STAT 511

Lecture : Simple linear regression

Devore: Section 12.1-12.4

Prof. Michael Levine

April 26, 2020

- A simple linear regression investigates the relationship between the two variables that is not deterministic. The variable whose value is fixed by the experimenter is called the **independent, predictor** or **explanatory** variable. For fixed x, the second variable is random; it is referred to as the **dependent** or **response** variable.
- The data is usually given as n pairs $(x_1, y_1), \ldots, (x_n, y_n)$. A **scatter plot** gives a good indication of the nature of the relationship between the two variables.

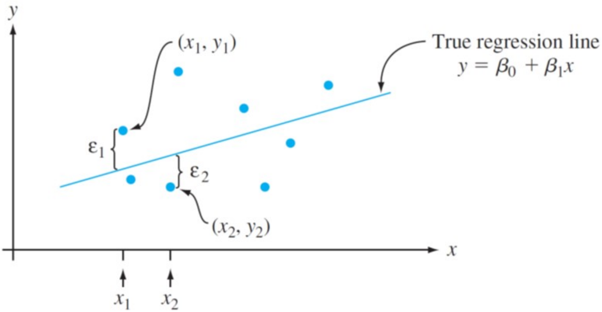
A Linear Model

- The usual linear regression model is

$$Y = \beta_0 + \beta_1 x + \epsilon$$

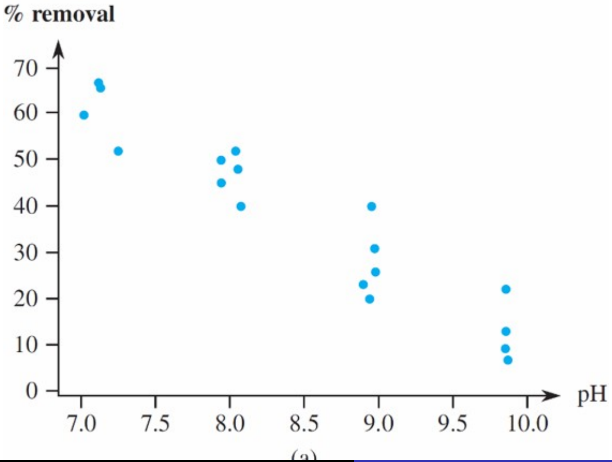
  where $\varepsilon \sim N(0, \sigma^2)$
- ε is the **random error term** or the **random deviation** while the line $y = \beta_0 + \beta_1 x$ is called the **true (or population) regression line**
- This model assumes that $E\,Y = \beta_0 + \beta_1 x$ while the deterministic model assumes that $y = \beta_0 + \beta_1 x$

Illustration



Appropriateness of the linear regression model

- Sometimes, such a model is suggested by physical considerations
- More commonly, it is simply suggested from an inspection of a scatter plot

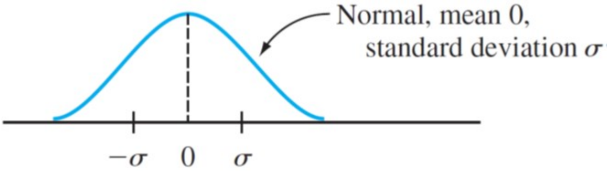


Implications of the linear regression model

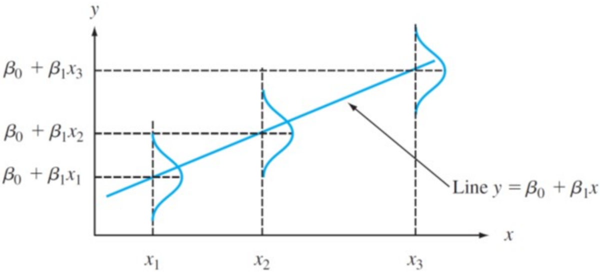
- Let x^* be a specific value of x,, corresponding mean and variance are $E\left(Y|x^*\right)$ and $V\left(Y|x^*\right)$
- E.g. x is the age of a child, Y is the size of the child's vocabulary
- The meaning: $E\left(Y|x^*\right) = \beta_0 + \beta_1 x^*$ and $V\left(Y|x^*\right) = \sigma^2$

Implications of the linear regression model

- Thus, the line $Y = \beta_0 + \beta_1 x$ is the line of of mean values
- The slope β_1 is the expected change in $E\,Y$ with one unit change in x
- σ^2 does not depend on x so the amount of variability in Y stays the same for all x



Illustration



Example

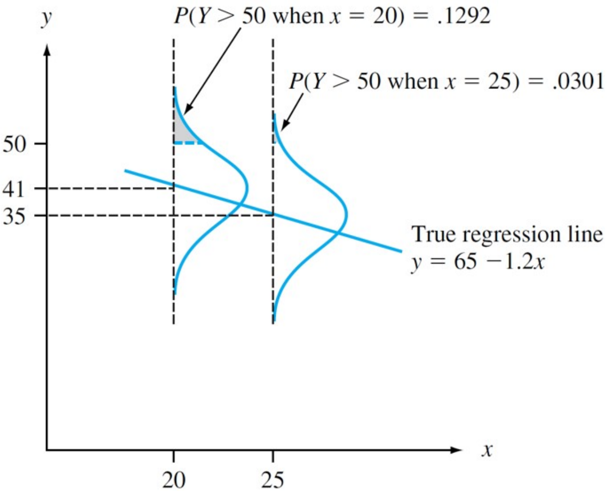
- The relationship between applied stress x and time-to-failure y is described by the simple linear regression model with true regression line $y = 65–1.2x$ and $\sigma = 8$
- For any fixed value x^* of stress, time-to-failure has a normal distribution with mean value $65–1.2x^*$ and standard deviation 8
- For $x = 20$,

$$E\left(Y|x^* = 20\right) = 65 - 1.2 * 20 = 41$$

- Thus, e.g.

$$P(Y > 50|x^* = 20) = P\left(Z > \frac{50 - 41}{8}\right) = 1 - \Phi(1.13) = .1292$$

Illustration



Example

- Suppose that Y_1 denotes an observation on time-to-failure made with $x = 25$ and Y_2 denotes an independent observation made with $x = 24$
- $Y_1 - Y_2$ is normally distributed, $E(Y_1 - Y_2) = \beta_1 = -1.2$; $V(Y_1 - Y_2) = 2\sigma^2 = 128$
- Thus,

$$P(Y_1 - Y_2 > 0) = P\left(Z > \frac{0 - (-1.2)}{11.314}\right) = P(Z > .11) = .4562$$

- Even though the slope is negative, it is not inconceivable that $Y_1 > Y_2$

Estimating Model Parameters

- The usual way to estimate parameters of a linear regression models is by using the **Least Squares** approach suggested by Gauss
- The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1 x$ is

$$y_i - (b_0 + b_1 x_i)$$

- The sum of squared deviations from the data points $(x_1, y_1), \ldots, (x_n, y_n)$ to the line is

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

Estimating Model Parameters

- The point estimates of the **true** model coefficients β_0 and β_1 are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$. They are the values that minimize $f(b_0, b_1)$.
- In other words, they are such that $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ for any b_0 and b_1.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the **least squares estimates**
- The **estimated regression line** is $y = \hat{\beta}_0 + \hat{\beta}_1 x$

System of normal equations

- $$nb_0 + b_1(\sum x_i) = \sum y_i$$
  $$b_0(\sum x_i) + b_1(\sum x_i^2) = \sum x_i y_i$$
- If not all x_i are identical, there is a unique solution - least squares

Solutions

- $$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$
- $$b_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Example

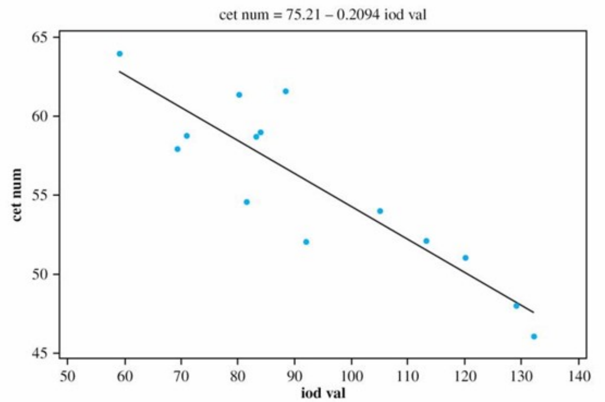
- The cetane number is a critical property in specifying the ignition quality of a fuel used in a diesel engine. Determination of this number for a biodiesel fuel is expensive and time-consuming.
- The iodine value is the amount of iodine necessary to saturate a sample of 100 g of oil
- x = iodine value (g) and y = cetane number for a sample of 14 biofuels.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

Example

- $\hat{\beta}_1 = -.20938742$
- Thus, expected change in true average cetane number associated with 1 g decrease in iodine value is about $-.209$
- The estimated $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 75.212432$

Example



cet num = 75.21 – 0.2094 iod val
65
60
55
50
45
cet num
50
60
70
80
90
100
110
120
130
140
iod val

Estimating σ^2

- Estimating σ^2 is needed to get confidence intervals and/or test hypotheses about coefficients of the regression model
- **The fitted values** are $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \ldots, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$
- **The residuals** are $y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n$
- The residuals are needed to estimate the variance of errors; specifically,

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

where the **error sum of squares** is $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

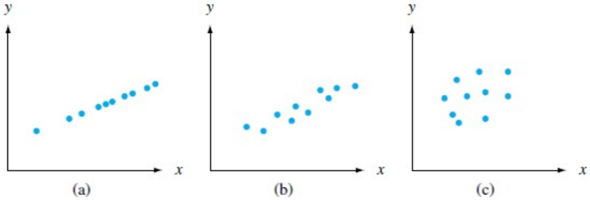
Computational formula for SSE

- The direct computation of SSE is rather involved
- A better option is to use the computation formula

$$SSE = S_{yy} - \hat{\beta}_1 S_{xy}$$

- This formula does not involve computation of predicted values and residuals
- It is, however, very sensitive to the rounding effect in $\hat{\beta}_0$ and $\hat{\beta}_1$

Variation in the data



R^2 - coefficient of determination I

- How much of the total variation can the linear regression model explain? That total variation will be described by the **total sum of squares**

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

- It is always true that SSE$<$ SST, so we can define

$$R^2 = 1 - \frac{SSE}{SST}$$

which is a number between 0 and 1 that suggests how much of the total variation is explained by the regression model

R^2 - coefficient of determination II

- Its alternative form is $R^2 = \frac{SSR}{SST}$ where $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the **regression sum of squares**
- The same identity as before in ANOVA analysis is true

$$SST = SSR + SSE$$

- Cetane number-iodine value example: high value of R^2

Parameter estimators

- An estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- An estimator of β_0 is

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_0 \bar{x}$$

- An estimator of σ^2 is

$$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i Y_i}{n-2}$$

$\hat{\beta}_1$ as a linear estimator of the slope

- Verify that

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$$

  where $c_i = \frac{x_i - \bar{x}}{S_{xx}}$

- Consequently, $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$
- The variance of $\hat{\beta}_1$ is estimated by $\frac{s}{\sqrt{S_{xx}}}$.

A confidence interval for β_1

- Note that
$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t_{n-2}$$
- Thus, the $100\%(1-\alpha)$ CI for β_1 is
$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \frac{S}{\sqrt{S_{xx}}}$$

Example

- When damage to a timber structure occurs, it may be more economical to repair the damaged area rather than replace the entire structure
- The dependent variable is y = rupture load (N) and the independent variable is anchorage length (the additional length of material used to bond at the junction), in mm

x	50	50	80	80	110	110	140	140	170	170
y	17,052	14,063	26,264	19,600	21,952	26,362	26,362	26,754	31,654	32,928

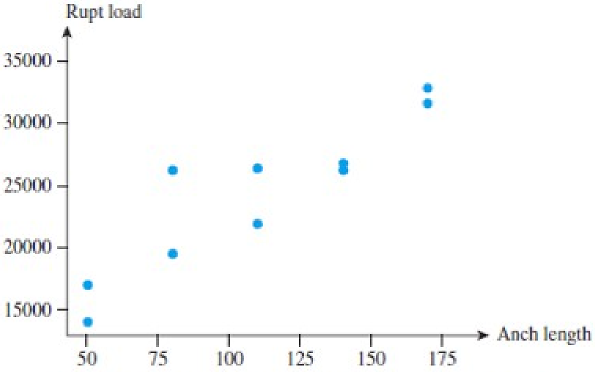


Figure 12.14 Scatterplot of the data from Example 12.11

Example

- Main quantities are $S_{xx} = 18,000$ error df $10 - 2 = 8$, $s = 2661.33$. The estimated standard error is

$$\frac{s}{\sqrt{S_{xx}}} = 19.836$$

- The 95% confidence interval is

$$123.64 \pm (2.306)(19.836) = (77.90, 169.38)$$

Hypothesis testing

- The most common is the **model utility test** $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
- The test statistic value is

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$$

- Then, if e.g. $H_a : \beta_1 > \beta_{10}$, we have the P-value as the area under the t_{n-2} curve to the right of t

Example

- Mopeds are very popular in Europe because of cost and ease of operation
- They can be dangerous if performance characteristics are modified. One of the features commonly manipulated is the maximum speed
- simple linear regression analysis of the variables x = test track speed (km/h) and y = rolling test speed

x	42.2	42.6	43.3	43.5	43.7	44.1	44.9	45.3	45.7
y	44	44	44	45	45	46	46	46	47

x	45.7	45.9	46.0	46.2	46.2	46.8	46.8	47.1	47.2
y	48	48	48	47	48	48	49	49	49

12.3 Inferences About the Slope Parameter β_1

$s_{\hat{\beta}_1}$ $\quad t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$

```
The regression equation is
roll spd = −2.22 + 1.08 trk spd

Predictor        Coef      SE Coef        T         P
Constant       −2.224       3.528      −0.63     0.537
trk spd       1.08342     0.07806      13.88     0.000 ← P-value for
                                                          model utility
S = 0.506890   R-Sq = 92.3%   R-Sq(adj) = 91.9%          test

Analysis of Variance

Source            DF       SS       MS        F       P
Regression         1   49.500   49.500   192.65   0.000
Residual Error    16    4.111    0.257
Total             17   53.611
```

Regression and ANOVA

Source of Variation	df	Sum of Squares	Mean Square	f
Regression	1	SSR	SSR	$\frac{\text{SSR}}{\text{SSE}/(n-2)}$
Error	$n-2$	SSE	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SST		

- Note that $t^2 = f$ for the test of $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

Source of Variation	df	Sum of Squares	Mean Square	f
Regression	1	SSR	SSR	$\frac{\text{SSR}}{\text{SSE}/(n-2)}$
Error	$n-2$	SSE	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SST		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10698	2338.67544	4.57	0.0018
Anch Lngth	1	123.64333	19.83639	6.23	0.0003

Figure 12.15 SAS output for the data of Example 12.11

Inference about the mean $\mu_{Y \cdot x^*}$

- For a given value x^*, the estimated average value of Y is $\hat{\beta}_0 + \hat{\beta}_1 x^*$
- It can also be viewed as the prediction at the given point x^*
- It is possible to represent the estimated average value of Y as

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = \sum_{i=1}^{n} d_i Y_i$$

where $d_i = \frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Summary

- $E(\hat{Y}) = \beta_0 + \beta_1 x^*$, and the variance is

$$V(\hat{Y}) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

- The estimated variance results from the above by replacing σ^2 with s; $\hat{Y}$ is also normally distributed
- To construct a confidence interval or to test a hypothesis, just note that

$$T = \frac{\hat{Y} - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S_{\hat{Y}}} \sim t_{n-2}$$

Inferences concerning $\mu_{Y \cdot x^*}$

- The variable

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}}$$

  has a t distribution with $n - 2$ df
- The $100\%(1-\alpha)$ CI for $E\left(Y|x^*\right) = \mu_{Y \cdot x^*}$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2} s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{y} \pm t_{\alpha/2, n-2} s_{\hat{Y}}$$

Example

- Corrosion of steel reinforcing bars is the most important durability problem for reinforced concrete structures
- Representative data on x = carbonation depth (mm) and y = strength (MPa) for a sample of core specimens from a building in Singapore
- The scatter plot supports the use of simple linear regression; thus, let us obtain 95% CI for $\beta_0 + \beta_1 45$ for $x = 45$ mm

Example

- First, $\hat{\beta}_1 = -.297561$ and $\hat{\beta}_0 = 27.182936$ so

$$\hat{y} = 27.182936 - .297561 * 45 = 13.79$$

- The estimated

$$s_{\hat{Y}} = 2.8640\sqrt{\frac{1}{18} + \frac{45 - 36.6111)^2}{4840.7778}} = .7582$$

- Th 16 df t-critical value is 2.120 and so

$$13.79 \pm (2.120)(.7582) = (12.18, 15.40)$$

Example

- The following output results from a request to fit the simple linear regression model and calculate confidence intervals for the mean value of strength at depths of 45 mm and 35 mm

The regression equation is strength = 27.2 − 0.298 depth

Predictor	Coef	Stdev	t-ratio	P
Constant	27.183	1.651	16.46	0.000
depth	−0.29756	0.04116	−7.23	0.000

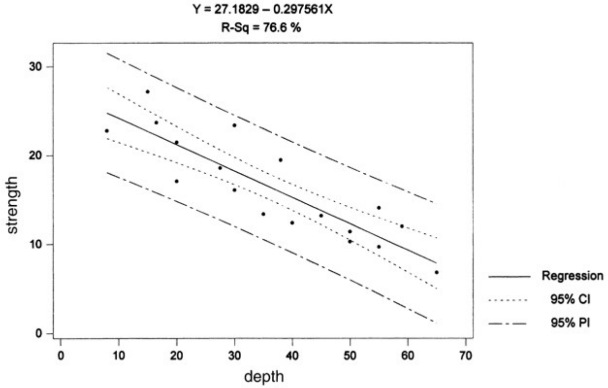
s = 2.864 R-sq = 76.6% R-sq(adj) = 75.1%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	1	428.62	428.62	52.25	0.000
Error	16	131.24	8.20		
Total	17	559.86			

Fit	Stdev.Fit	95.0% C.I.	95.0% P.I.
13.793	0.758	(12.185, 15.401)	(7.510, 20.075)

Fit	Stdev.Fit	95.0% C.I.	95.0% P.I.
16.768	0.678	(15.330, 18.207)	(10.527, 23.009)


Y = 27.1829 – 0.297561X
R-Sq = 76.6 %
30
20
10
0
strength
0
10
20
30
40
50
60
70
depth
Regression
95% CI
95% PI

CI's for multiple values of x

- In some situations, a CI is desired not just for a single x value but for two or more x values
- Suppose an investigator wishes a CI both for $\mu_{Y \cdot v}$ and for $\mu_{Y \cdot w}$, where v and w are two different values of the independent variable
- The intervals are not independent because the same $\hat{\beta}_0$, $\hat{\beta}_1$ and S are used in each. We therefore cannot assert that the joint confidence level for the two intervals is exactly 90% even if we select $\alpha = 0.05$
- It can be shown, though, that if the $100\%(1-\alpha)$ CI is computed both for $x = v$ and $x = w$ to obtain joint CIs for $\mu_{Y \cdot v}$ow and for $\mu_{Y \cdot w}$, then the joint confidence level on the resulting pair of intervals is at least $100\%(1-2\alpha)$.

A prediction interval for a future value of Y

- Sometimes, an investigator may wish to obtain an interval of plausible values for the value of Y associated with some future observation when the independent variable has value x^*
- We may want to relate vocabulary size y to the age of a child x. The CI with $x^* = 6$ would provide an estimate of true average vocabulary size for all 6-year-old children
- Alternatively, we might wish an interval of plausible values for the vocabulary size of a particular 6-year-old child

The error of prediction

- The error of prediction is $Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$
- The variance of the prediction error is

$$V(Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

- The expected value of the prediction error is $E\left(Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)\right) = 0$ and

$$T = \frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

Prediction interval

- The prediction interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

- This interval is always wider than the corresponding confidence interval

Example

- Let's return to the carbonation depth-strength data and calculate a 95% PI for a strength value that would result from selecting a single core specimen whose carbonation depth is 45 mm
- The relevant quantities are $\hat{y} = 13.79$, $s_{\hat{Y}} = .7582$, $s = 2.8640$
- For a prediction level of 95% based on $n - 2 = 16$ df the critical value is 2.120
- The prediction interval is then

$$13.79 \pm 2.120\sqrt{(2.8640)^2 + (.7582)^2} = (7.51, 20.07)$$