatter plot with **residuals** on the *y* axis and the values of a **predictor** on the *x* axis.
- If the predictor on the *x* axis is the same predictor used in model, offers nothing new.
- If the predictor on the *x* axis is a new and different predictor, can help to determine whether the predictor should be added to model.

A residuals vs. predictor plot offering nothing new.

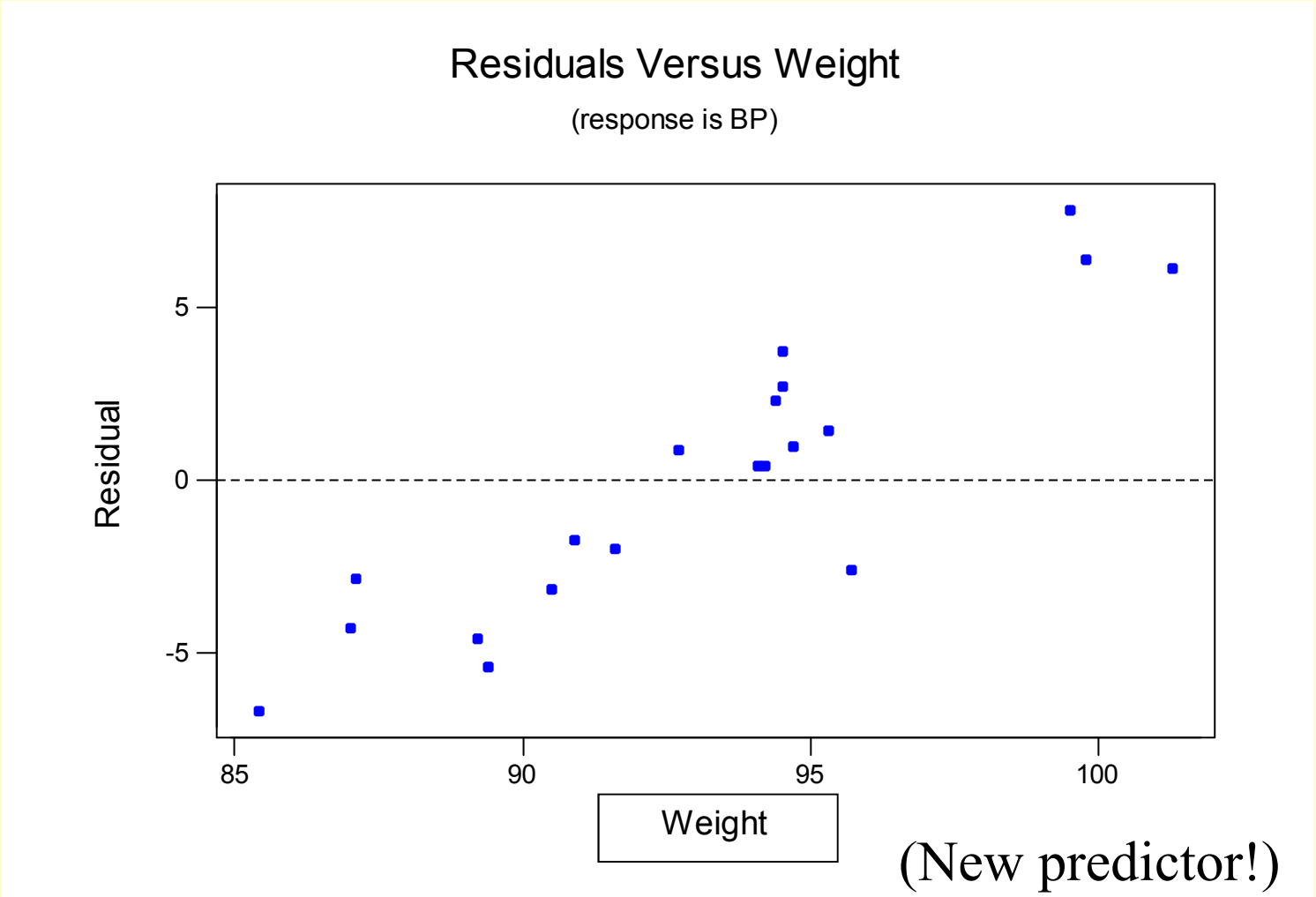


(Same predictor!)

Example: What are good predictors of blood pressure?

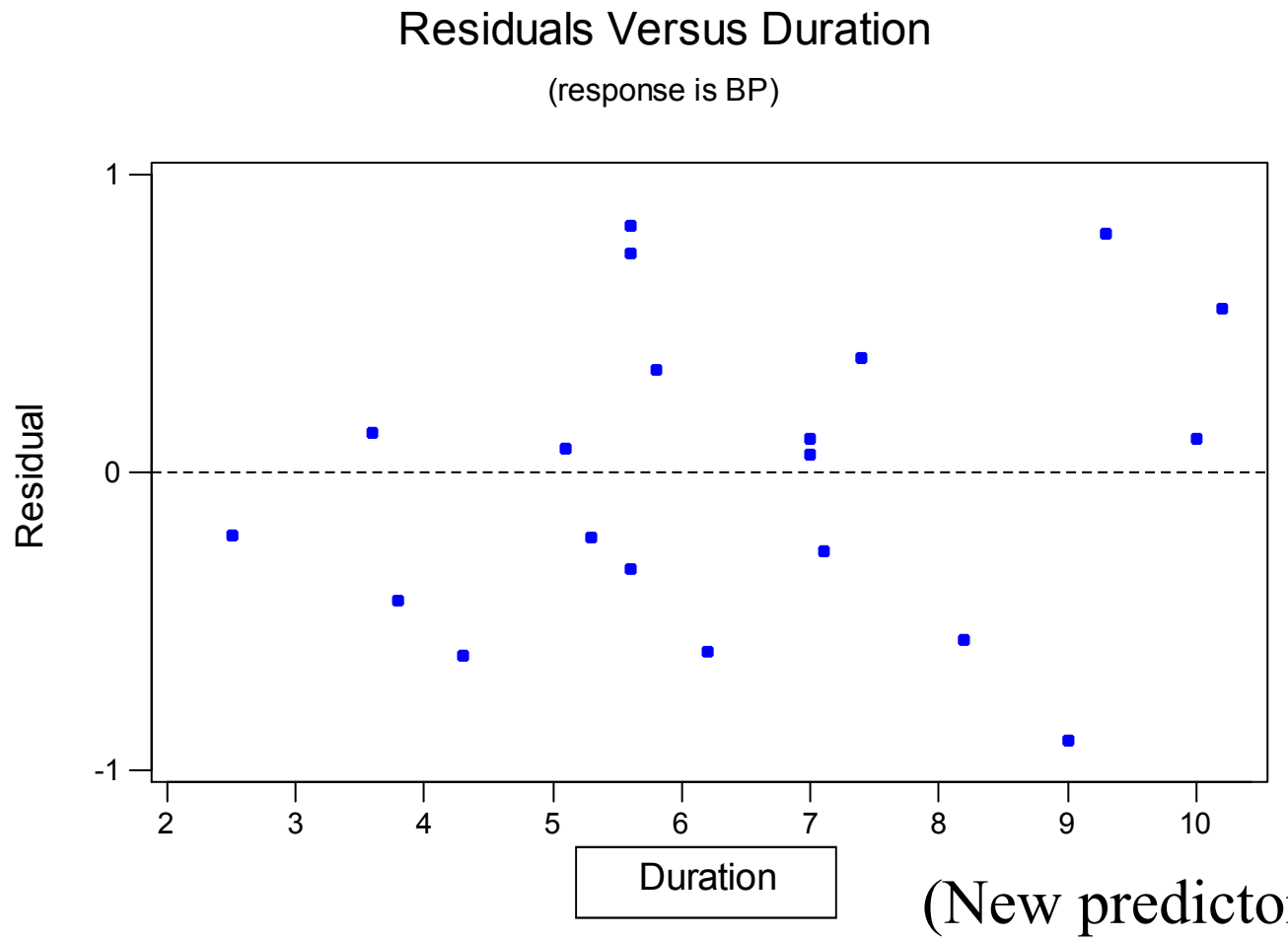
- n = 20 hypertensive individuals
- **weight** = weight of individual
- **duration** = years with high blood pressure

Residuals (age only) vs. weight plot



(New predictor!)

Residuals (age, weight) vs. duration plot



(New predictor!)

How a **non-linear function** shows up on a residual vs. fits plot

- The residuals depart from 0 in some systematic manner:
  - such as, being positive for small x values, negative for medium x values, and positive again for large x values

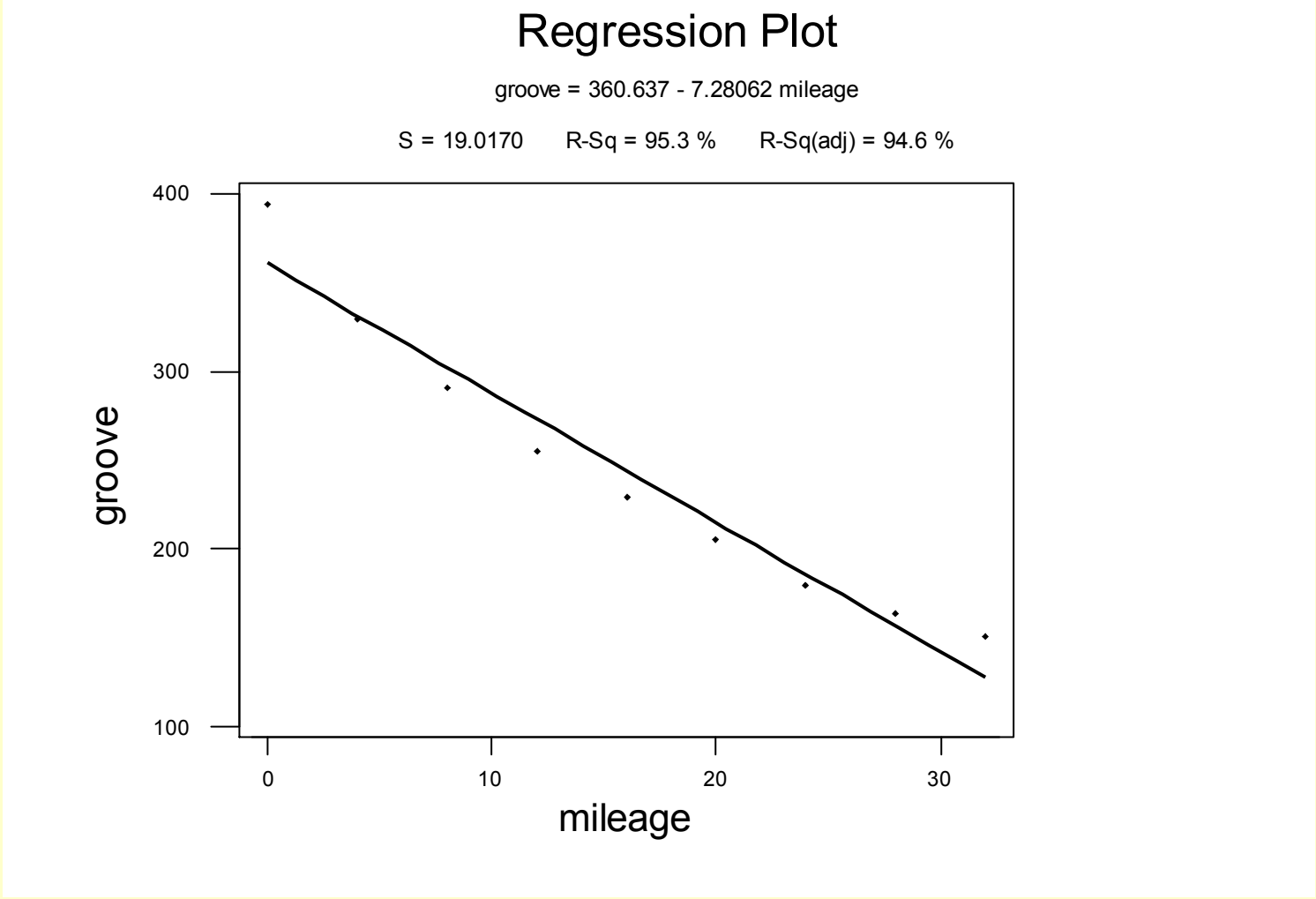
Example: A linear relationship between tread wear and mileage?

mileage	groove
0	394.33
4	329.50
8	291.00
12	255.17
16	229.33
20	204.83
24	179.00
28	163.83
32	150.33

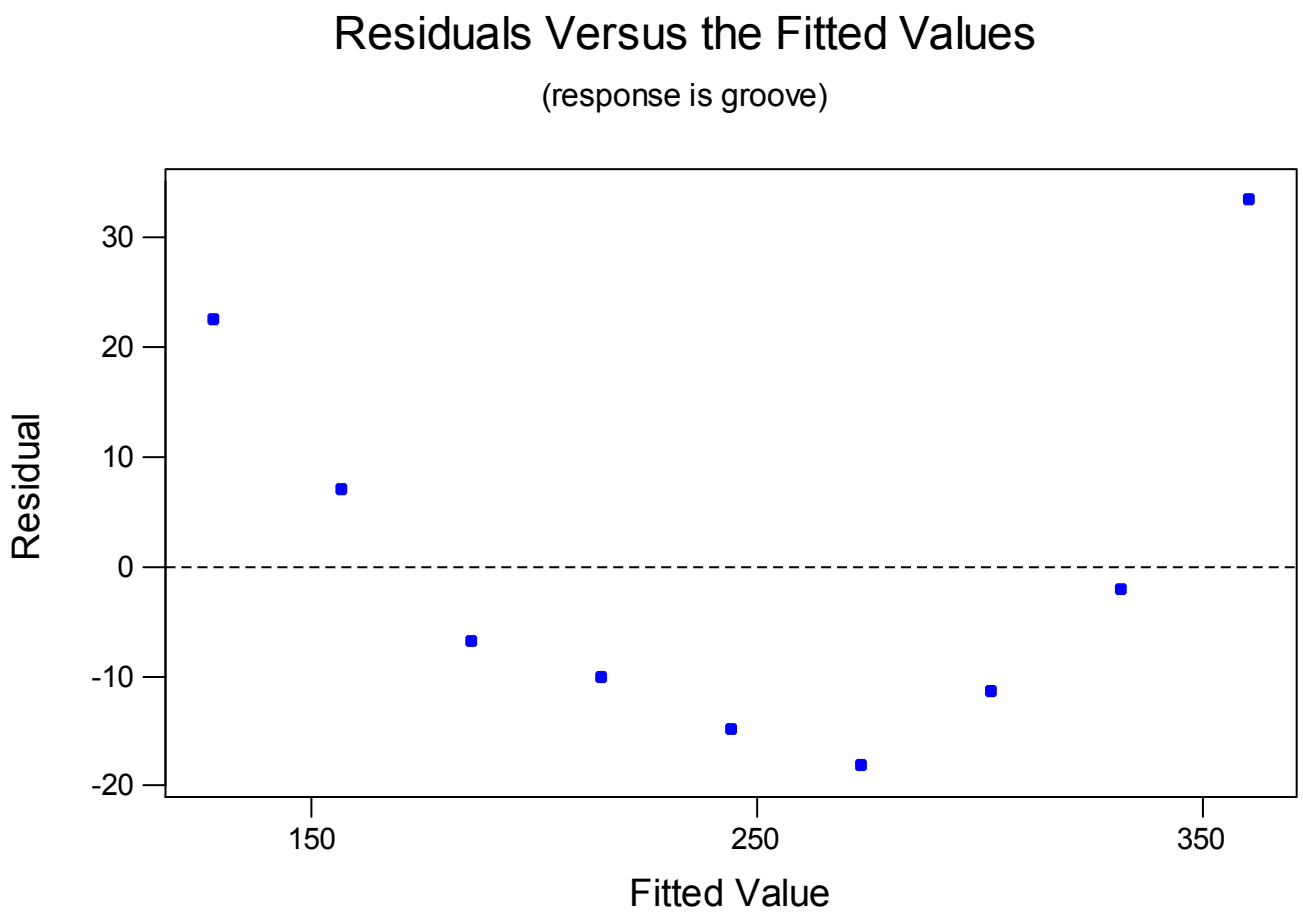
X = mileage in 1000 miles

Y = groove depth in mils

Is tire tread wear linearly related to mileage?



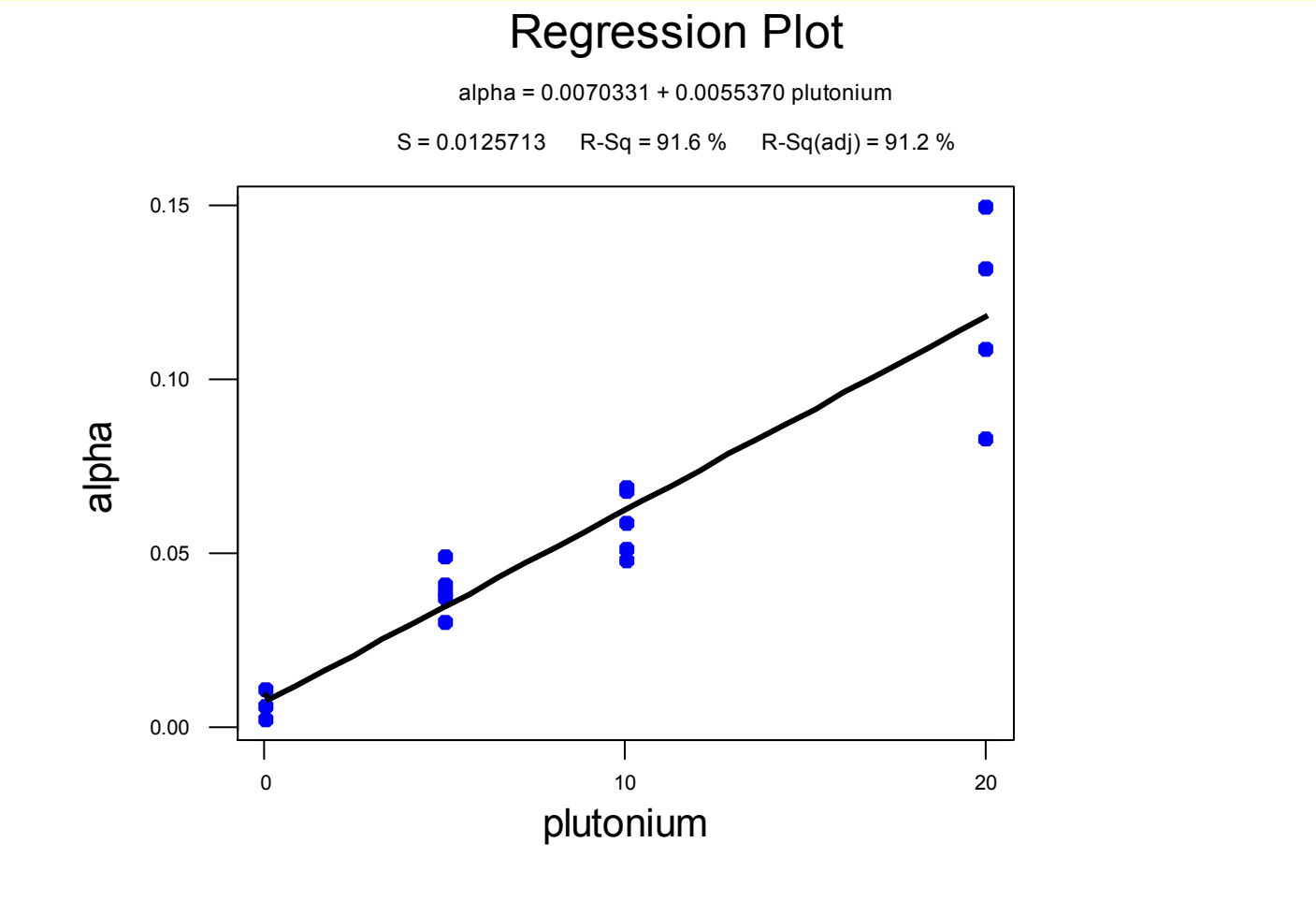
A residual vs. fits plot suggesting relationship is not linear



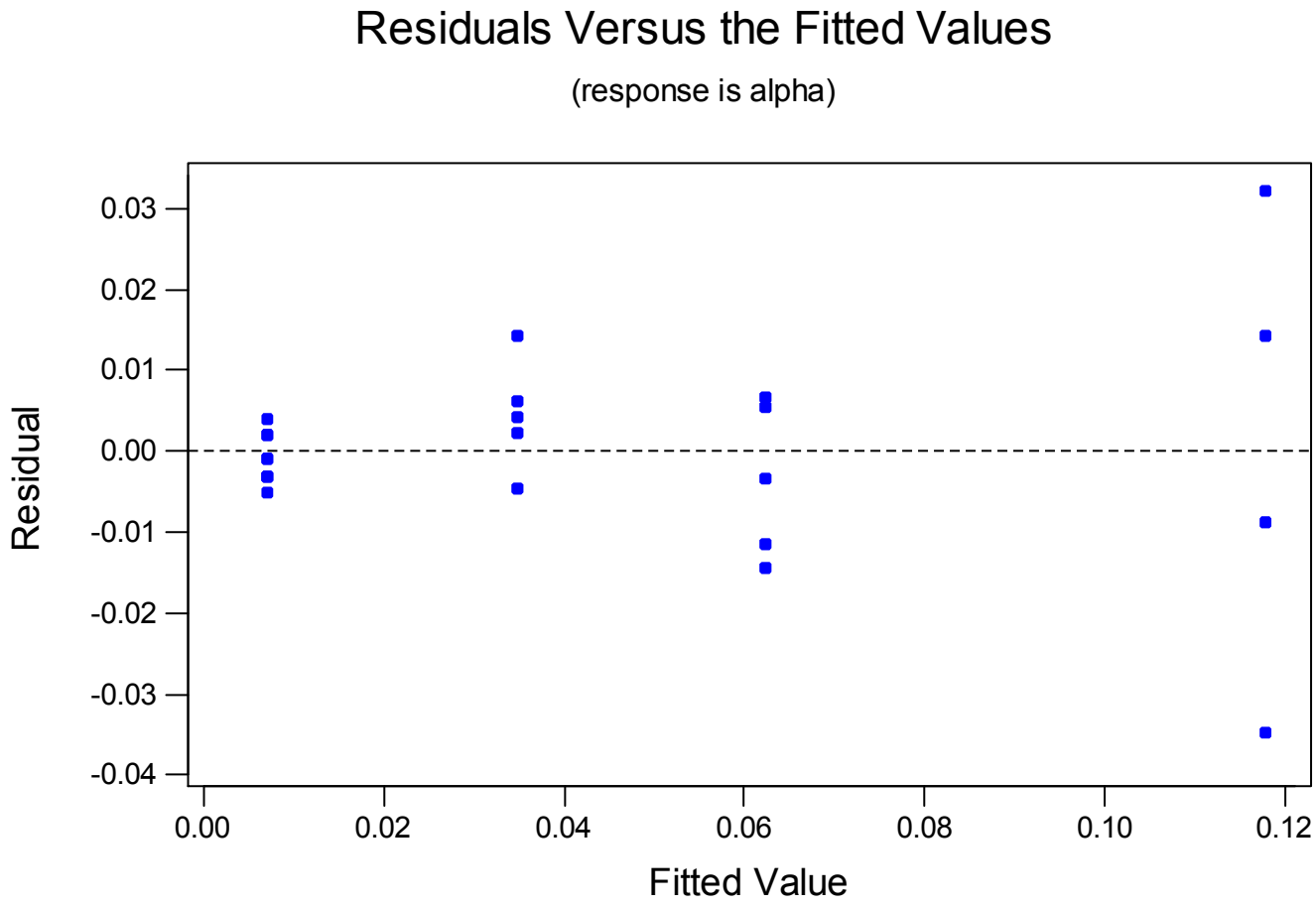
How **non-constant error variance** shows up on a residual vs. fits plot

- The plot has a "**fanning**" effect.
  - Residuals are close to 0 for small x values and are more spread out for large x values.
- The plot has a "**funneling**" effect
  - Residuals are spread out for small x values and close to 0 for large x values.
- Or, the spread of the residuals can vary in some complex fashion.

Example: How is plutonium activity related to alpha particle counts?



A residual vs. fits plot suggesting non-constant error variance



How an **outlier** shows up on a residuals vs. fits plot

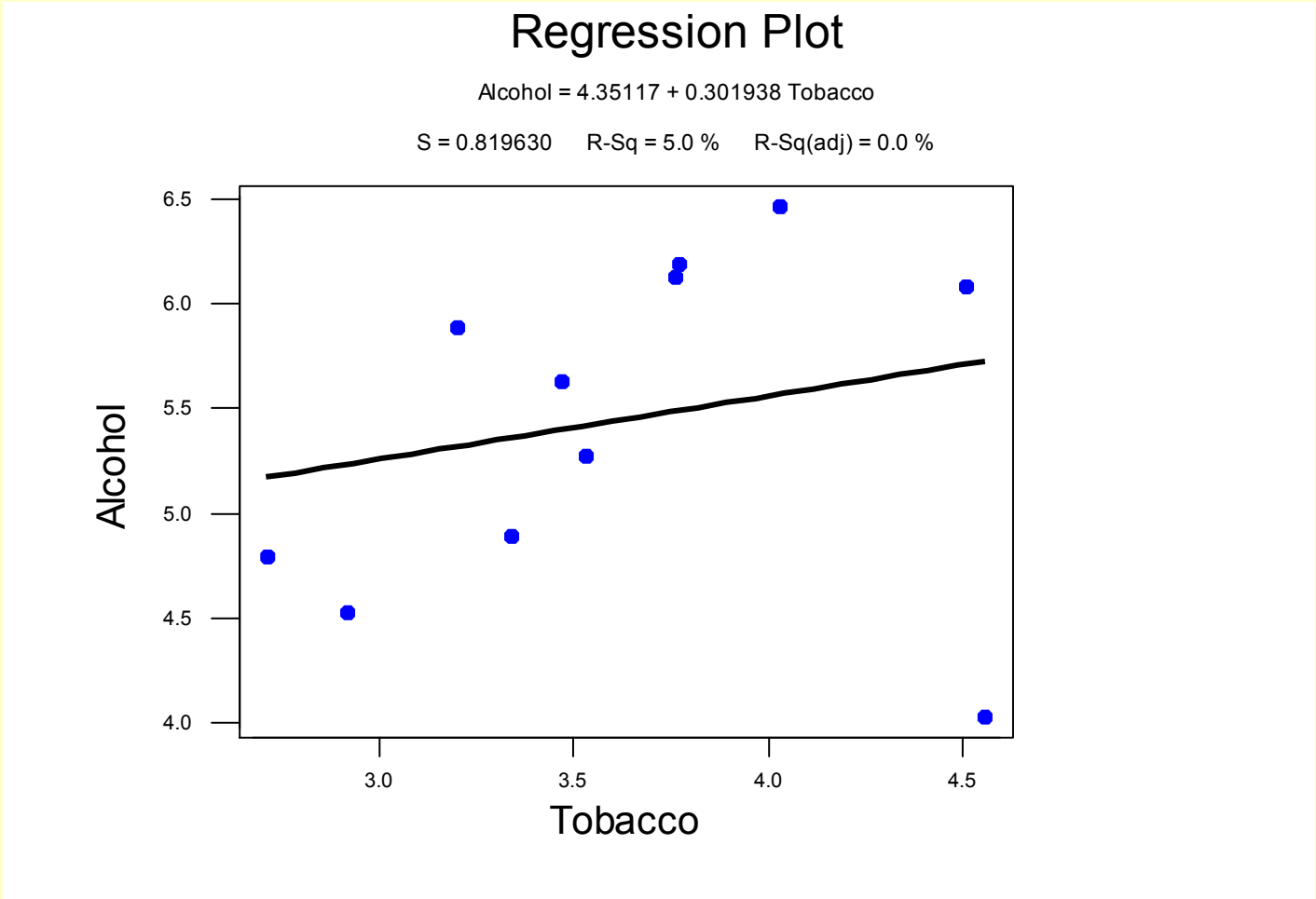
- The observation's residual stands apart from the basic random pattern of the rest of the residuals.
- The random pattern of the residual plot can even disappear if one outlier really deviates from the pattern of the rest of the data.

Example: Relationship between tobacco use and alcohol use?

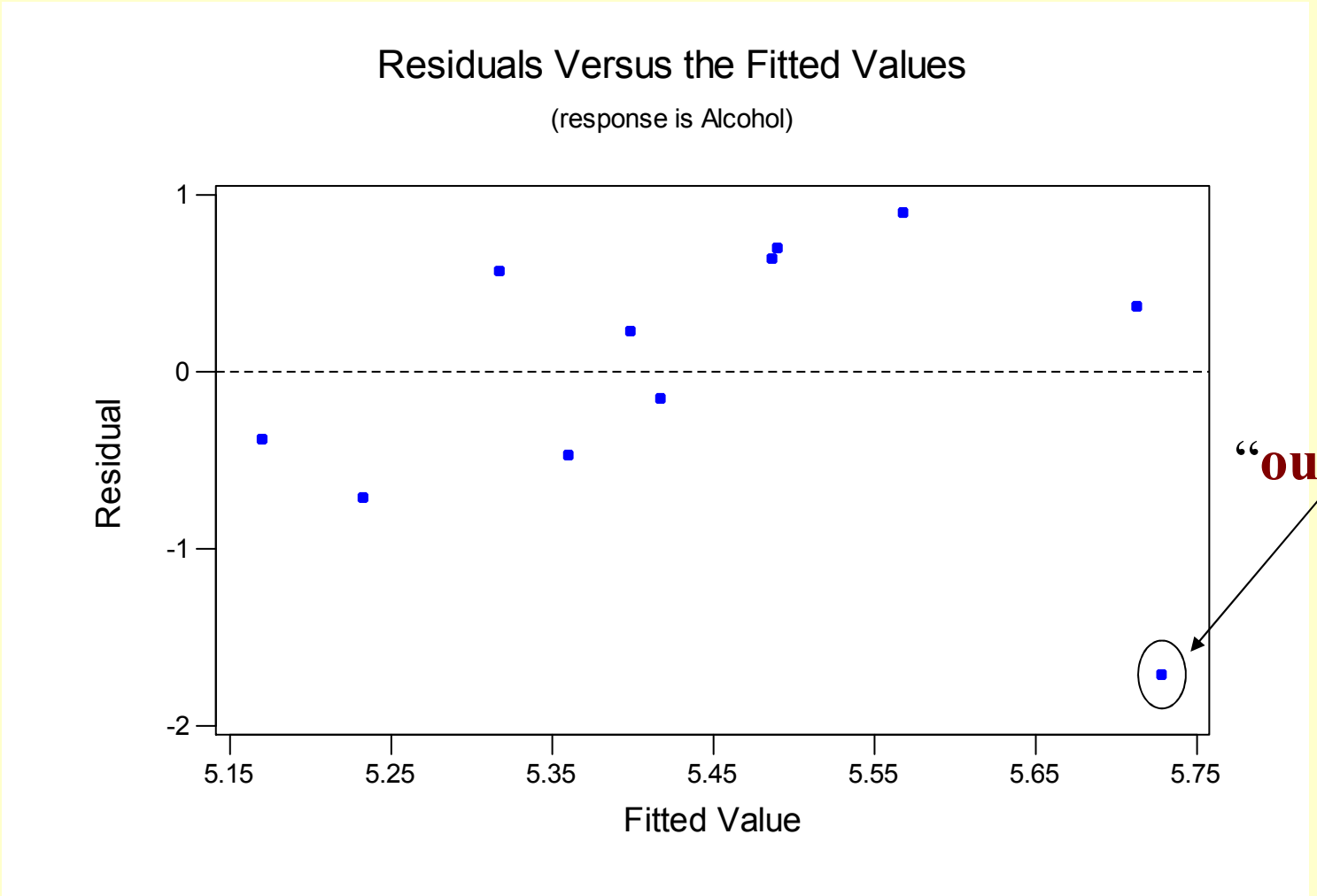
Region	Alcohol	Tobacco
North	6.47	4.03
Yorkshire	6.13	3.76
Northeast	6.19	3.77
EastMidlands	4.89	3.34
WestMidlands	5.63	3.47
EastAnglia	4.52	2.92
Southeast	5.89	3.20
Southwest	4.79	2.71
Wales	5.27	3.53
Scotland	6.08	4.51
Northern Ireland	4.02	4.56

- Family Expenditure Survey of British Dept. of Employment
- X = average weekly expenditure on tobacco
- Y = average weekly expenditure on alcohol

Example: Relationship between tobacco use and alcohol use?



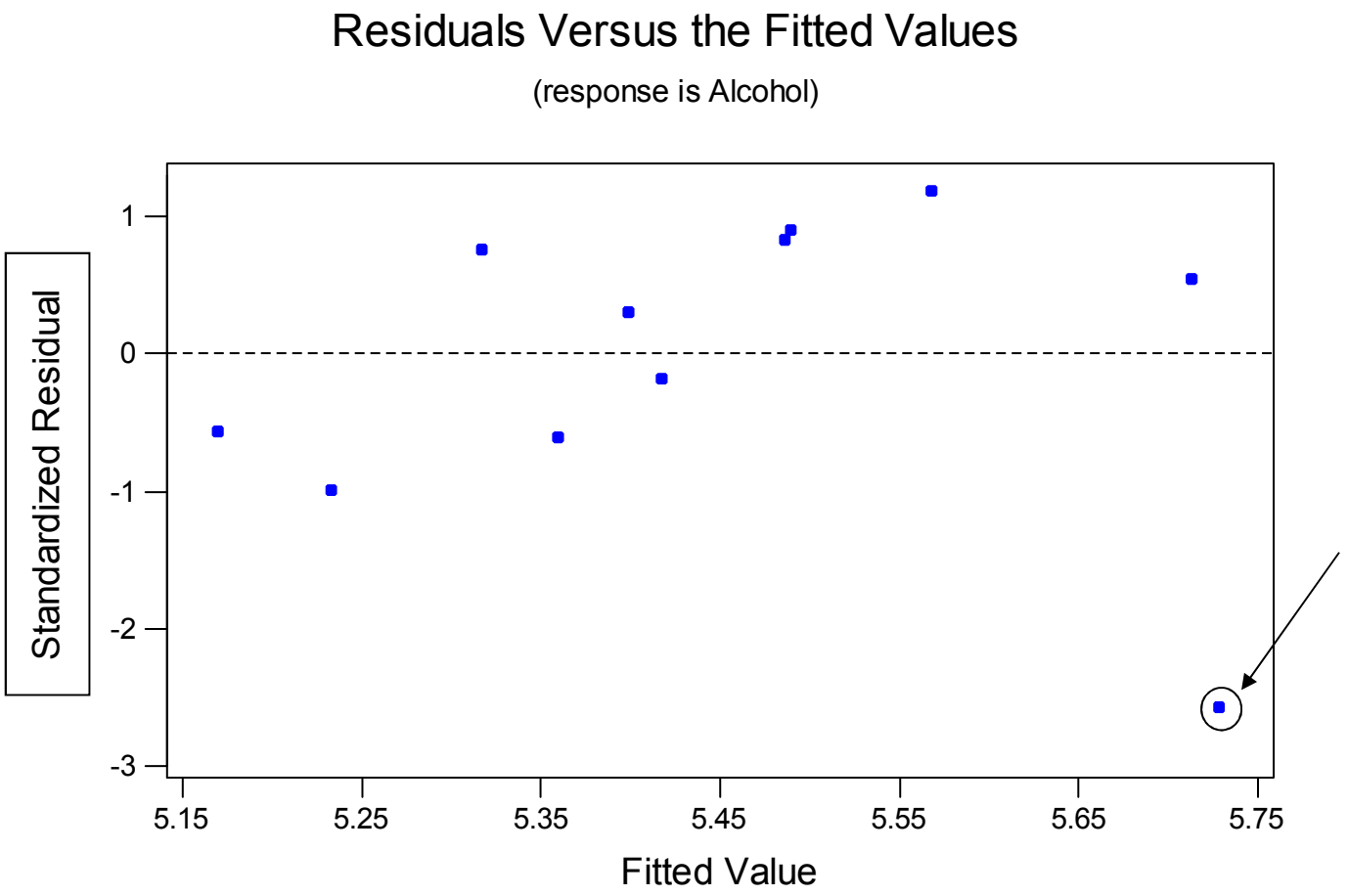
A residual vs. fits plot suggesting an outlier exists



How large does a residual need to be before being flagged?

- The magnitude of the residuals depends on the units of the response variable.
- Make the residuals "unitless" by dividing by their standard deviation. That is, use "**standardized residuals**."
- Then, an observation with a standardized residual greater than 2 or smaller than -2 should be flagged for further investigation.

Standardized residuals vs. fits plot

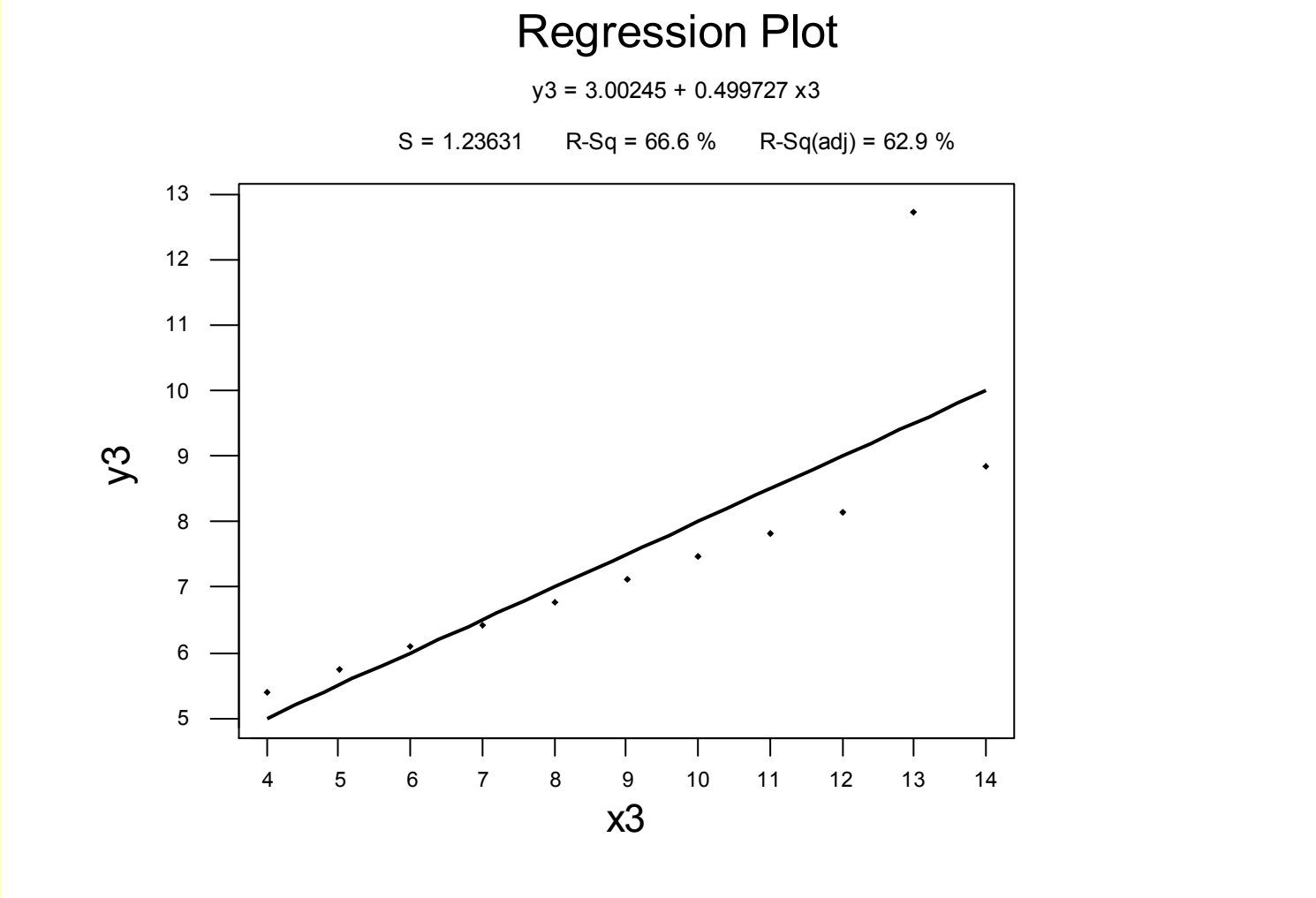


Identifies observations with large standardized residuals

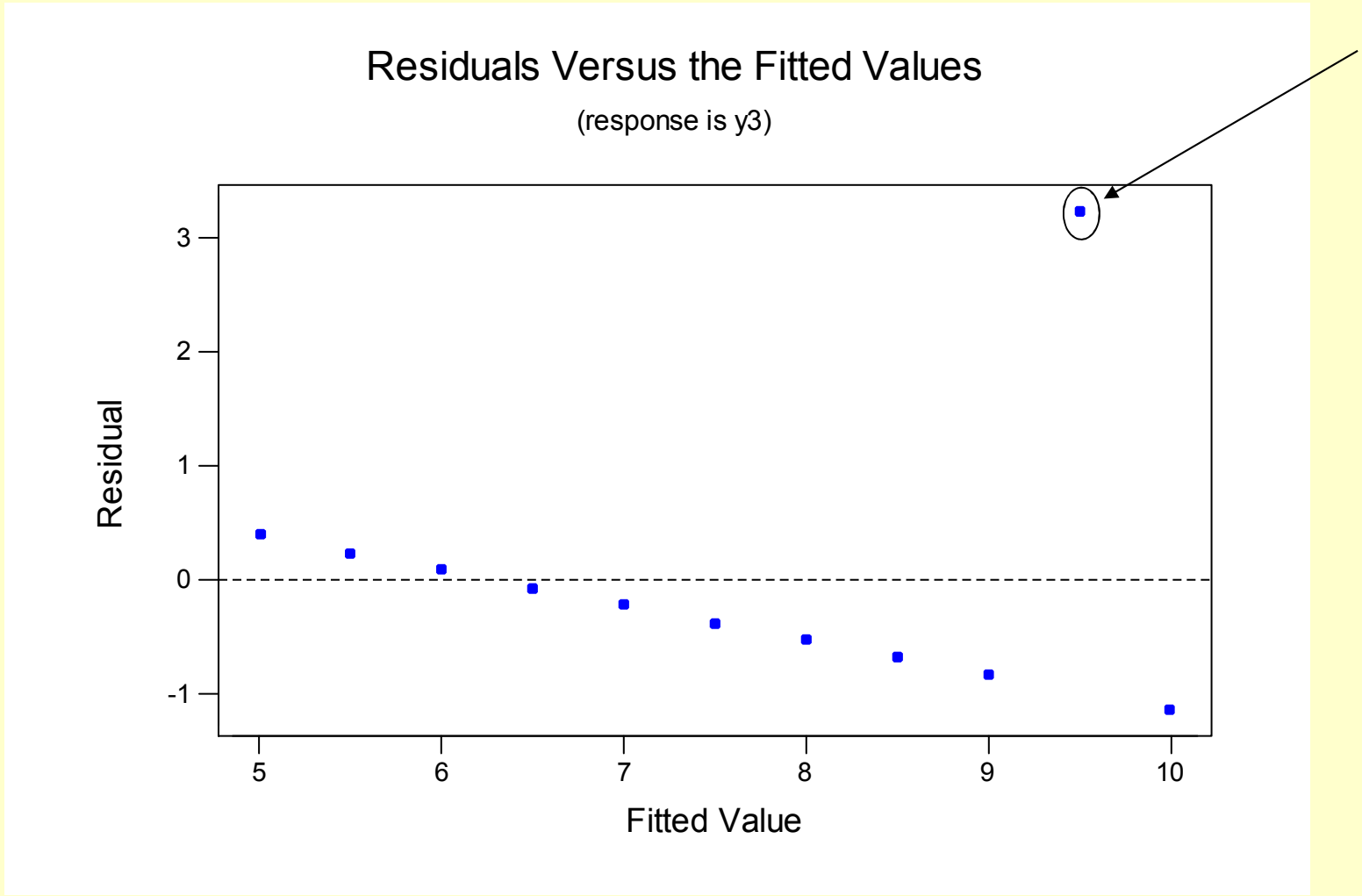
```
Unusual Observations
Obs  Tobacco  Alcohol  Fit  SE Fit  Resid  St Resid
11      4.56    4.020  5.728   0.482 -1.708   -2.58R

R denotes an observation with a large standardized
residual.
```

Anscombe data set #3



A residual vs. fits plot suggesting an outlier exists

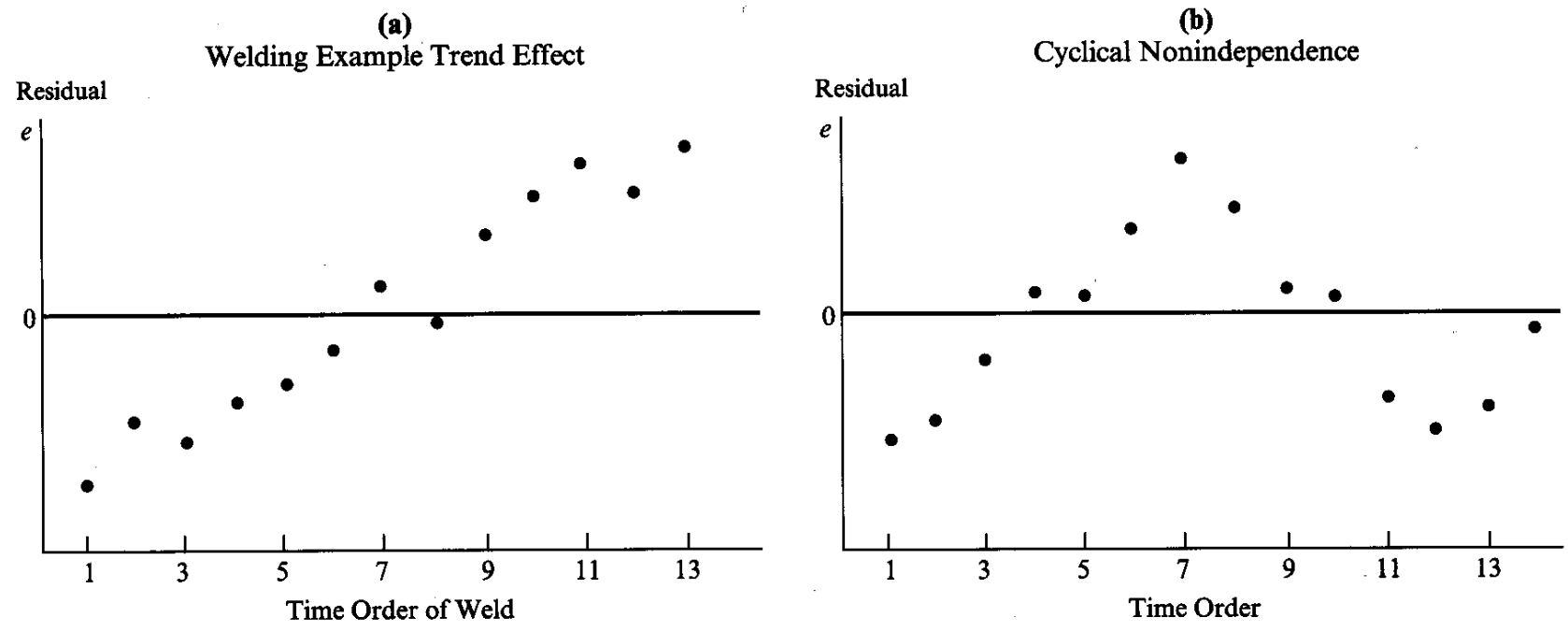


Residuals vs. order plot

- Helps assess **serial correlation** of error terms.
- If the data are obtained in a time (or space) sequence, a "residuals vs. order" plot helps to see if there is any correlation between error terms that are near each other in the sequence.
- A horizontal band bouncing randomly around 0 suggests errors are independent, while a systematic pattern suggests not.

Residuals vs. order plots suggesting non-independence of error terms

FIGURE 3.8 Residual Time Sequence Plots Illustrating Nonindependence of Error Terms.



Normal (probability) plot of residuals

- Helps assess normality of error terms.
- If data are Normal(μ, σ^2), then **percentiles of the normal distribution** should plot linearly against **sample percentiles** (with sampling variation).
- The parameters μ and σ^2 are unknown. Theory shows it's okay to assume $\mu = 0$ and $\sigma^2 = 1$.

Normal (probability) plot
of residuals

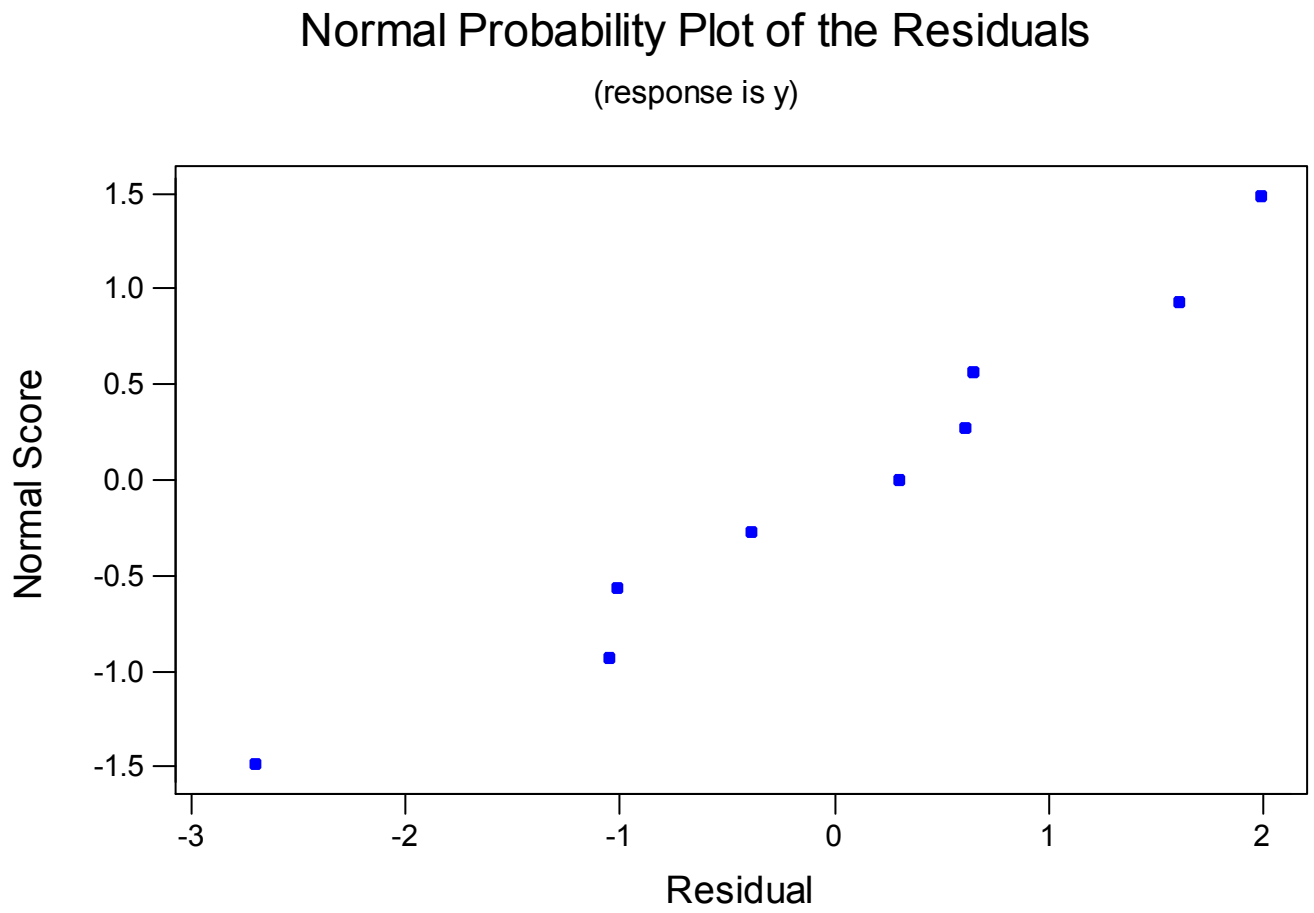
Ordered! $\left(\frac{i}{n+1}\right)$ $\left(\frac{i-3/8}{n+1/4}\right)$

x	y	i	RESI1	PCT	MTB_PCT	NSCORE
3	12	1	-2.70103	0.1	0.060976	-1.54664
2	10	2	-1.04639	0.2	0.158537	-1.00049
5	21	3	-1.01031	0.3	0.256098	-0.65542
1	7	4	-0.39175	0.4	0.353659	-0.37546
3	15	5	0.29897	0.5	0.451220	-0.12258
1	8	6	0.60825	0.6	0.548780	0.12258
4	19	7	0.64433	0.7	0.646341	0.37546
1	9	8	1.60825	0.8	0.743902	0.65542
5	24	9	1.98969	0.9	0.841463	1.00049

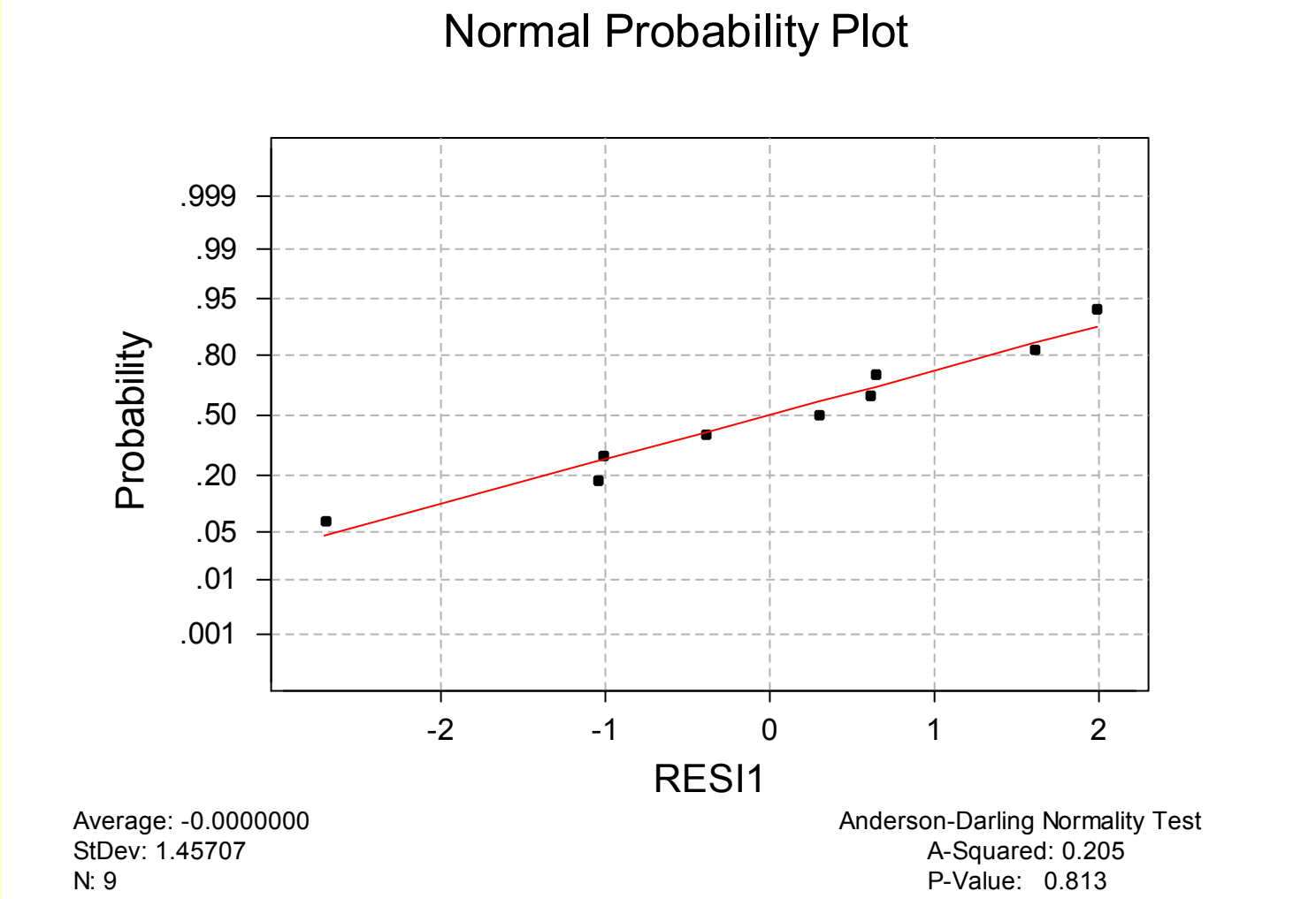
Normal (probability) plot
of residuals (cont'd)

- Plot **normal scores** (theoretical percentiles) on vertical axis against **ordered residuals** (sample percentiles) on horizontal axis.
- Plot that is nearly linear suggests normality of error terms.

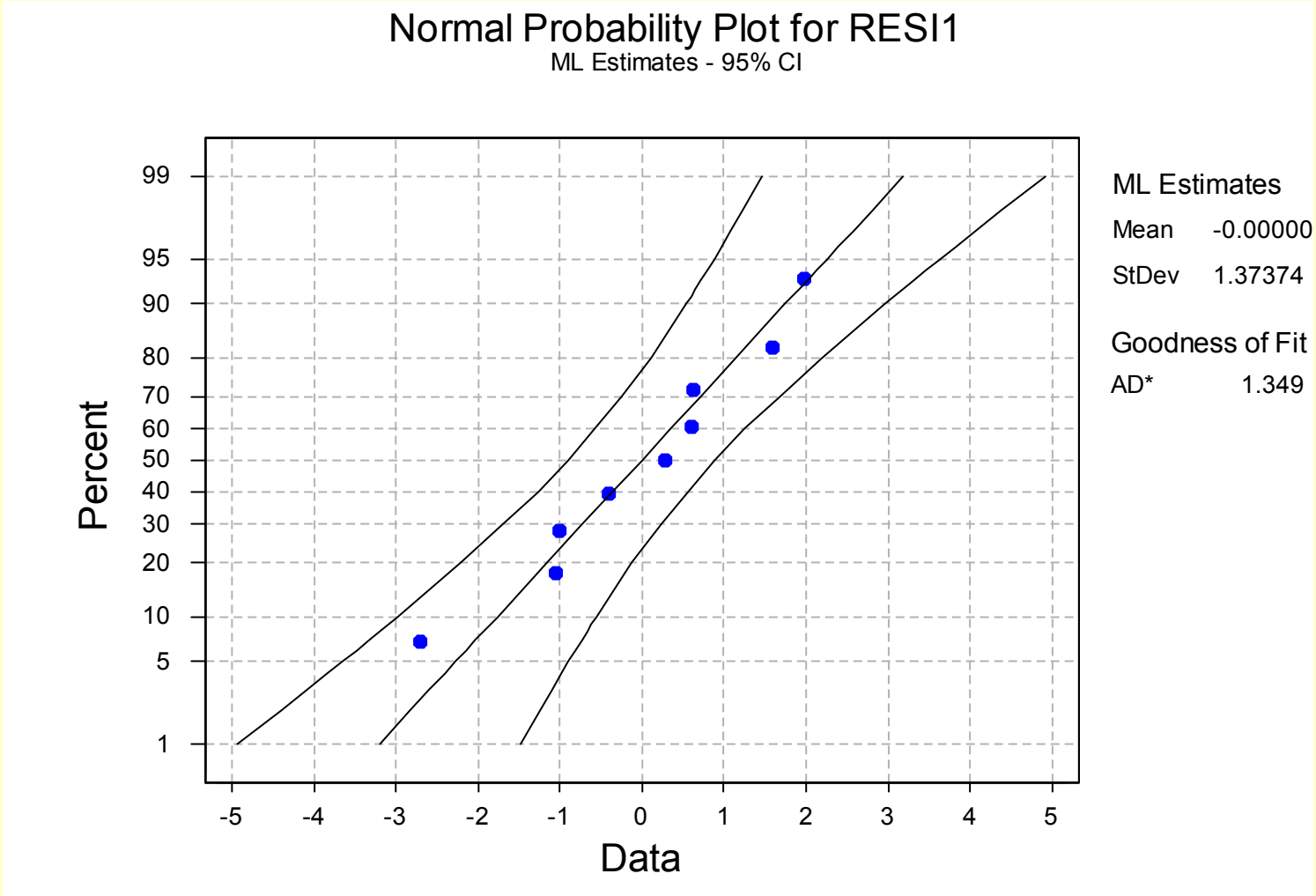
Normal (probability) plot



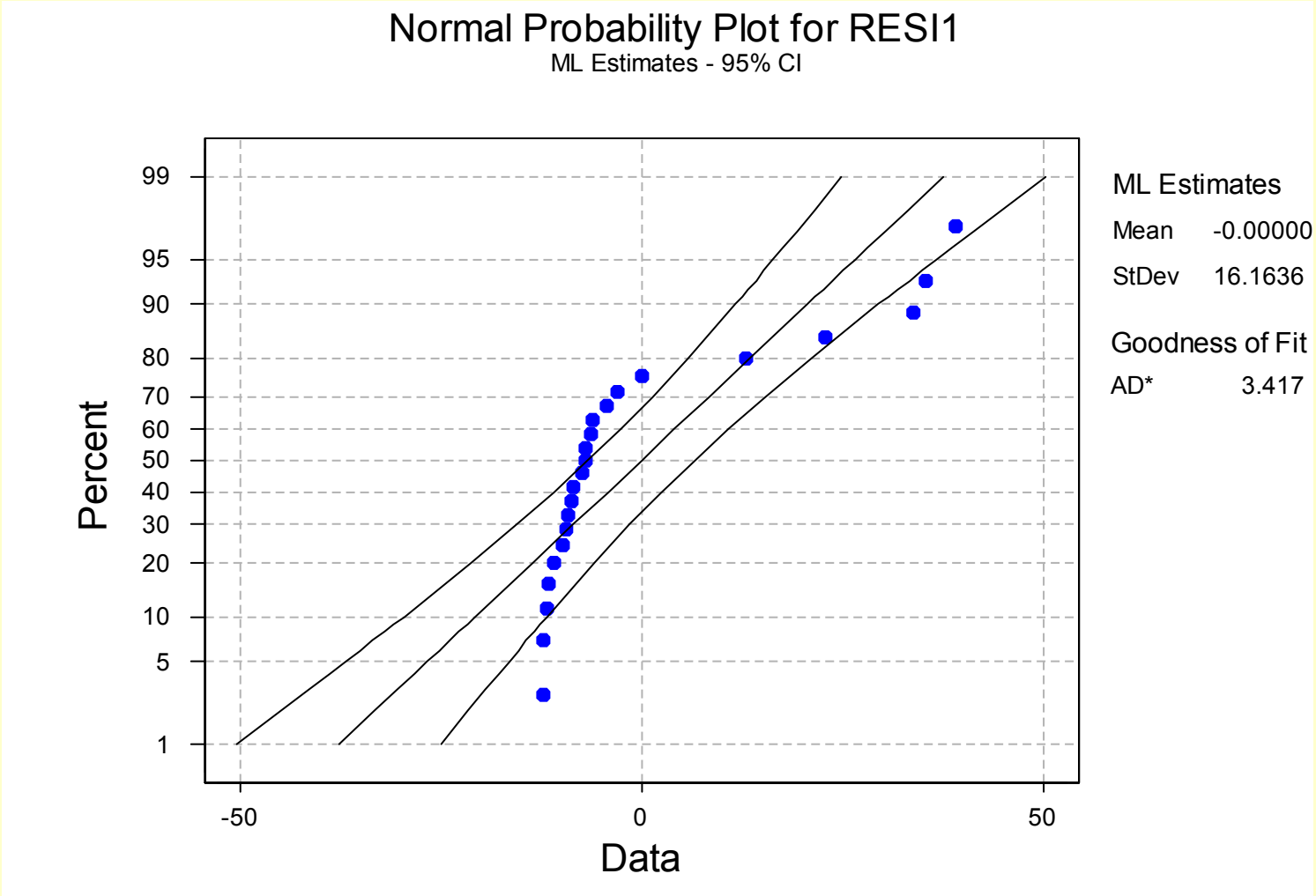
Normal (probability) plot



Normal (probability) plot



A normal (probability) plot with non-normal error terms



Fixing problems with the model

Transforming the data so that the simple linear regression model is okay for the transformed data.

Options for fixing problems with the model

- Abandon simple linear regression model and find a more appropriate – but typically more complex – model.
- **Transform the data** so that the simple linear regression model works for the transformed data.

Abandoning the model

- If **not linear**: try a different function, like a quadratic (Ch. 7) or an exponential function (Ch. 13).
- If **unequal error variances**: use weighted least squares (Ch. 10).
- If **error terms are not independent**: try fitting a time series model (Ch. 12).
- If **important predictor variables omitted**: try fitting a multiple regression model (Ch. 6).
- If **outlier**: use robust estimation procedure (Ch. 10).

Choices for transforming the data

- Transform *X* values only.
- Transform *Y* values only.
- Transform both the *X* and the *Y* values.

Transforming the *X* values only

Transforming the X values only

- Appropriate **when non-linearity is the only problem** – normality and equal variance okay – with the model.
- Transforming the *Y* values would likely change the well-behaved error terms into badly-behaved error terms.

Example 1

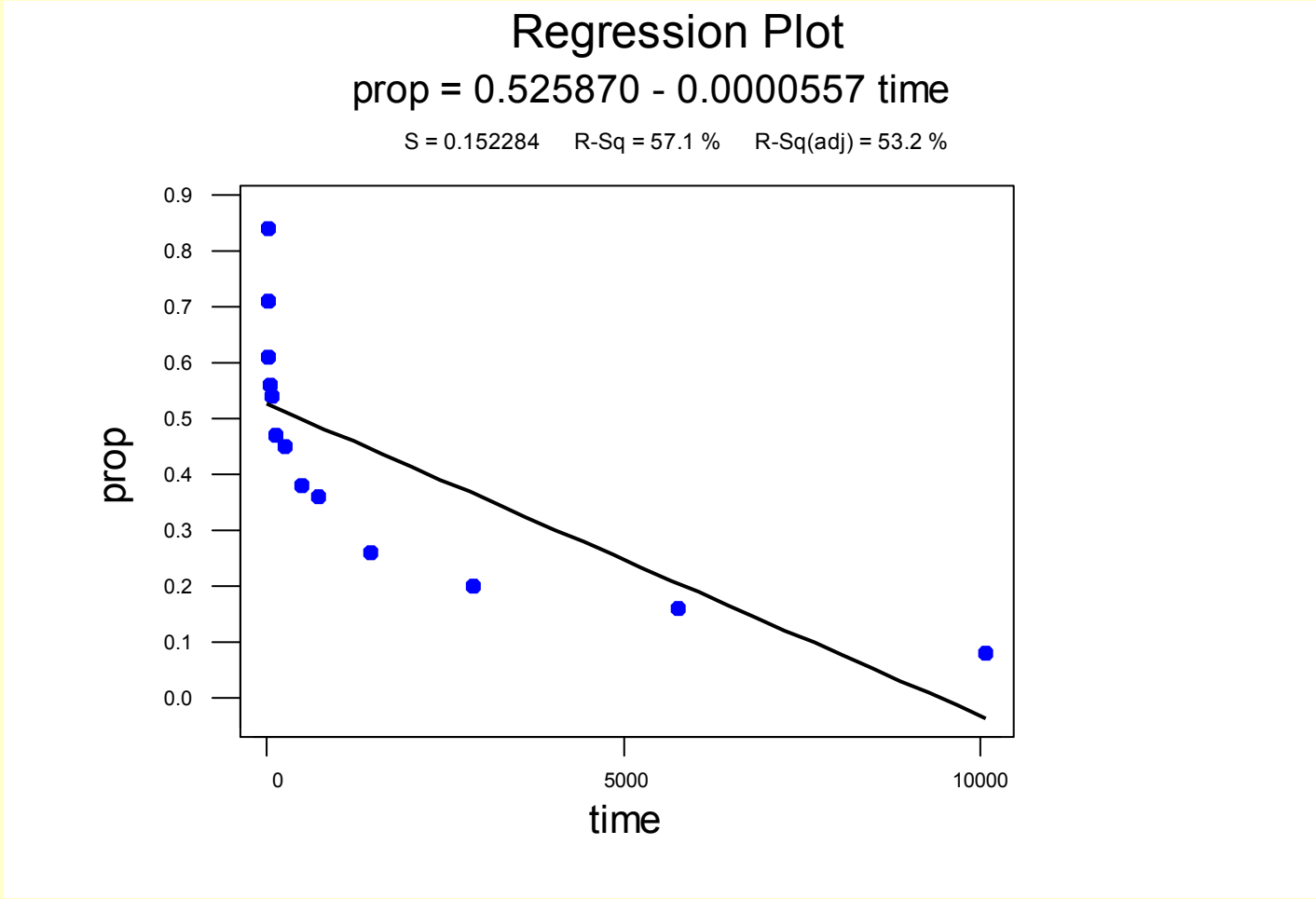
Memory retention

time	prop
1	0.84
5	0.71
15	0.61
30	0.56
60	0.54
120	0.47
240	0.45
480	0.38
720	0.36
1440	0.26
2880	0.20
5760	0.16
10080	0.08

- Subjects asked to memorize a list of disconnected items. Asked to recall them at various times up to a week later
- Predictor **time** = time, in minutes, since initially memorized the list.
- Response **prop** = proportion of items recalled correctly.

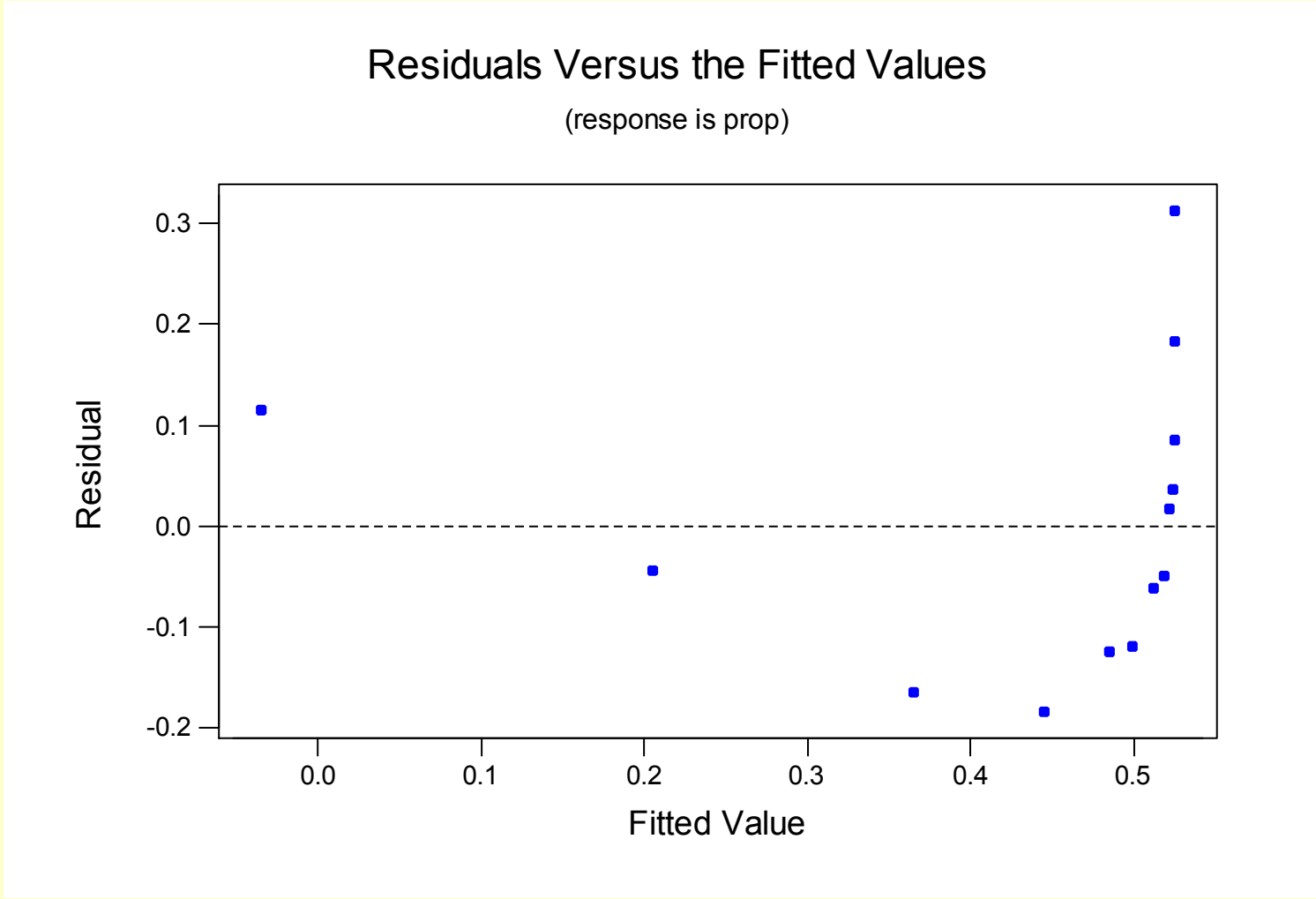
Example 1

Fitted line plot



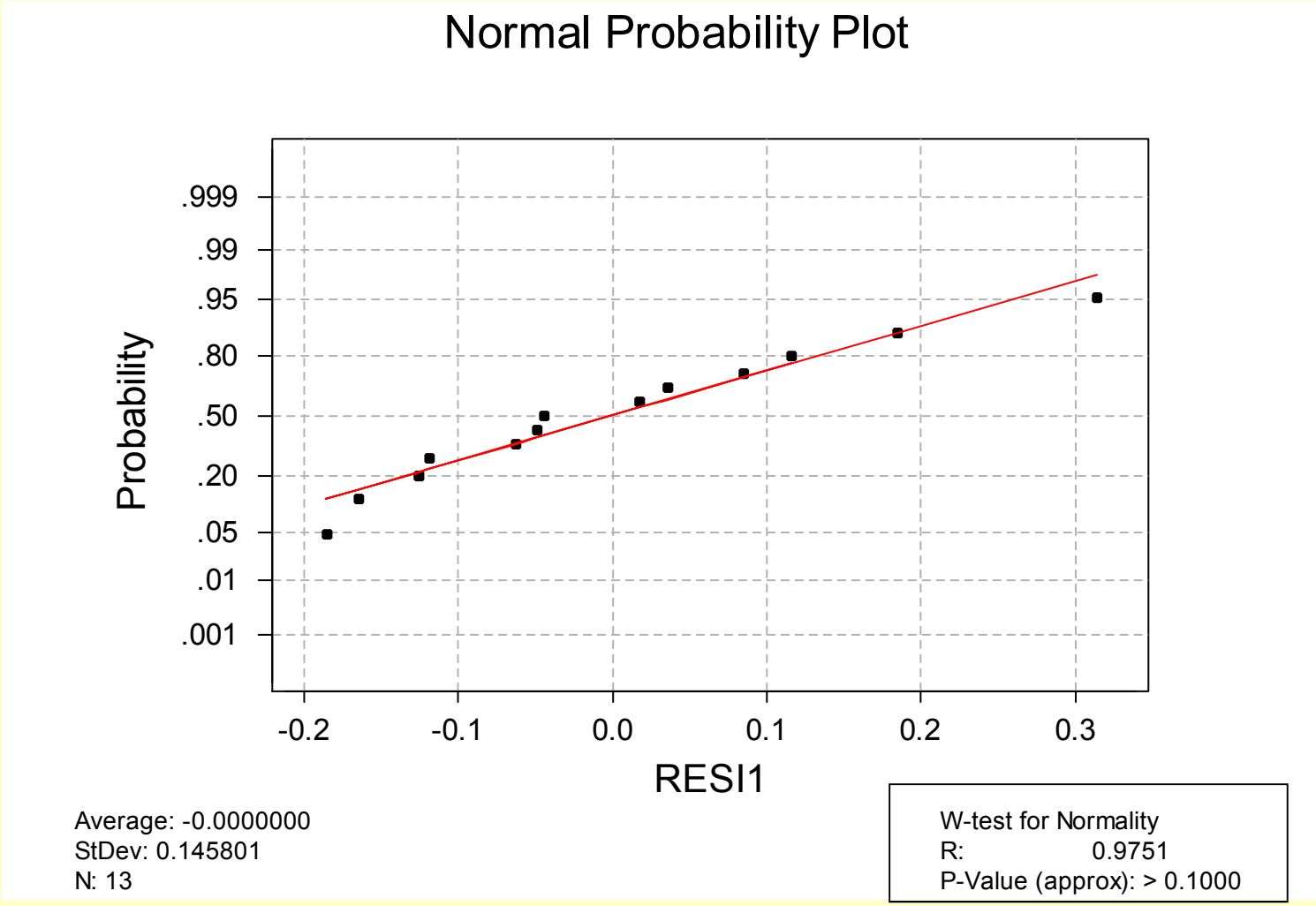
Example 1

Residual vs. fits plot



Example 1

Normal probability plot



Example 1

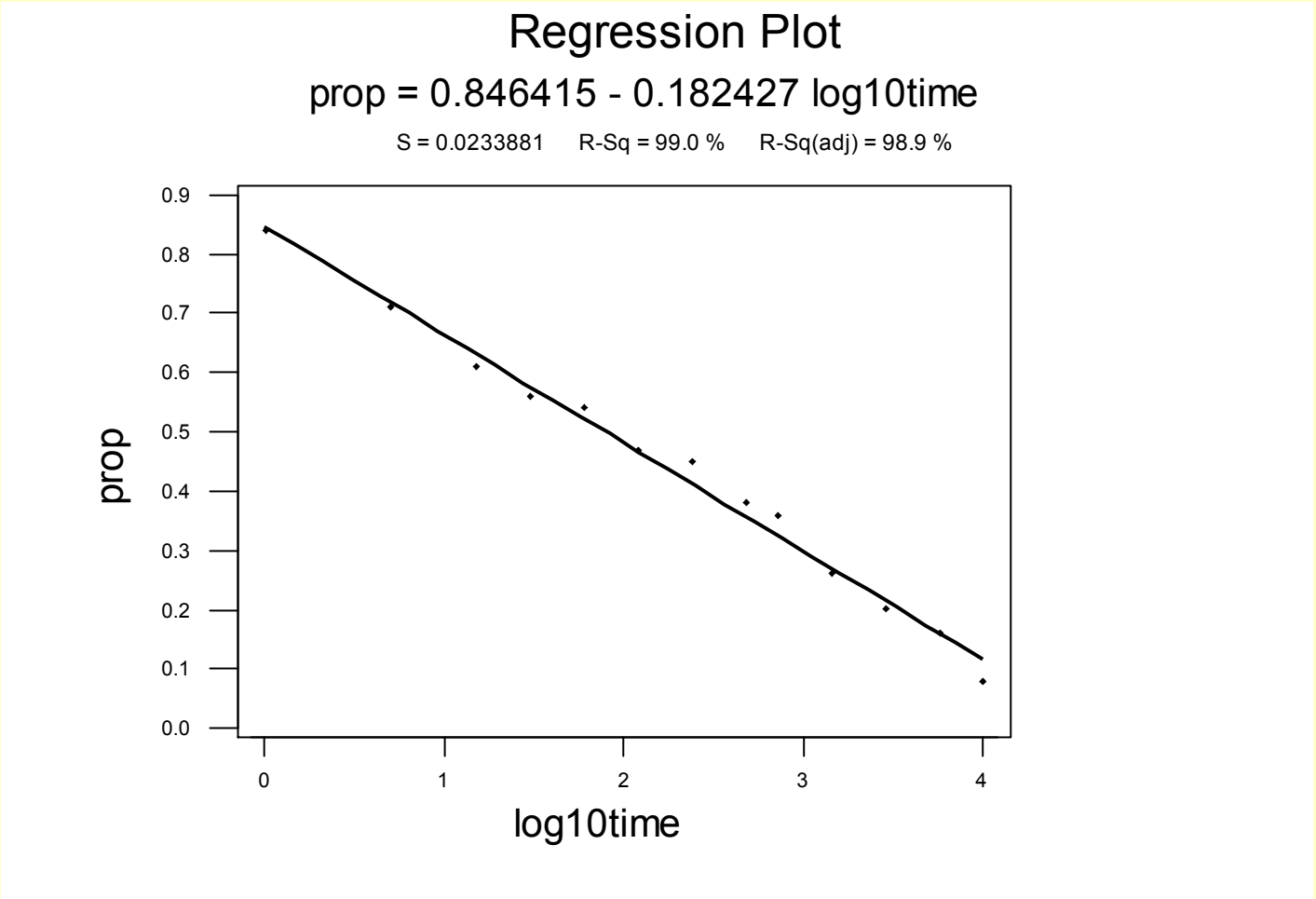
Transform the *X* values

time	prop	log10_time
1	0.84	0.00000
5	0.71	0.69897
15	0.61	1.17609
30	0.56	1.47712
60	0.54	1.77815
120	0.47	2.07918
240	0.45	2.38021
480	0.38	2.68124
720	0.36	2.85733
1440	0.26	3.15836
2880	0.20	3.45939
5760	0.16	3.76042
10080	0.08	4.00346

Change ("transform") the predictor **time** to **$\log_{10}$(time)**.

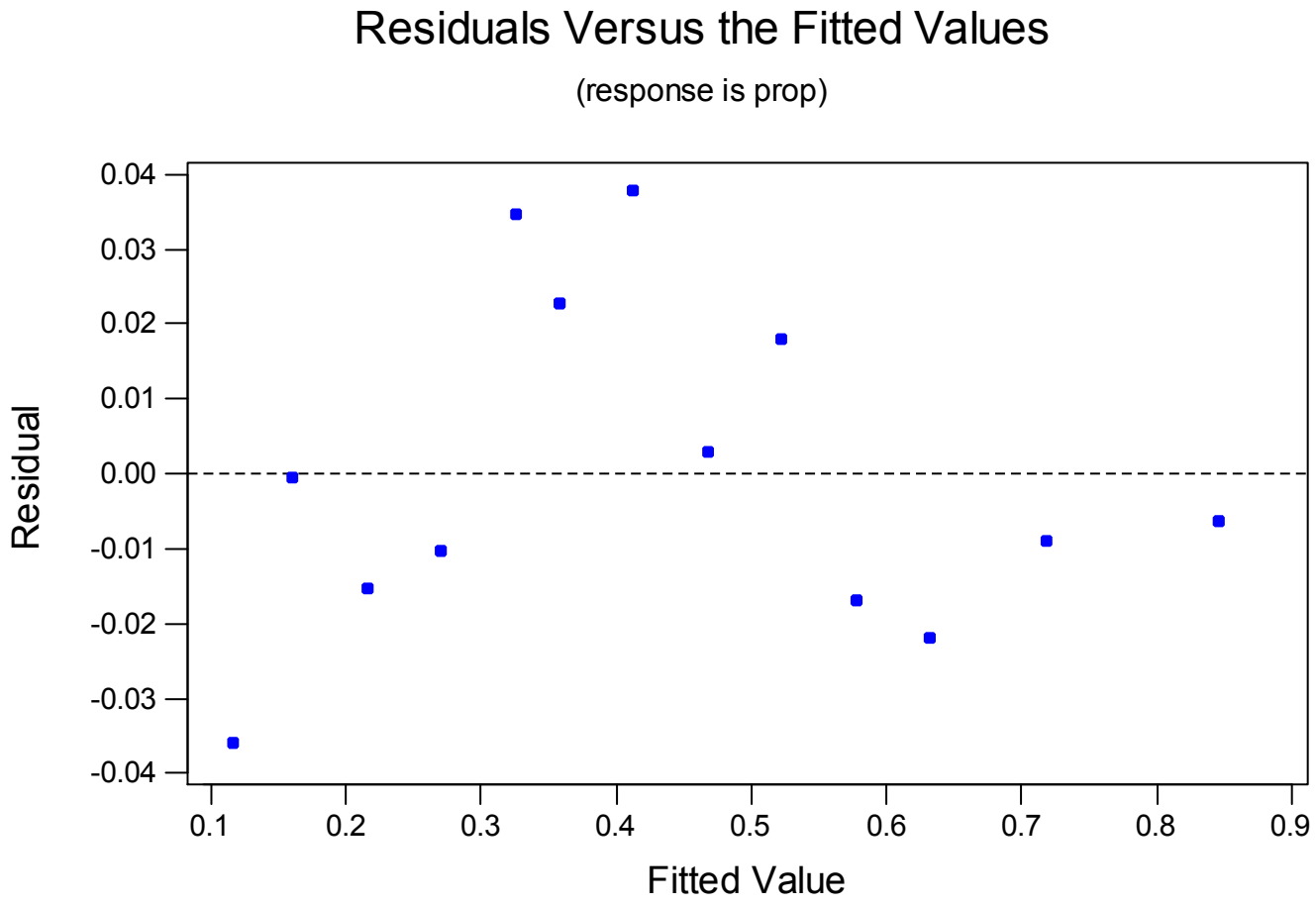
Example 1

Fitted line plot using transformed *X* values



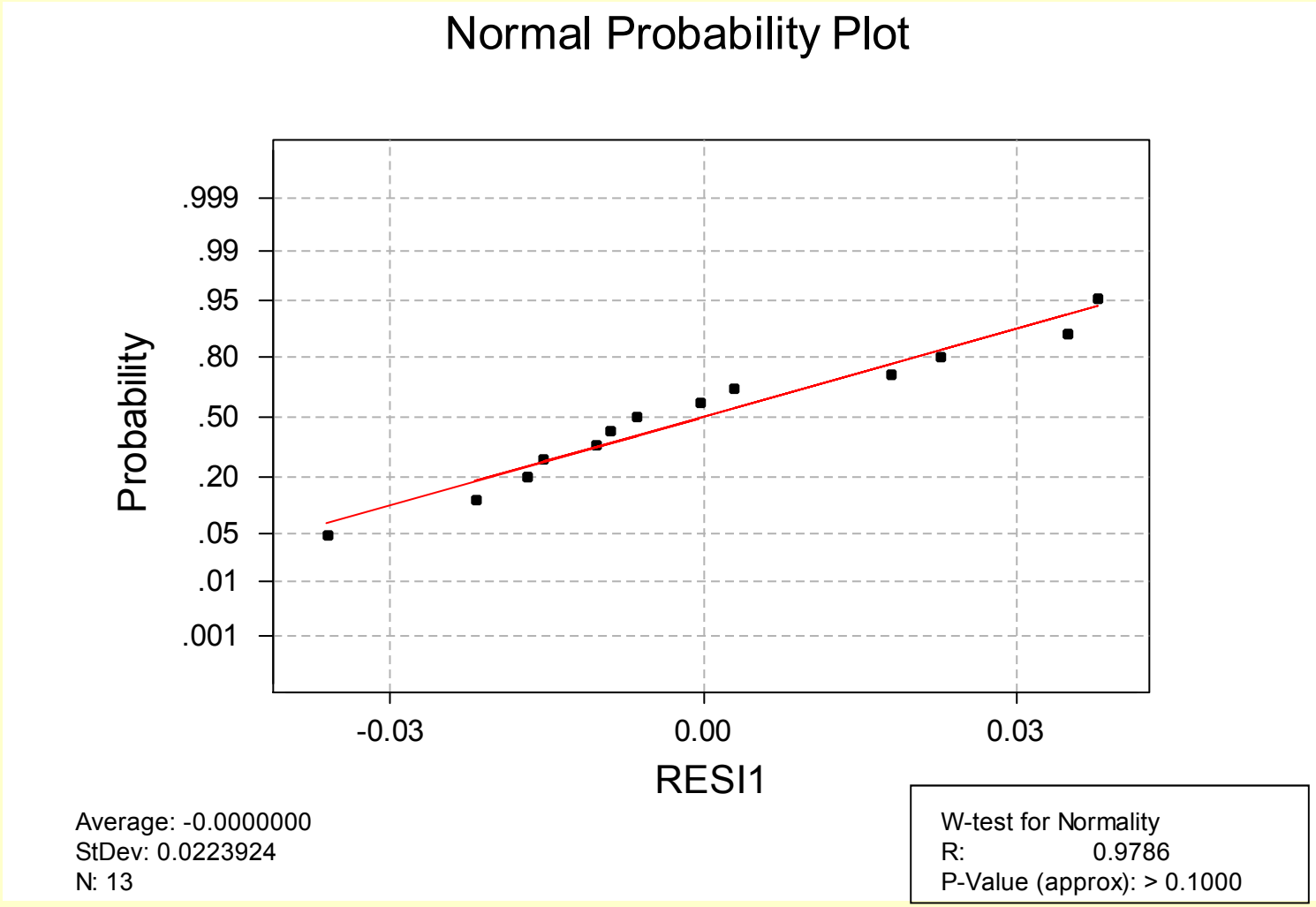
Example 1

Residuals vs. fits plot using transformed *X* values



Example 1

Normal probability plot using transformed *X* values



Example 1

Predicting new proportion

Estimated regression function:

$$\hat{\boldsymbol{Y}} = 0.846 - 0.182 \times \log_{10}(\boldsymbol{time})$$

Therefore, we predict the proportion of words recalled after 1000 minutes is:

$$\hat{\boldsymbol{Y}} = 0.846 - 0.182 \times \log_{10}(1000)$$

$$\hat{\boldsymbol{Y}} = 0.846 - 0.182 \times 3 = 0.30$$

Example 1

Predicting new proportion

```
Predicted Values for New Observations

New     Fit     SE Fit         95.0% CI         95.0% PI
1     0.299    0.00765    (0.282, 0.316)   (0.245, 0.353)

Values of Predictors for New Observations

New Obs  log10tim
1              3.00
```

We can be 95% confident that a person will recall between 24.5% and 35.3% of the words after 1000 minutes.

Transforming the *Y* values only

Transforming the *Y* values only

- Appropriate when **non-normality** and/or **unequal variances** are the problems.
- The transformation on *Y* may also help to "straighten out" a curved relationship.

Example 2

Gestation time and birth weight for mammals

Mammal	Birthwgt	Gestation
Goat	2.75	155
Sheep	4.00	175
Deer	0.48	190
Porcupine	1.50	210
Bear	0.37	213
Hippo	50.00	243
Horse	30.00	340
Camel	40.00	380
Zebra	40.00	390
Giraffe	98.00	457
Elephant	113.00	670

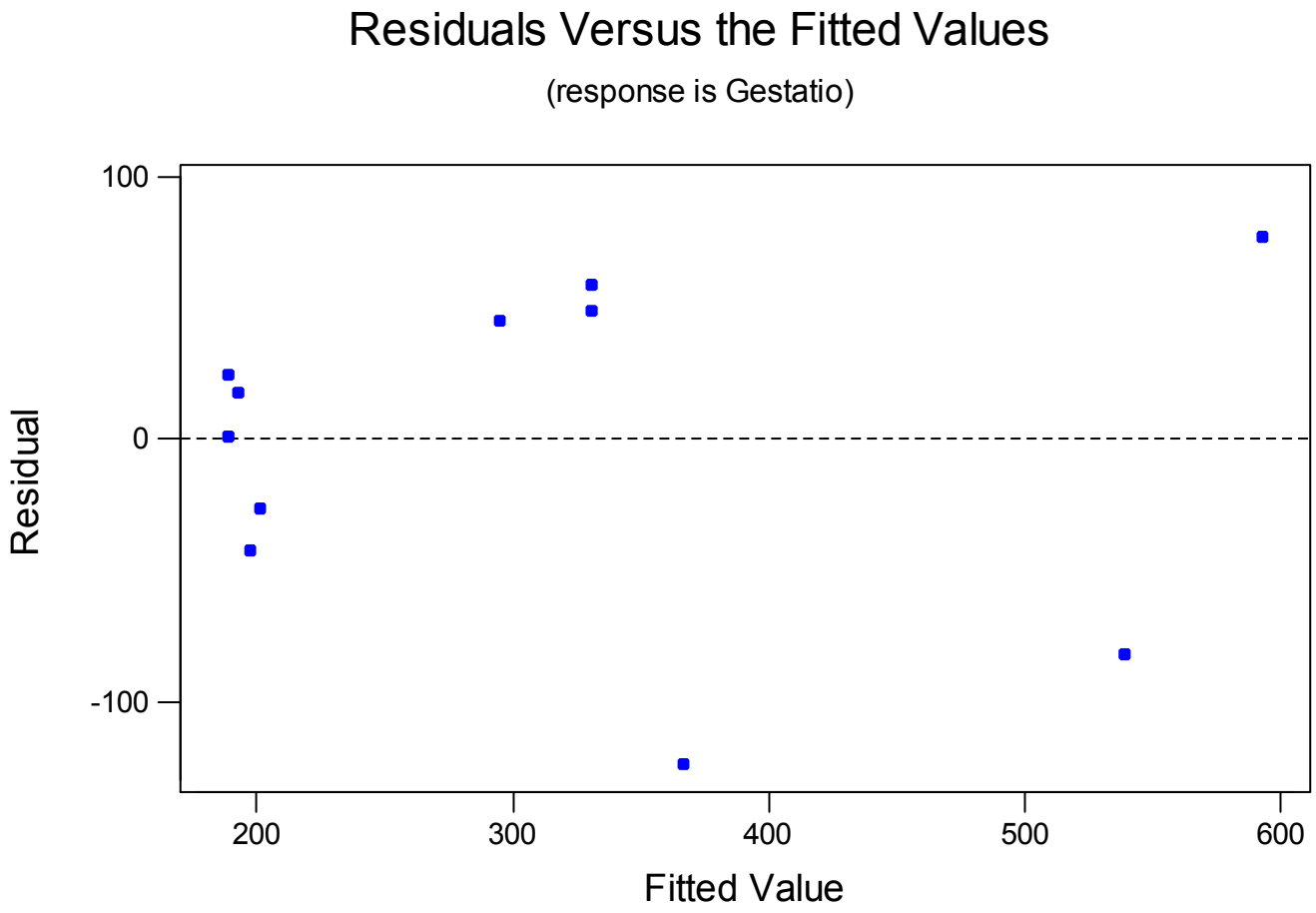
- Predictor **Birthwgt** = birth weight, in kg, of mammal.
- Response **Gestation** = number of days until birth

Example 2

Fitted line plot

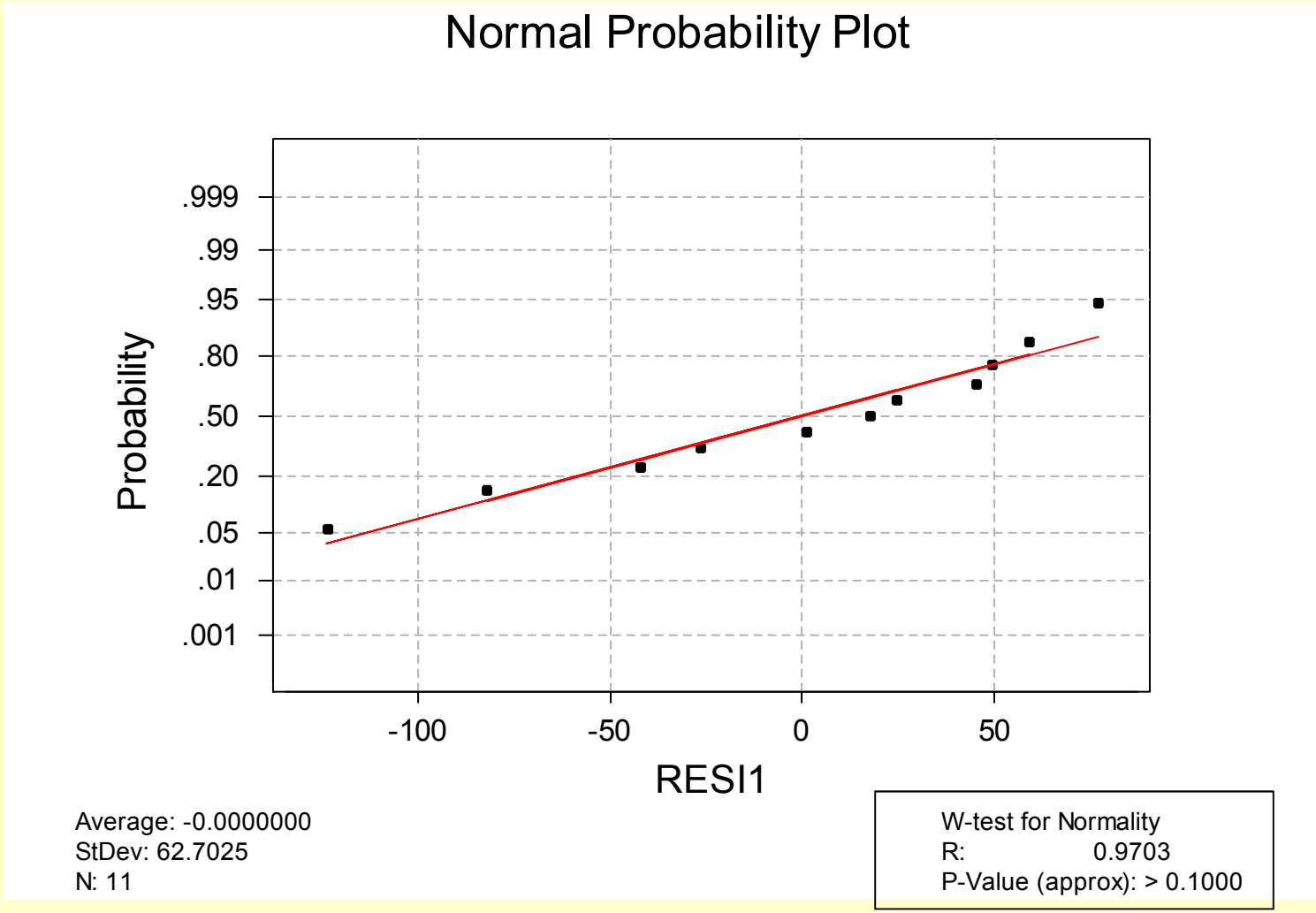
Example 2

Residual vs. fits plot



Example 2

Normal probability plot



Example 2

Transform the *Y* values

Mammal	Birthwgt	Gestation	log10Gest
Goat	2.75	155	2.19033
Sheep	4.00	175	2.24304
Deer	0.48	190	2.27875
Porcupine	1.50	210	2.32222
Bear	0.37	213	2.32838
Hippo	50.00	243	2.38561
Horse	30.00	340	2.53148
Camel	40.00	380	2.57978
Zebra	40.00	390	2.59106
Giraffe	98.00	457	2.65992
Elephant	113.00	670	2.82607

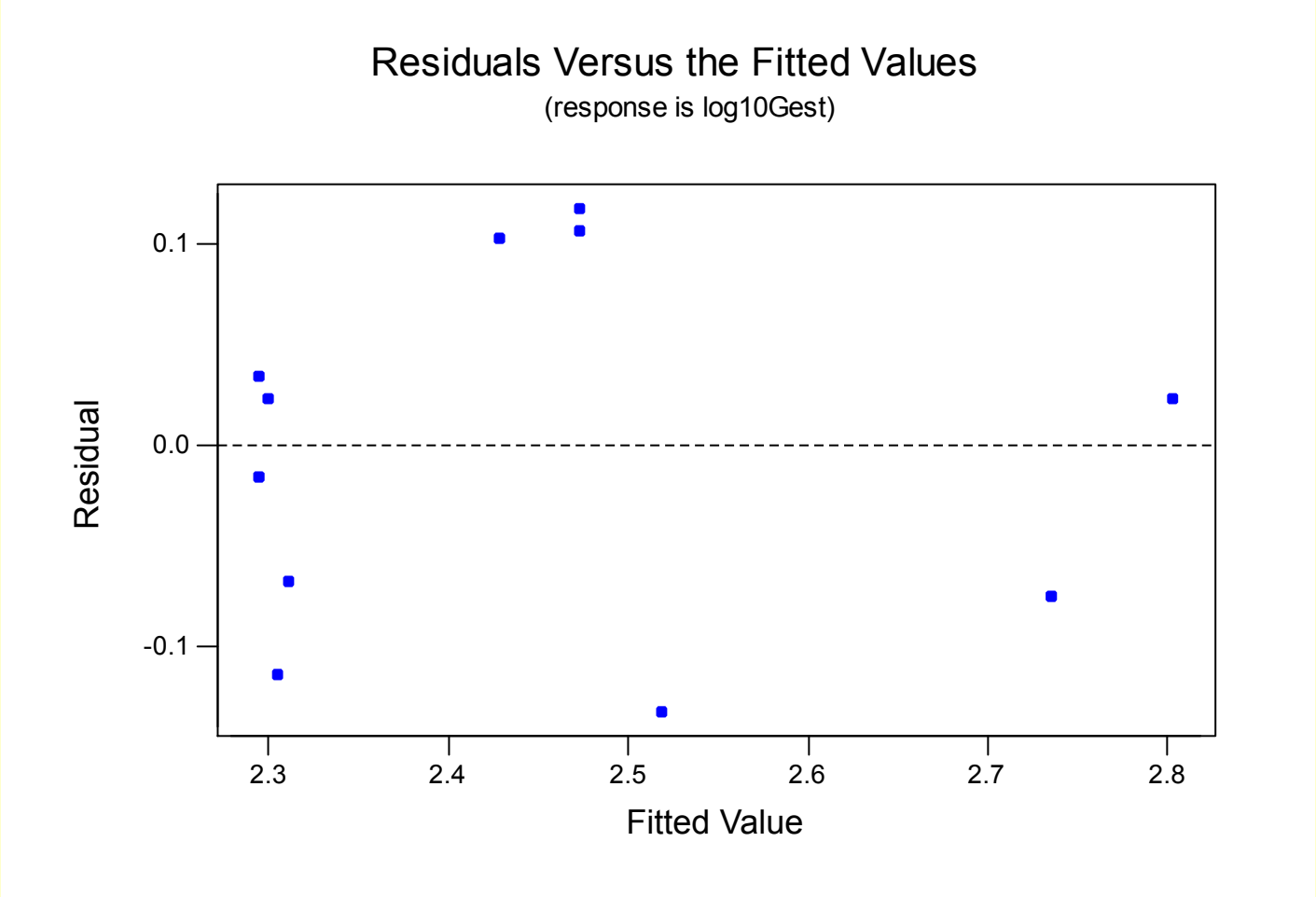
Change ("transform") the response **Gestation** to **$\log_{10}$(Gestation)**.

Example 2

Fitted line plot using transformed *Y* values

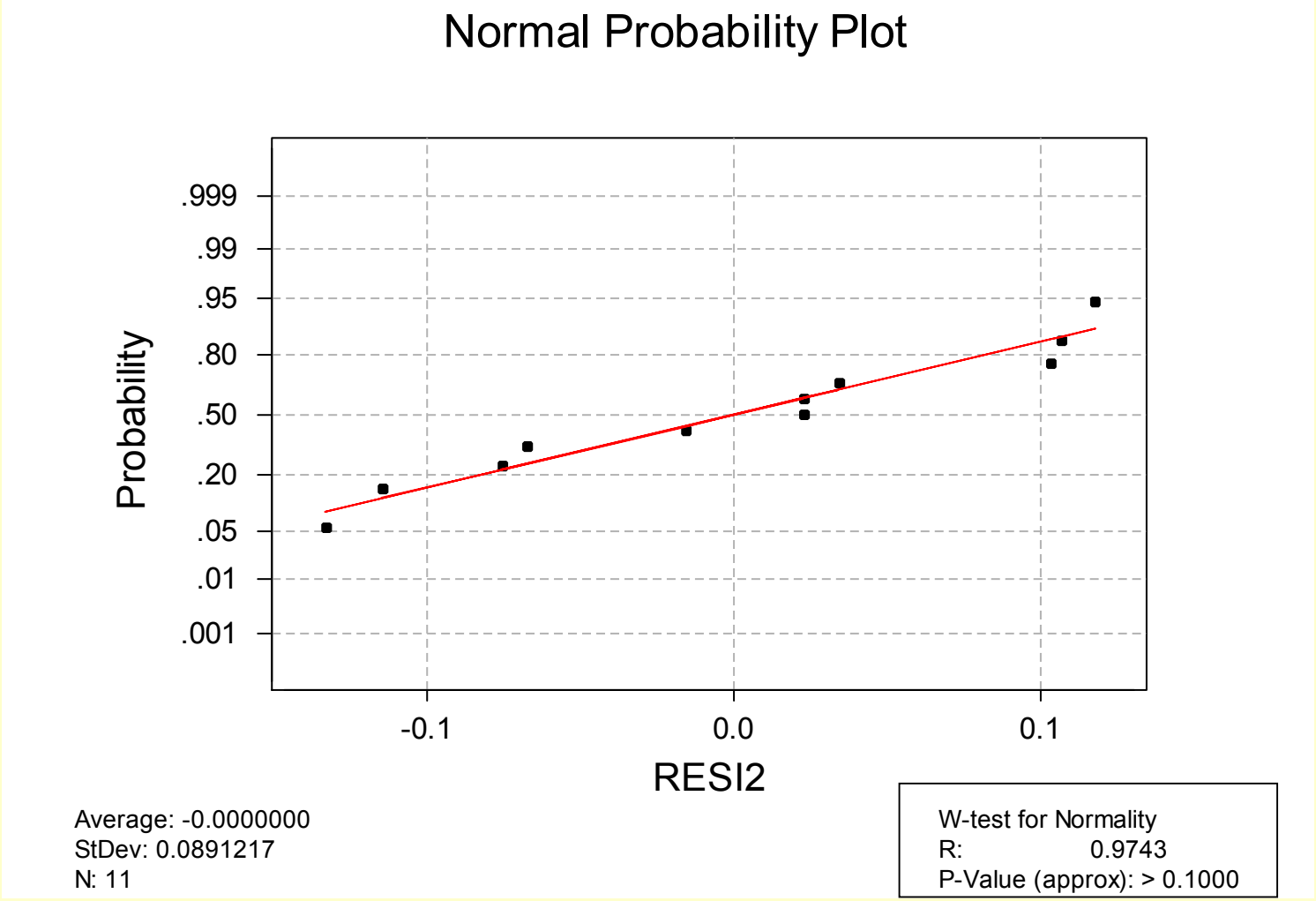
Example 2

Residual vs. fits plot using transformed *Y* values



Example 2

Normal probability plot using transformed *Y* values



Example 2

Predicting new gestation

Estimated regression function:

$$\log_{10}(\hat{\boldsymbol{Gest}}) = 2.29 + 0.0045 \times \boldsymbol{Birthwgt}$$

Therefore, since:

$$\log_{10}(\hat{\boldsymbol{Gest}}) = 2.29 + 0.0045 \times 50 = 2.515$$

we predict the gestation length of another mammal at 50 kgs to be:

$$\hat{\boldsymbol{Gest}} = 10^{\log_{10}(\hat{\boldsymbol{Gest}})} = 10^{2.515} = 327.3$$

Example 2

Predicting new gestation

```
Predicted Values for New Observations

New   Fit     SE Fit        95.0% CI            95.0% PI
1    2.5186   0.0306   (2.4494, 2.5878)   (2.2951, 2.7421)

Values of Predictors for New Observations

New    Birthwgt
1         50.0
```

$10^{2.2951} = 197.3$

$10^{2.7421} = 552.2$

We can be 95% confident that the gestation length for a new mammal at 50 kgs will be between 197.3 and 552.2 days.

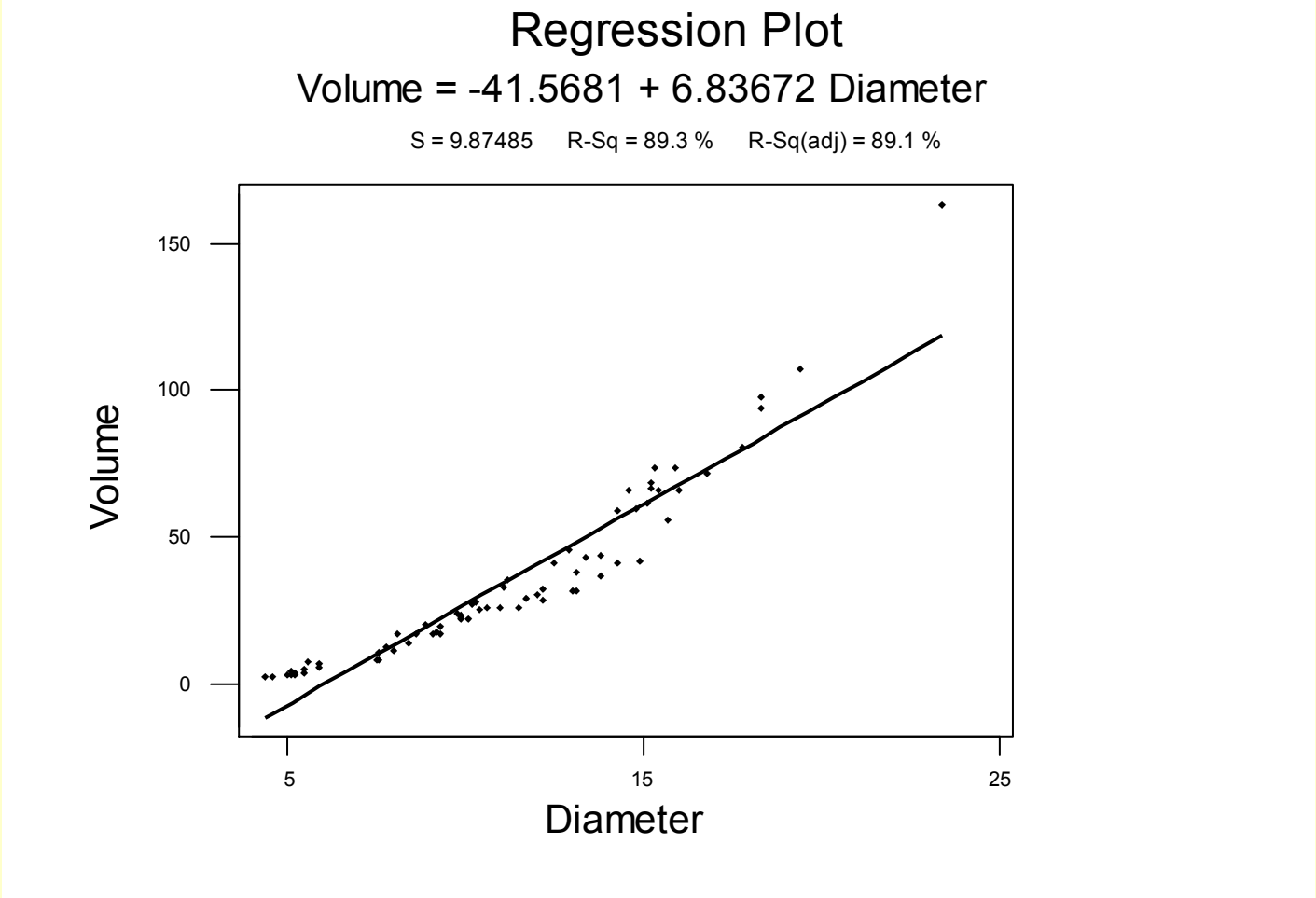
Transforming both the X and Y values

Transforming both the *X* and *Y* values

- Appropriate when the error terms are **not normal**, have **unequal variances**, and the function is **not linear**.
- Transforming the *Y* values corrects the problems with the error terms (and may help the non-linearity).
- Transforming the *X* values corrects the non-linearity.

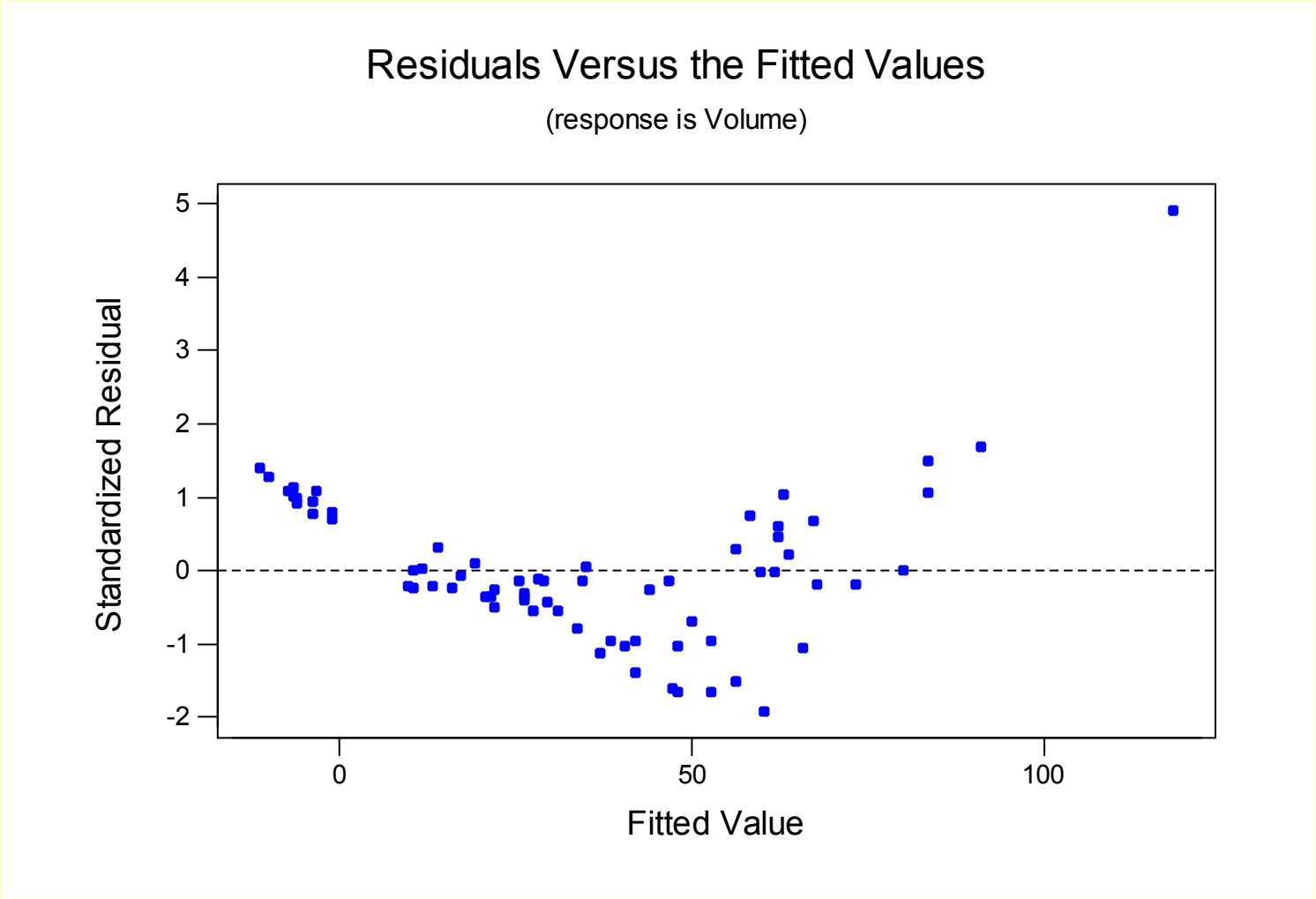
Example 3

Diameter (inches) and volume (cu. ft.) of 70 shortleaf pines



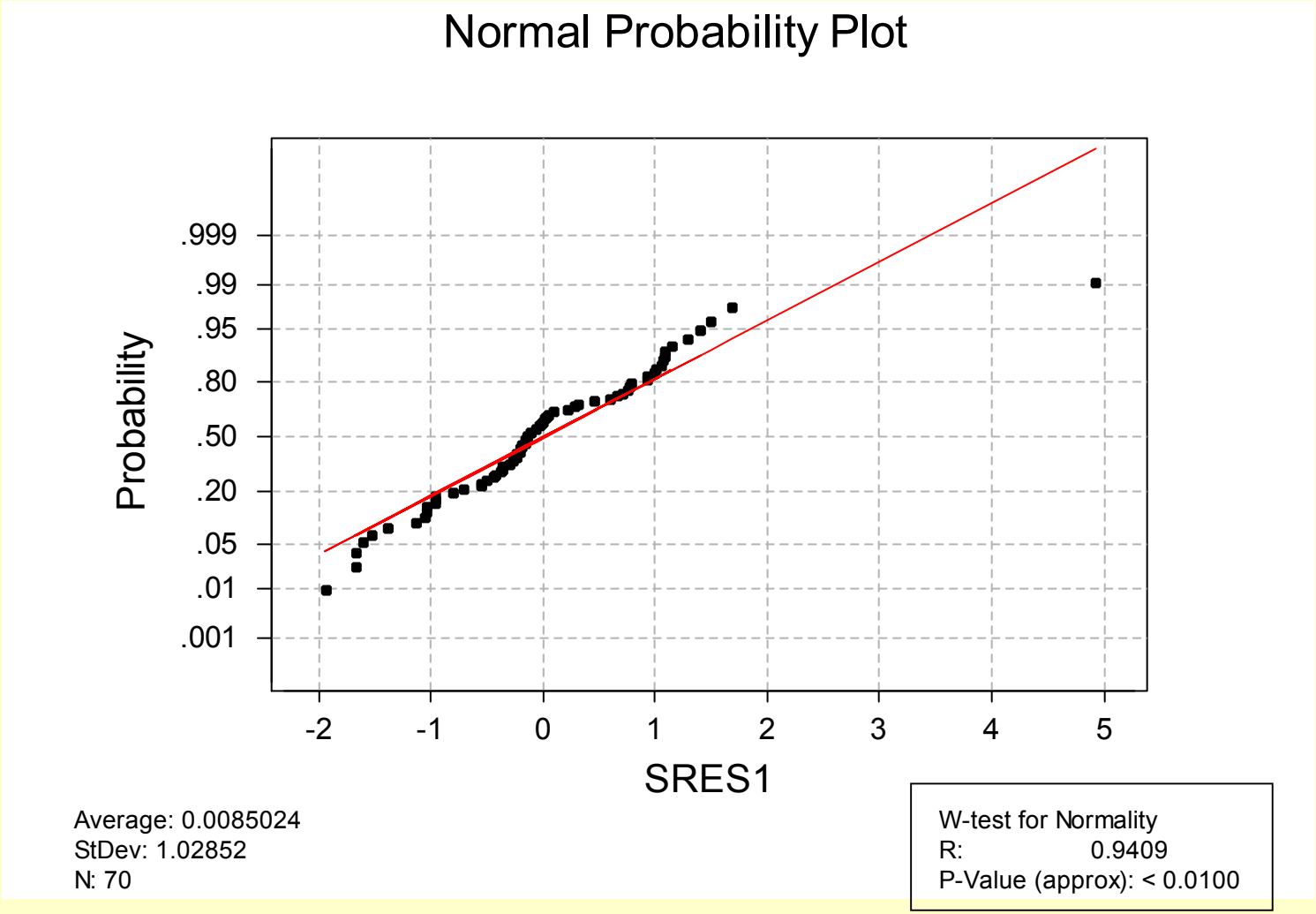
Example 3

Residuals vs. fits plot



Example 3

Normal probability plot



Example 3

Transform the *Y* values only

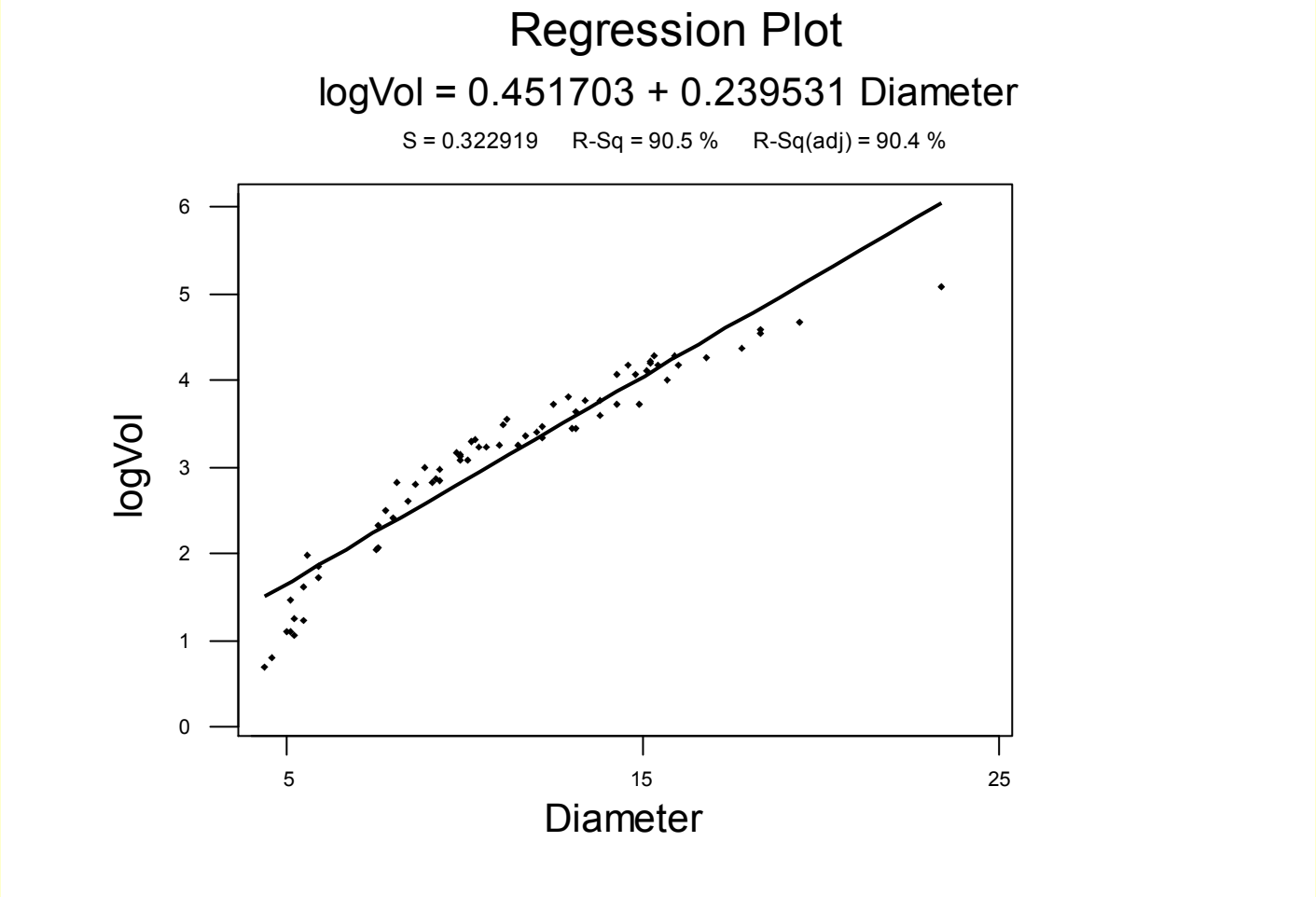
Diameter	Volume	logVol
4.4	2.0	0.69315
4.6	2.2	0.78846
5.0	3.0	1.09861
5.1	4.3	1.45862
5.1	3.0	1.09861
5.2	2.9	1.06471
5.2	3.5	1.25276
5.5	3.4	1.22378
5.5	5.0	1.60944
5.6	7.2	1.97408
5.9	6.4	1.85630
5.9	5.6	1.72277
7.5	7.7	2.04122
7.6	10.3	2.33214

… **and so on** …

Transform response
volume to **$\log_e$(volume)**

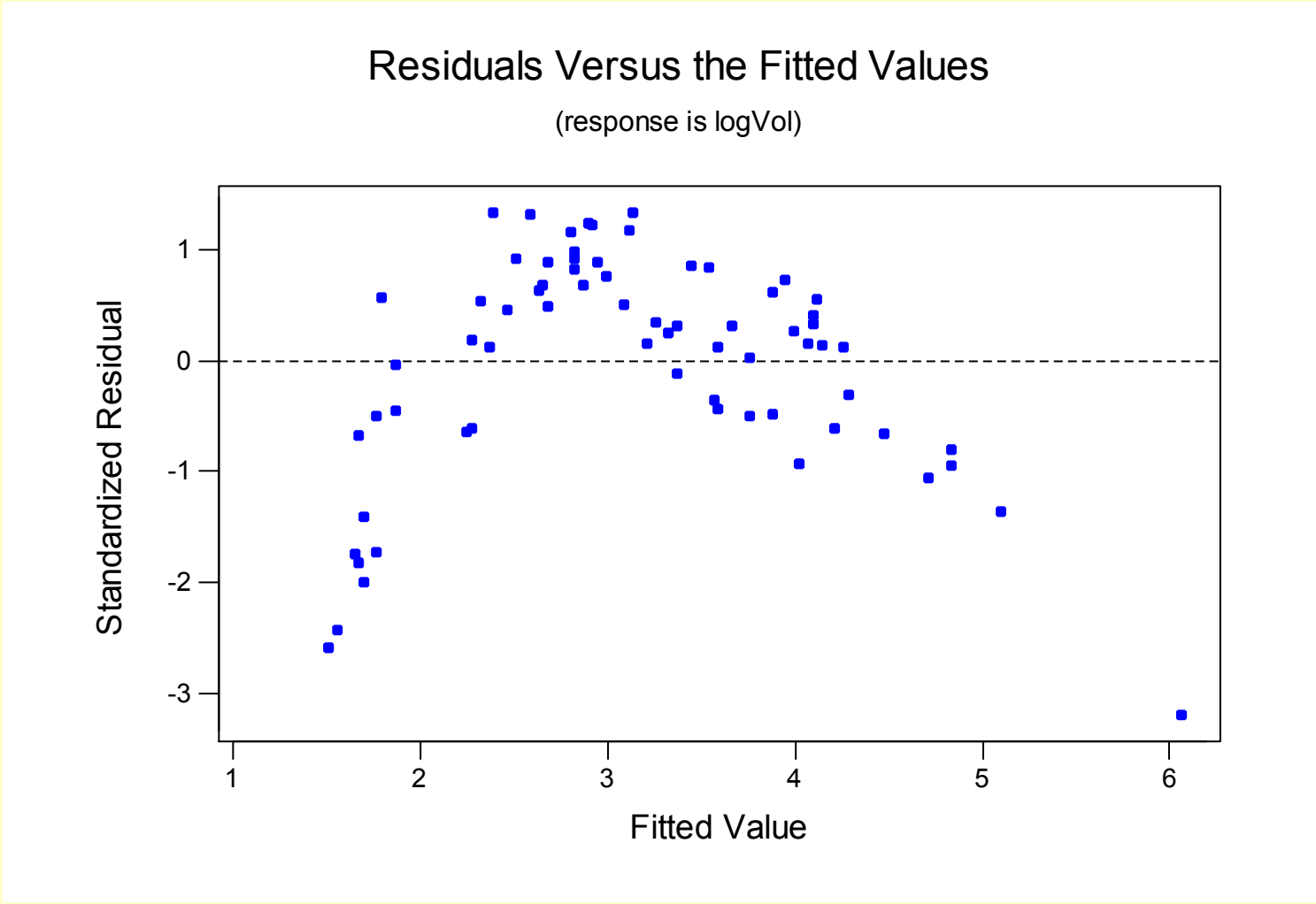
Example 3

Fitted line plot using transformed *Y* values



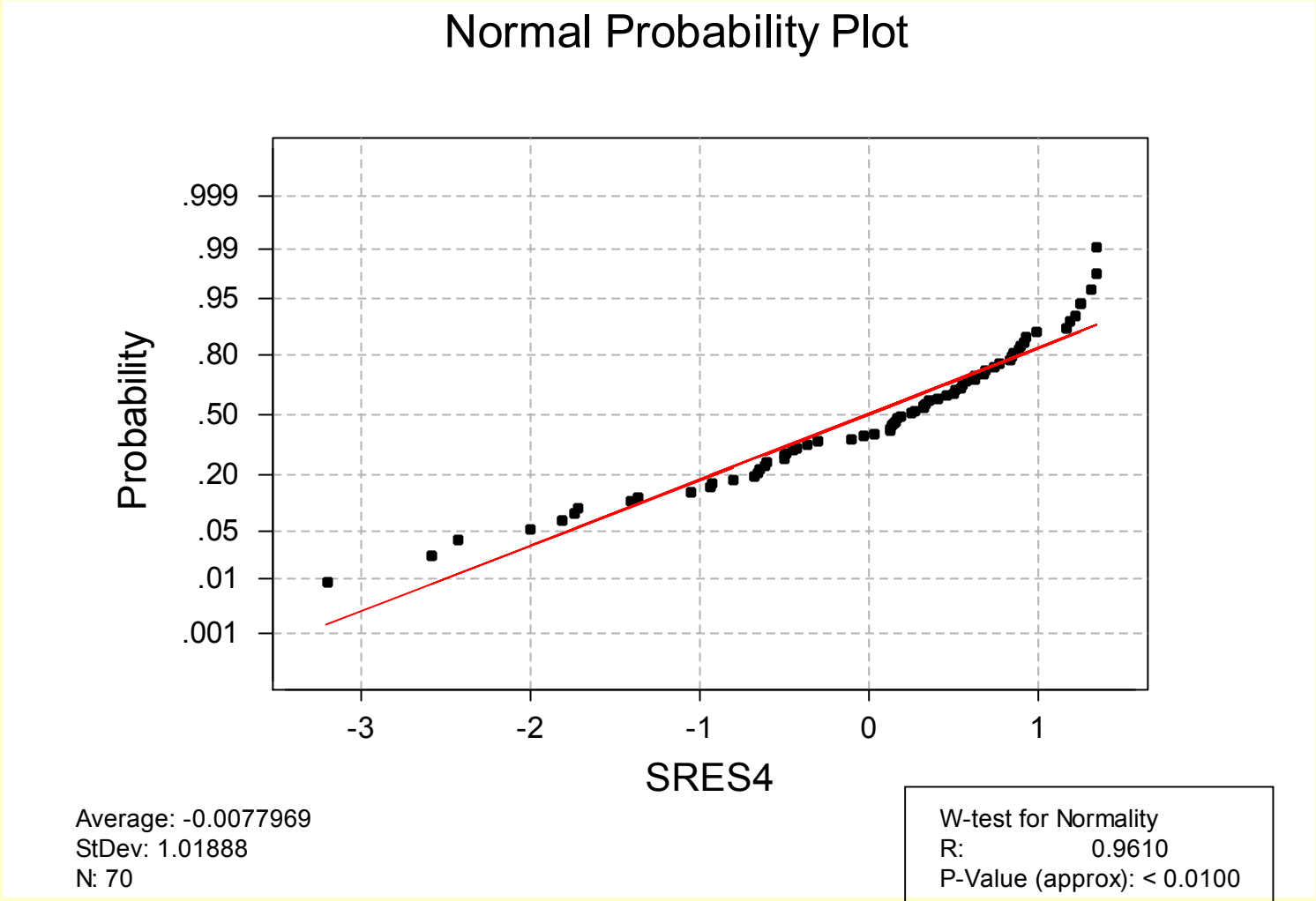
Example 3

Residuals vs. fits plot using transformed *Y* values



Example 3

Normal probability plot using transformed *Y* values



Example 3

Transform both the *X* and *Y* values

Diameter	Volume	logDiam	logVol
4.4	2.0	1.48160	0.69315
4.6	2.2	1.52606	0.78846
5.0	3.0	1.60944	1.09861
5.1	4.3	1.62924	1.45862
5.1	3.0	1.62924	1.09861
5.2	2.9	1.64866	1.06471
5.2	3.5	1.64866	1.25276
5.5	3.4	1.70475	1.22378
5.5	5.0	1.70475	1.60944
5.6	7.2	1.72277	1.97408
5.9	6.4	1.77495	1.85630
5.9	5.6	1.77495	1.72277
7.5	7.7	2.01490	2.04122
7.6	10.3	2.02815	2.33214

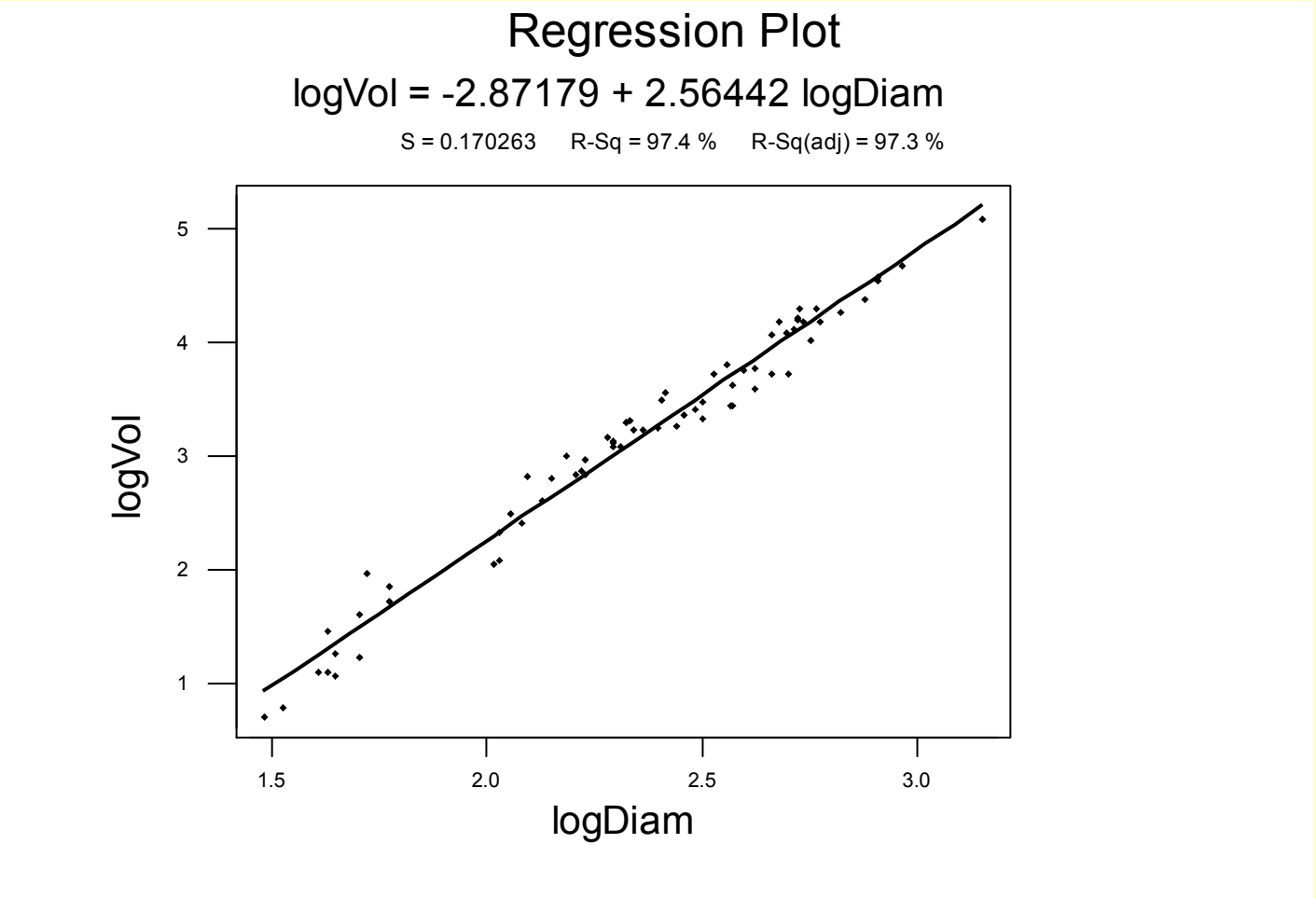
… **and so on** …

Transform predictor **diameter** to **$\log_e$(diameter)**

Transform response **volume** to **$\log_e$(volume)**

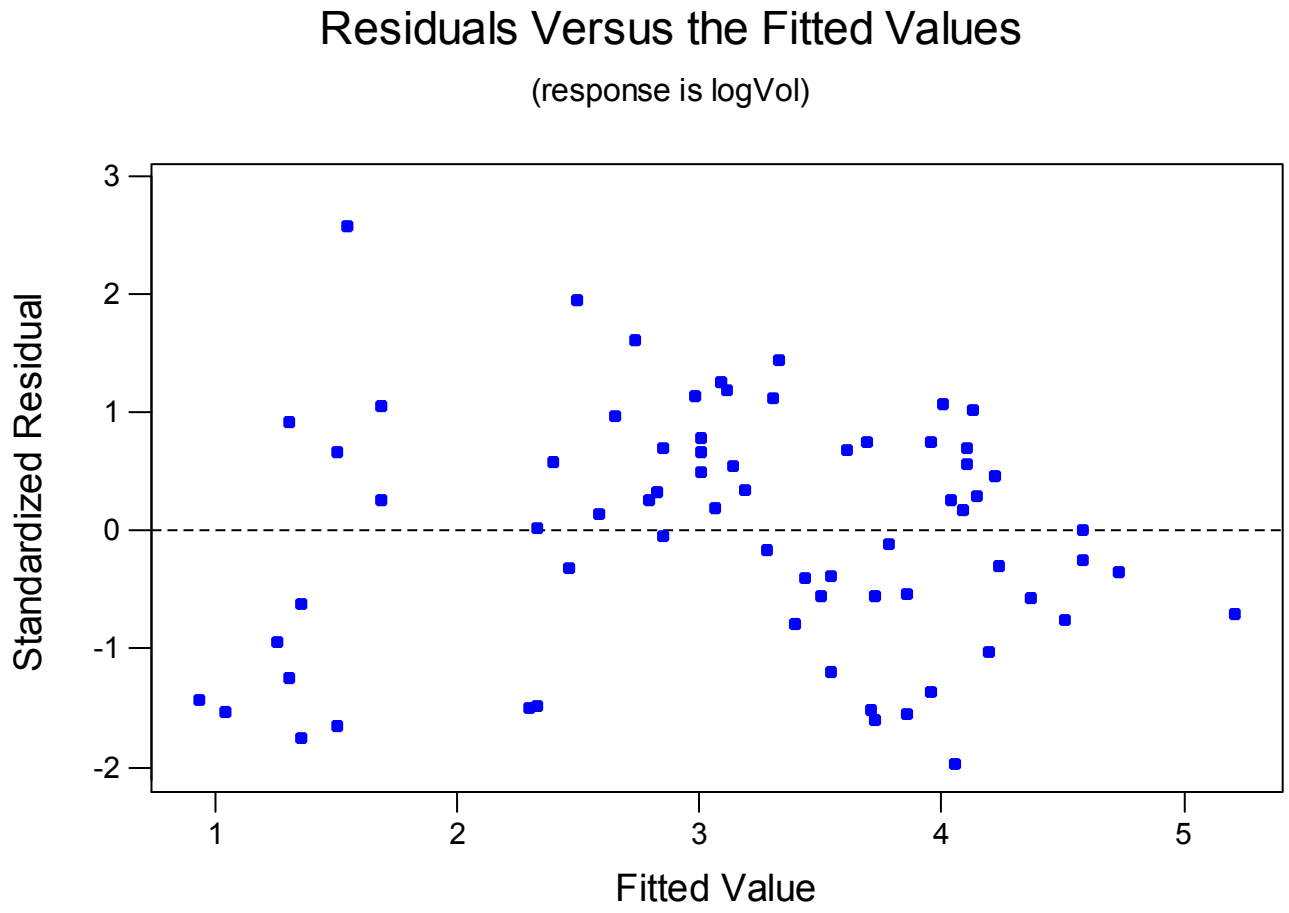
Example 3

Fitted line plot using transformed *X* and *Y* values



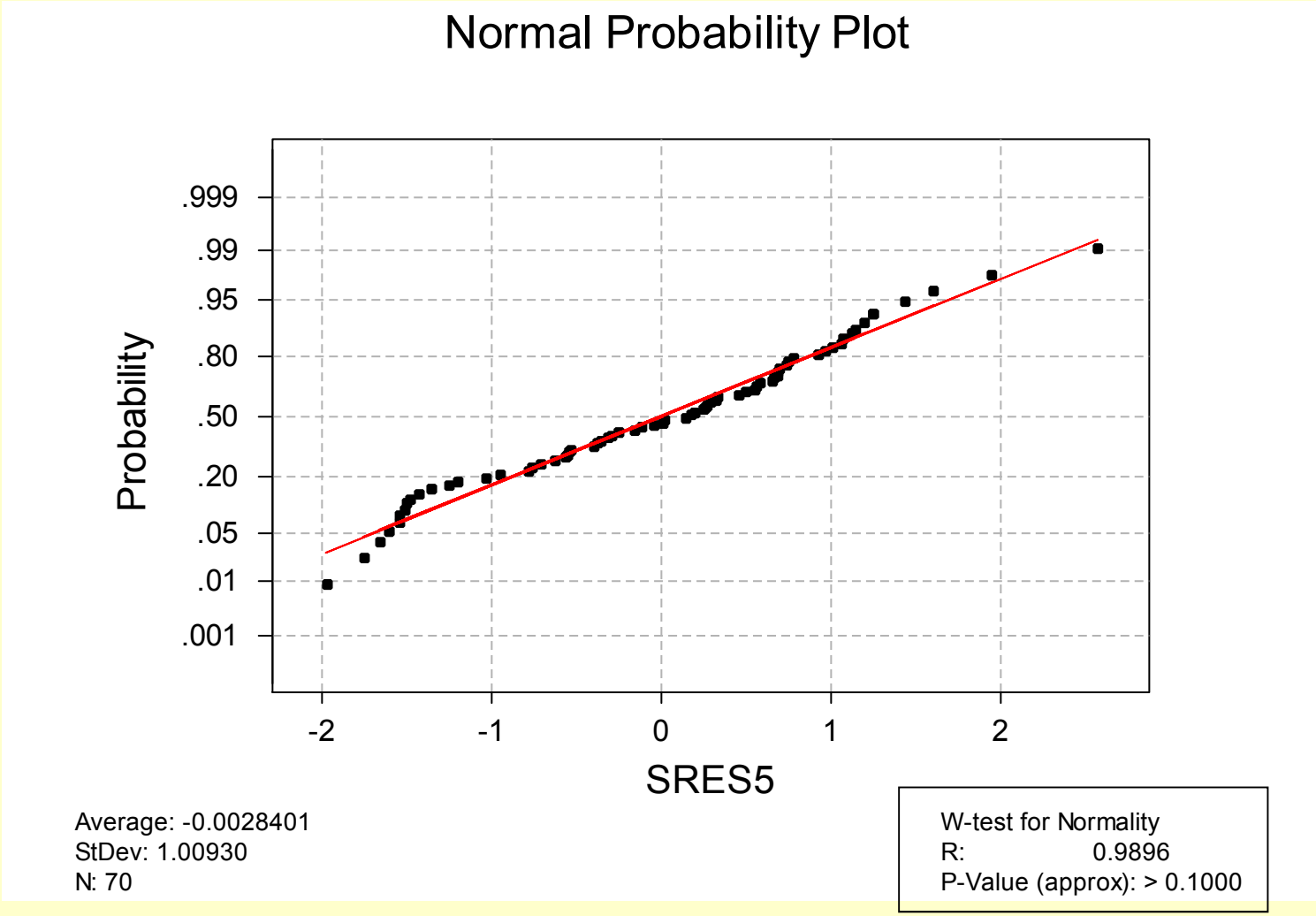
Example 3

Residual plot using transformed *X* and *Y* values



Example 3

Normal probability plot using transformed *X* and *Y* values



Transformation strategies

Effects of transformations

- **Transforming the *Y* values** corrects the problems with the error terms – and may simultaneously help non-linearity.
- **Transforming the *X* values** can only correct non-linearity.

Transformation strategies

- If form of the relationship between x and y is known, then it may be possible to **find a linearizing transformation analytically**.
- Fitting a regression model **empirically** generally requires **trial and error** – try different transformations to see which does best.

Transformation strategies

Finding a linearizing transformation analytically

Knowing functional relationship is of the power form

If the relationship between x and y is of the **power form**:

$$y = \alpha x^{\beta}$$

taking log of both sides transforms it into a linear form:

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

Knowing functional relationship is of the exponential form

If the relationship between x and y is of **exponential form**:

$$\boldsymbol{y = \alpha e^{\beta x}}$$

taking log of both sides transforms it into a linear form:

$$\log_e \boldsymbol{y} = \log_e \alpha + \beta \boldsymbol{x}$$

Transformation strategies

Finding a transformation by trial and error

Family of **power transformations**

The **most common transformation** involves transforming the response by taking it to some power λ. That is:

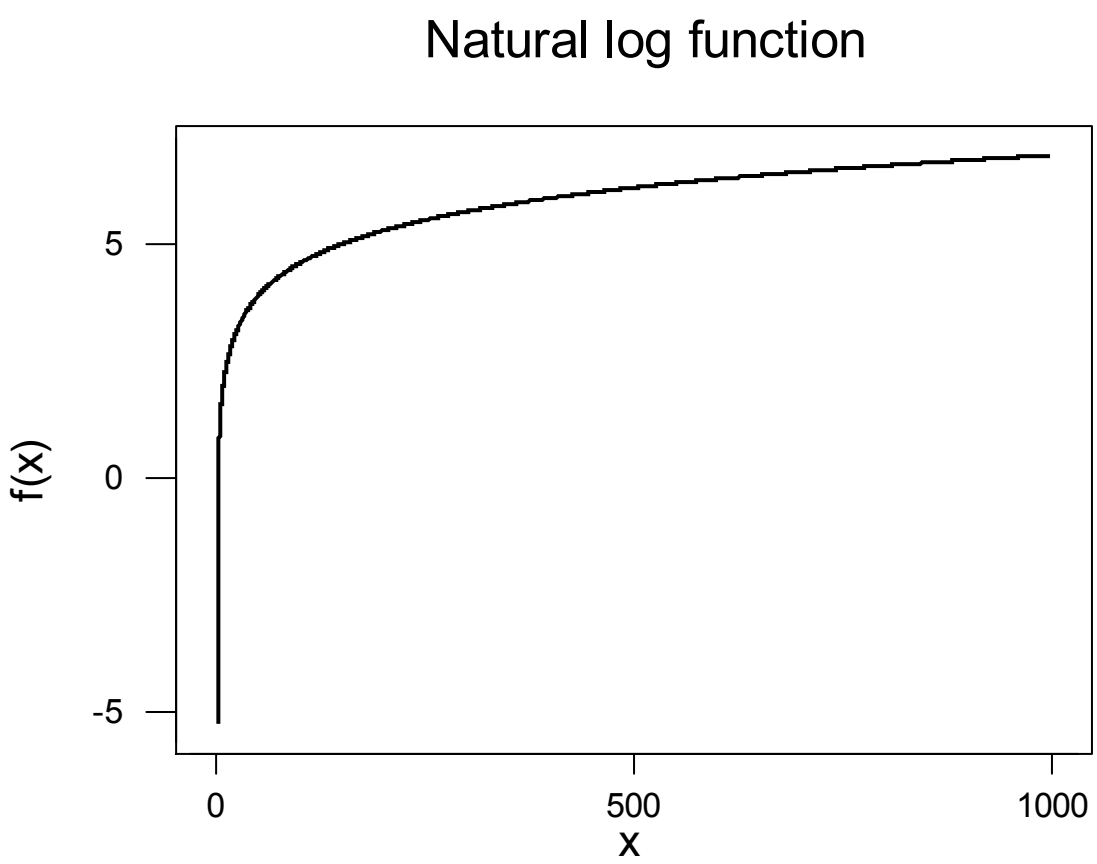
$$y' = y^{\lambda}$$

Most commonly, for interpretation reasons, λ is a number between -1 and 2, such as -1, -0.5, 0, 0.5, (1), 1.5, and 2.

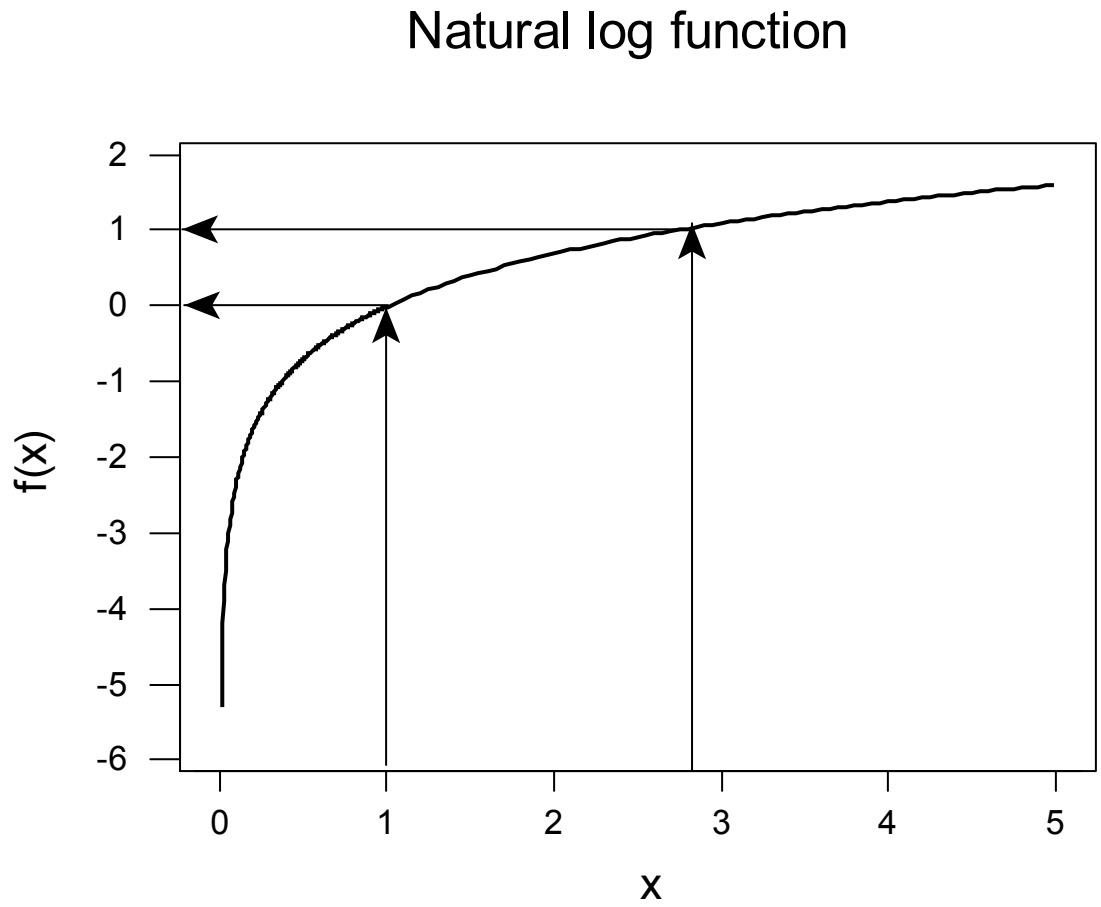
When λ = 0, the transformation is taken to be the log transformation. That is:

$$y' = \log_e y$$

Effect of $\log_e$ transformation



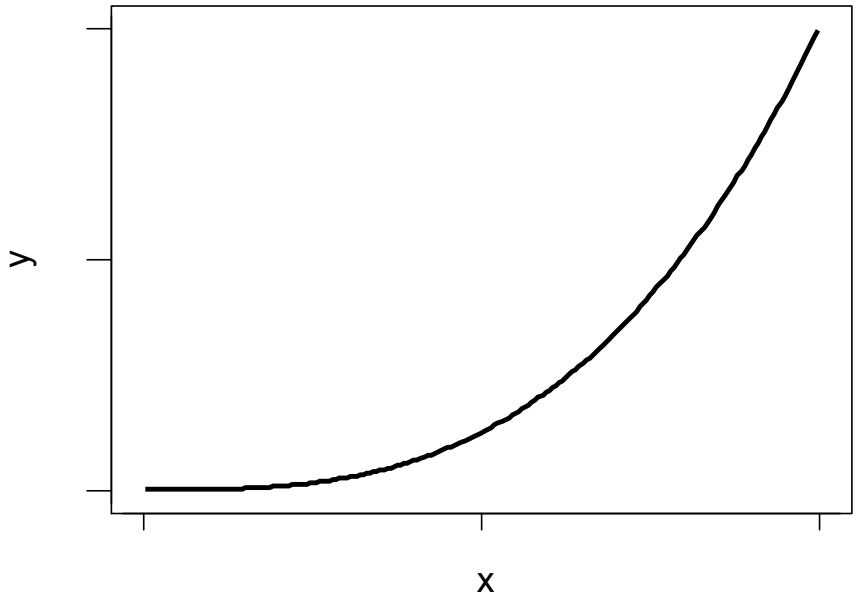
Effect of $\log_e$ transformation



Some guidelines for specifying λ

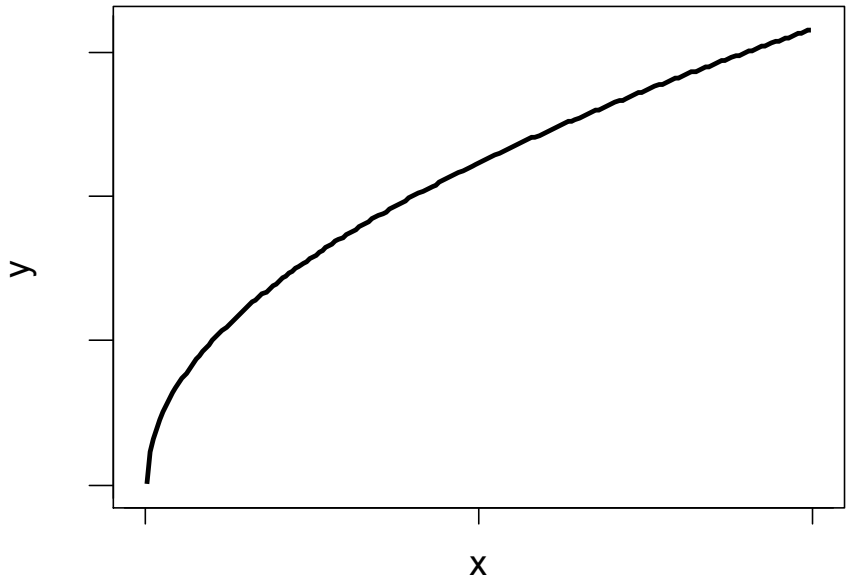
- To make **smaller values** more spread out, use a **smaller λ**.
- To make **larger values** more spread out, use a **larger λ**.

Possible transformations



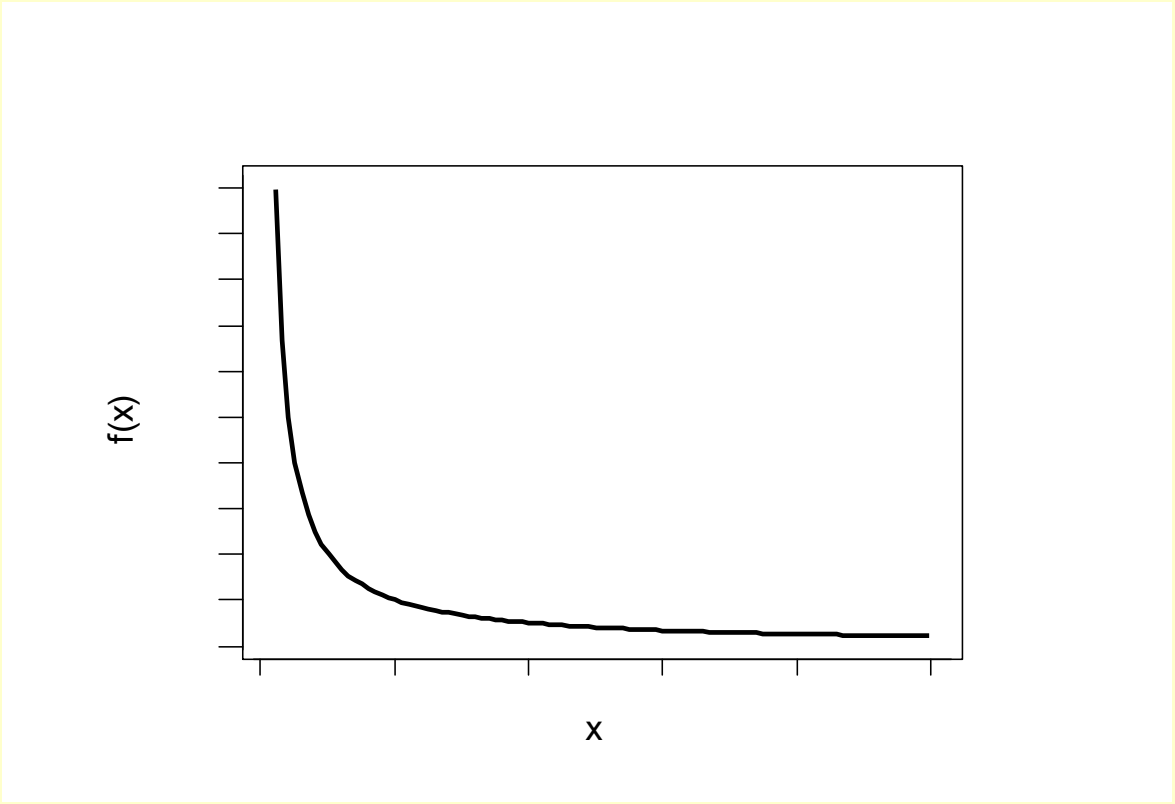
x	y
x^2	y
x^3	y
x	$\log y$
x	$-1/y$

Possible transformations



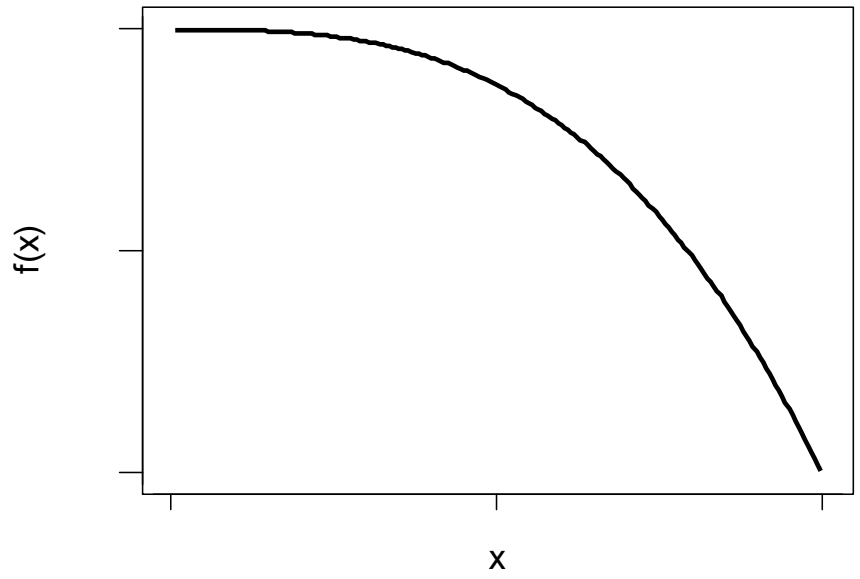
x	y
$\log x$	y
$-1/x$	y
x	y^3
x	y^2

Possible transformations



x	y
$\log x$	y
$-1/x$	y
x	$\log y$
x	$-1/y$
$\log x$	$\log y$

Possible transformations



x	y
x^2	y
x^3	y
x	y^2
x	y^3

Transformation strategies

Variance stabilizing transformations

Common variance stabilizing transformations

If the response is a Poisson count, so that the variance is proportional to the mean, use the **square root transformation**:

$$y' = y^{1/2} = \sqrt{y}$$

If the response is a binomial proportion, use the **arcsine square root transformation**:

$$\hat{p}' = \sin^{-1}\left(\sqrt{\hat{p}}\right)$$

Common variance stabilizing transformations

If the variance is proportional to the mean squared, use the **natural log transformation**:

$$y' = \log_e(y)$$

If the variance is proportional to the mean to the fourth power, use the **reciprocal transformation**:

$$y' = -\frac{1}{y}$$