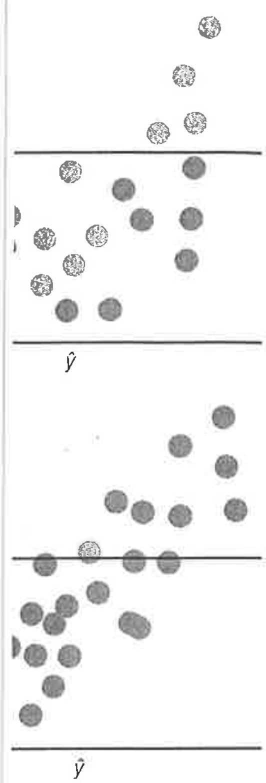


ersus fits.

he "Rorschach effect" of see-
ı examining these plots is that
ly identifiable patterns in the

lso several diagnostic hypoth-
of the regression assumptions.
ate fit of the residuals to a nor-
variance assumption has been
ısed. For determining whether
er the Durban–Watson test or
these diagnostic tests may be

ture needed to perform infer-
, such as point estimates, and
relation, standard error of the

ey Publishers, Hoboken, New Jersey,

estimate, and r^2, may still be undertaken even if these assumptions are not met, if the results are cross-validated. What is not allowed by violated assumptions is statistical inference. But we as data miners and big data scientists understand that inference is not our primary *modus operandi*. Rather, data mining seeks confirmation through cross-validation of the results across data partitions. For example, if we are examining the relationship between *outdoor event ticket sales* and *rainfall amounts*, and if the training data set and test data set both report correlation coefficients of about −0.7, and there is graphical evidence to back this up, then we may feel confident in reporting to our client *in a descriptive manner* that the variables are negatively correlated, even if both variables are not normally distributed (which is the assumption for the correlation test). We just cannot say that the correlation coefficient has a statistically significant negative value, because the phrase "statistically significant" belongs to the realm of inference. So, for data miners, the keys are to (i) cross-validate the results across partitions, and (ii) restrict the interpretation of the results to descriptive language, and avoid inferential terminology.

8.10 INFERENCE IN REGRESSION

Inference in regression offers a systematic framework for assessing the significance of linear association between two variables. Of course, analysts need to keep in mind the usual caveats regarding the use of inference in general for big data problems. For very large sample sizes, even tiny effect sizes may be found to be statistically significant, even when their practical significance may not be clear.

We shall examine five inferential methods in this chapter, which are as follows:

1. The *t-test* for the relationship between the response variable and the predictor variable.
2. The correlation coefficient test.
3. The confidence interval for the slope, β_1.
4. The confidence interval for the mean of the response variable, given a particular value of the predictor.
5. The prediction interval for a random value of the response variable, given a particular value of the predictor.

In Chapter 9, we also investigate the *F-test* for the significance of the regression as a whole. However, for simple linear regression, the *t-test* and the *F-test* are equivalent.

How do we go about performing inference in regression? Take a moment to consider the form of the true (population) regression equation.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

This equation asserts that there is a linear relationship between y on the one hand, and some function of x on the other. Now, β_1 is a model parameter, so that it is a constant whose value is unknown. Is there some value that β_1 could take such that, if β_1 took that value, there would no longer exist a linear relationship between x and y?

Consider what would happen if β_1 was zero. Then the true regression equation would be as follows:

$$y = \beta_0 + (0)x + \varepsilon$$

In other words, when $\beta_1 = 0$, the true regression equation becomes:

$$y = \beta_0 + \varepsilon$$

That is, a linear relationship between x and y no longer exists. However, if β_1 takes on any conceivable value other than zero, then a linear relationship of some kind exists between the response and the predictor. Much of our regression inference in this chapter is based on this key idea, that the linear relationship between x and y depends on the value of β_1.

8.11 *t*-TEST FOR THE RELATIONSHIP BETWEEN *x* AND *y*

Much of the inference we perform in this section refers to the regression of *rating* on *sugars*. The assumption is that the residuals (or standardized residuals) from the regression are approximately normally distributed. Figure 8.15 shows that this assumption is validated. There are some strays at either end, but the bulk of the data lie within the confidence bounds.

The least squares estimate of the slope, b_1, is a statistic, because its value varies from sample to sample. Like all statistics, it has a sampling distribution with

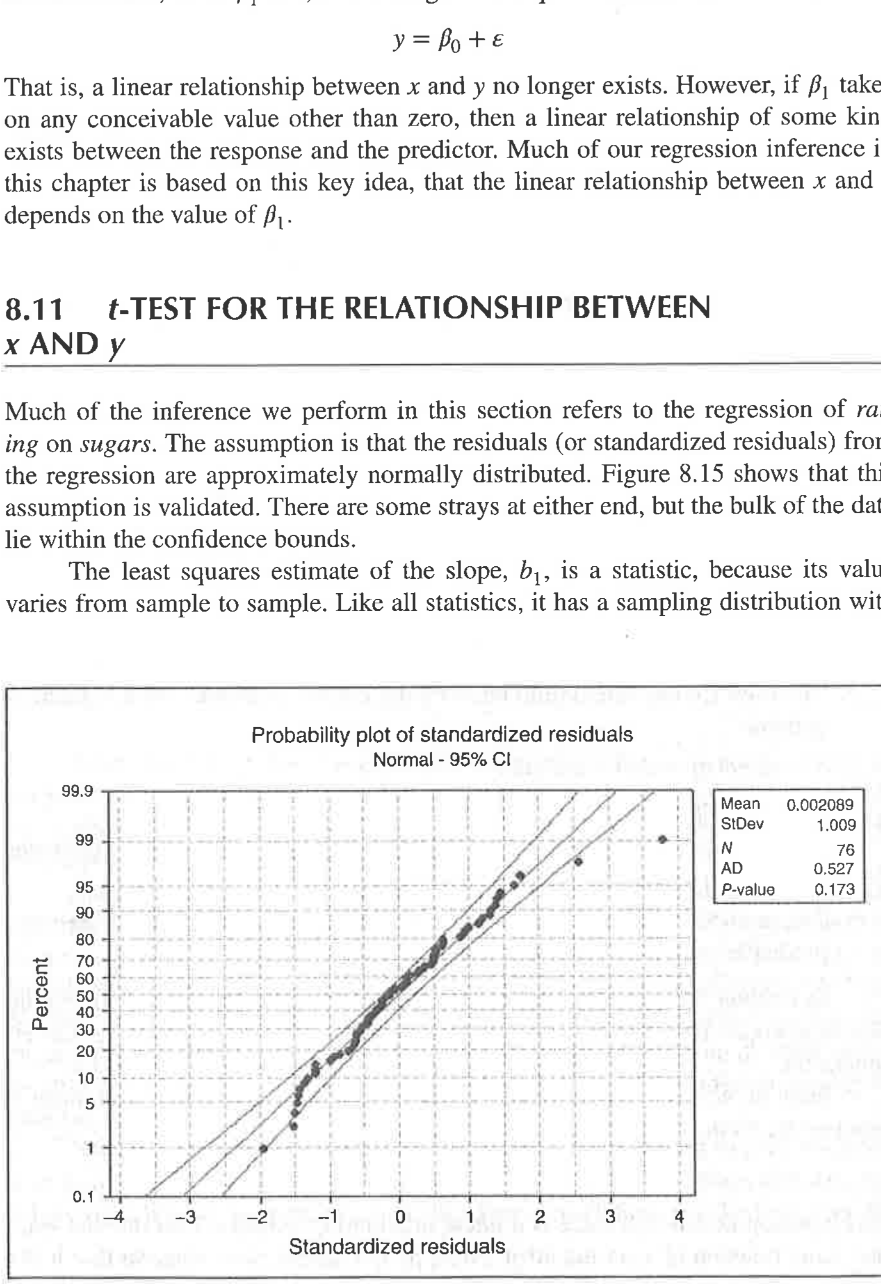


Figure 8.15 Normal probability plot of the residuals for the regression of *rating* on *sugars*.

regression equation would

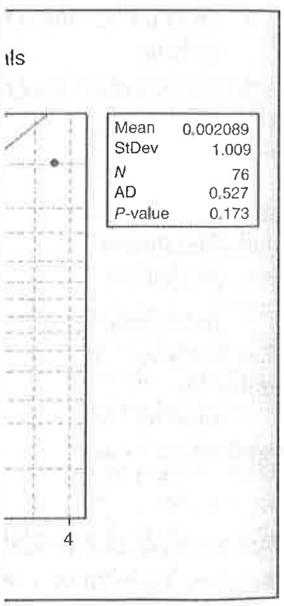
becomes:

xists. However, if β_1 takes relationship of some kind our regression inference in ationship between x and y

TWEEN

s to the regression of *rat-* andardized residuals) from igure 8.15 shows that this nd, but the bulk of the data

statistic, because its value sampling distribution with



gression of *rating* on *sugars*.

a particular mean and standard error. The sampling distribution of b_1 has as its mean the (unknown) value of the true slope β_1, and has as its standard error, the following:

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum x^2 - \left(\sum x\right)^2 / n}}$$

Just as one-sample inference about the mean μ is based on the sampling distribution of $\bar{x}$, so regression inference about the slope β_1 is based on this sampling distribution of b_1. The point estimate of σ_{b_1} is s_{b_1}, given by

$$s_{b_1} = \frac{s}{\sqrt{\sum x^2 - \left(\sum x\right)^2 / n}}$$

where s is the standard error of the estimate, reported in the regression results. The s_{b_1} statistic is to be interpreted as a measure of the variability of the slope. Large values of s_{b_1} indicate that the estimate of the slope b_1 is unstable, while small values of s_{b_1} indicate that the estimate of the slope b_1 is precise. The *t-test* is based on the distribution of $t = \frac{(b_1-\beta_1)}{s_{b_1}}$, which follows a t-distribution with $n-2$ degrees of freedom. When the null hypothesis is true, the test statistic $t = \frac{b_1}{s_{b_1}}$ follows a t-distribution with $n-2$ degrees of freedom. The *t-test* requires that the residuals be normally distributed.

To illustrate, we shall carry out the *t-test* using the results from Table 8.7, the regression of nutritional rating on sugar content. For convenience, Table 8.7 is reproduced here as Table 8.11. Consider the row in Table 8.11, labeled "Sugars."

- Under "Coef" is found the value of b_1, -2.4614.
- Under "SE Coef" is found the value of s_{b_1}, the standard error of the slope. Here, $s_{b_1} = 0.2417$.
- Under "T" is found the value of the *t-statistic*; that is, the test statistic for the *t-test*, $t = \frac{b_1}{s_{b_1}} = \frac{-2.4614}{0.2417} = -10.18$.
- Under "P" is found the p-value of the *t-statistic*. As this is a two-tailed test, this p-value takes the following form: p-value $= P(|t| > |t_{obs}|)$, where t_{obs} represent the observed value of the *t-statistic* from the regression results. Here, p-value $= P(|t| > |t_{obs}|) = P(|t| > |-10.18|) \approx 0.000$, although, of course, no continuous p-value ever precisely equals zero.

The hypotheses for this hypothesis test are as follows. The null hypothesis asserts that no linear relationship exists between the variables, while the alternative hypothesis states that such a relationship does indeed exist.

H_0: $\beta_1 = 0$ (There is no linear relationship between sugar content and nutritional rating.)

H_a: $\beta_1 \neq 0$ (Yes, there is a linear relationship between sugar content and nutritional rating.)

We shall carry out the hypothesis test using the p-value method, where the null hypothesis is rejected when the p-value of the test statistic is small. What determines

TABLE 8.11 Results for regression of *nutritional rating* versus *sugar content*

```
The regression equation is
Rating = 59.9 - 2.46 Sugars


Predictor      Coef   SE Coef       T      P
Constant     59.853     1.998   29.96  0.000
Sugars      -2.4614    0.2417  -10.18  0.000


S = 9.16616    R-Sq = 58.4%    R-Sq(adj) = 57.8%


Analysis of Variance

Source           DF       SS      MS       F      P
Regression        1   8711.9  8711.9  103.69  0.000
Residual Error   74   6217.4    84.0
Total            75  14929.3


Unusual Observations

Obs  Sugars  Rating    Fit  SE Fit  Residual  St Resid
  1     6.0   68.40  45.08    1.08     23.32     2.56R
  4     0.0   93.70  59.85    2.00     33.85     3.78R

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

how small is small depends on the field of study, the analyst, and domain experts, although many analysts routinely use 0.05 as a threshold. Here, we have p-value $\approx$ 0.00, which is surely smaller than any reasonable threshold of significance. We therefore reject the null hypothesis, and conclude that a linear relationship exists between sugar content and nutritional rating.

8.12 CONFIDENCE INTERVAL FOR THE SLOPE OF THE REGRESSION LINE

Researchers may consider that hypothesis tests are too black-and-white in their conclusions, and prefer to estimate the slope of the regression line β_1, using a confidence interval. The interval used is a *t-interval*, and is based on the above sampling distribution for b_1. The form of the confidence interval is as follows.[5]

[5]The notation $100(1-\alpha)\%$ notation may be confusing. But suppose we let $\alpha = 0.05$, then the confidence level will be $100(1-\alpha)\% = 100(1-0.05)\% = 95\%$.

THE 100(1 − α)% CONFIDE OF THE REGRESSION LIN

We can be $100(1-\alpha)\%$ confide

where $t_{\alpha/2,n-2}$ is based on $n-2$

For example, let us cons regression line, β_1. We have th value for 95% confidence and Figure 8.16, we have $s_{b_1} = 0.$

$$b_1 - (t_{n-2})(s_{b_1}) =$$

$$b_1 + (t_{n-2})(s_{b_1}) =$$

We are 95% confident that the and −1.9780. That is, for eve decrease by between 1.9780 a within this interval, we can be variables, with 95% confiden

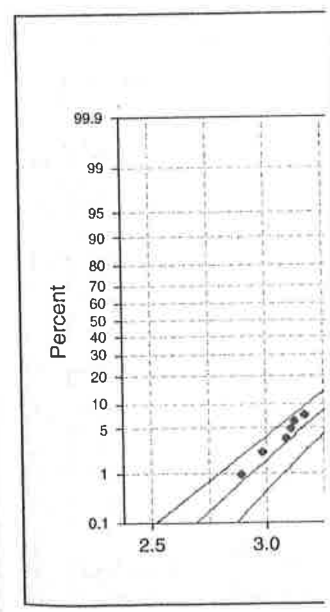


Figure 8.16 Probability plot c

THE 100(1 − α)% CONFIDENCE INTERVAL FOR THE TRUE SLOPE β_1 OF THE REGRESSION LINE

We can be 100(1 − α)% confident that the true slope β_1 of the regression line lies between:

$$b_1 \pm (t_{\alpha/2,n-2})(s_{b_1})$$

where $t_{\alpha/2,n-2}$ is based on $n-2$ degrees of freedom.

For example, let us construct a 95% confidence interval for the true slope of the regression line, β_1. We have the point estimate given as $b_1 = -2.4614$. The *t-critical* value for 95% confidence and $n-2 = 75$ degrees of freedom is $t_{75,95\%} = 2.0$. From Figure 8.16, we have $s_{b_1} = 0.2417$. Thus, our confidence interval is as follows:

$$b_1 - (t_{n-2})(s_{b_1}) = -2.4614 - (2.0)(0.2417) = -2.9448, \quad \text{and}$$

$$b_1 + (t_{n-2})(s_{b_1}) = -2.4614 + (2.0)(0.2417) = -1.9780.$$

We are 95% confident that the true slope of the regression line lies between −2.9448 and −1.9780. That is, for every additional gram of sugar, the nutritional rating will decrease by between 1.9780 and 2.9448 points. As the point $\beta_1 = 0$ is not contained within this interval, we can be sure of the significance of the relationship between the variables, with 95% confidence.

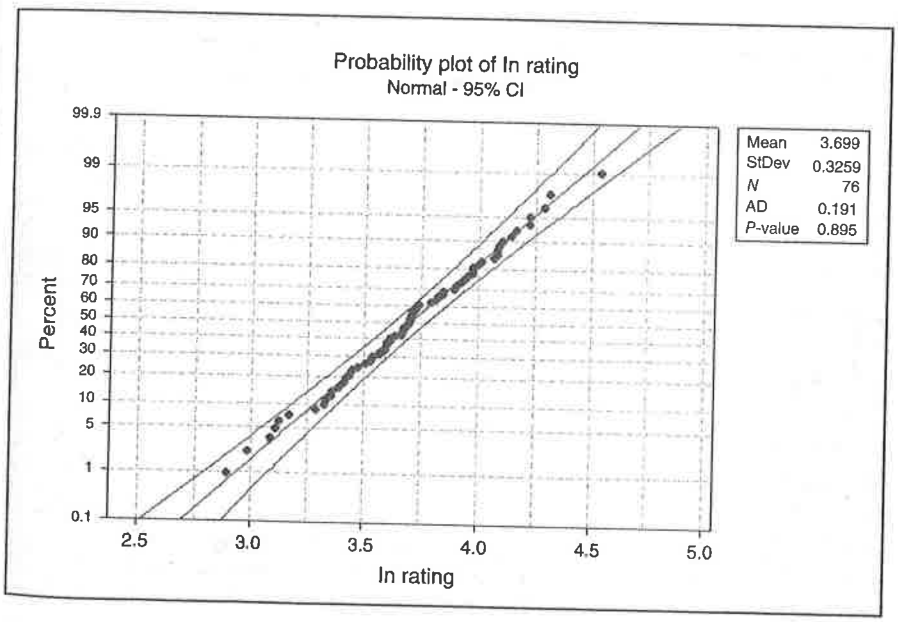


Figure 8.16 Probability plot of *ln rating* shows approximate normality.

8.13 CONFIDENCE INTERVAL FOR THE CORRELATION COEFFICIENT ρ

Let ρ ("rho") represent the population correlation coefficient between the x and y variables for the entire population. Then the confidence interval for ρ is as follows.

THE 100(1 − α)% CONFIDENCE INTERVAL FOR THE POPULATION CORRELATION COEFFICIENT ρ

We can be $100(1-\alpha)\%$ confident that the population correlation coefficient ρ lies between:

$$r \pm t_{\alpha/2,\,n-2} \cdot \sqrt{\frac{1-r^2}{n-2}}$$

where $t_{\alpha/2,n-2}$ is based on $n-2$ degrees of freedom.

This confidence interval requires that both the x and y variables be normally distributed. Now, *rating* is not normally distributed, but the transformed variable *ln rating* is normally distributed, as shown in Figure 8.16. However, neither *sugars* nor any transformation of *sugars* (see the *ladder of re-expressions* later in this chapter) is normally distributed. *Carbohydrates*, however, shows normality that is just barely acceptable, with an AD p-value of 0.081, as shown in Figure 8.17. Thus, the assumptions are met for calculating the confidence interval for the population correlation coefficient between *ln rating* and *carbohydrates*, but not between *ln rating*

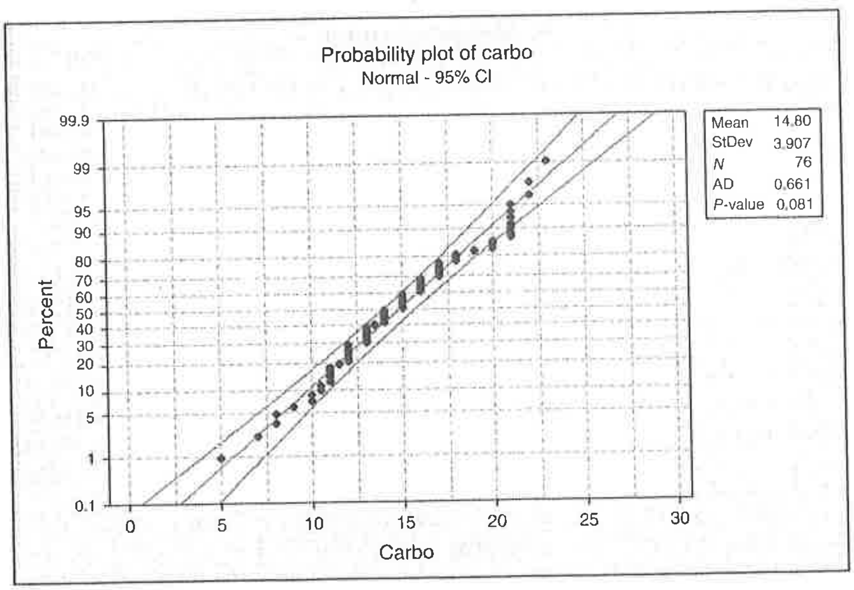


Figure 8.17 Probability plot of *carbohydrates* shows barely acceptable normality.

and *sugars*. Thus, let us proceed to construct a 95% confidence interval for ρ, the population correlation coefficient between *ln rating* and *carbohydrates*.

From Table 8.12, the regression output for the regression of *ln rating* on *carbohydrates*, we have $r^2 = 2.5\% = 0.025$, and the slope b_1 is positive, so that the sample correlation coefficient is $r = +\sqrt{r^2} = +\sqrt{0.025} = 0.1581$. The sample size is $n = 76$, so that $n - 2 = 74$. Finally, $t_{\alpha/2,\, n-2} = t_{0.025,\, 74}$ refers to the t-critical value with area 0.025 in the tail of the curve with 74 degrees of freedom. This value equals 1.99.[6] Thus, our 95% confidence interval for ρ is given by

$$r \pm t_{\alpha/2,\, n-2} \cdot \sqrt{\frac{1-r^2}{n-2}}$$

$$= 0.1581 \pm 1.99 \cdot \sqrt{\frac{1-0.025}{74}}$$

$$= (-0.0703,\ 0.3865)$$

[6]Use software such as Excel or Minitab to obtain this value, if desired.

TABLE 8.12 Regression of *ln rating* on carbohydrates

```
The regression equation is
ln rating = 3.50 + 0.0131 Carbo


Predictor        Coef   SE Coef      T      P
Constant       3.5043    0.1465  23.91  0.000
Carbo        0.013137  0.009576   1.37  0.174


S = 0.324030   R-Sq = 2.5%   R-Sq(adj) = 1.2%


Analysis of Variance

Source            DF      SS      MS     F      P
Regression         1  0.1976  0.1976  1.88  0.174
Residual Error    74  7.7697  0.1050
Total             75  7.9673


Unusual Observations

Obs  Carbo  ln rating     Fit  SE Fit  Residual  St Resid
  1    5.0     4.2254  3.5699  0.1010    0.6555    2.13RX
  4    8.0     4.5402  3.6094  0.0750    0.9308    2.95R
 11   12.0     2.8927  3.6619  0.0458   -0.7692   -2.40R
 13   13.0     2.9869  3.6750  0.0410   -0.6882   -2.14R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

We are 95% confident that the population correlation coefficient lies between −0.0703 and 0.3865. As zero is included in this interval, then we conclude that *ln rating* and *carbohydrates* are not linearly correlated. We generalize this interpretation method as follows.

USING A CONFIDENCE INTERVAL TO ASSESS CORRELATION

- If both endpoints of the confidence interval are positive, then we conclude with confidence level $100(1-\alpha)\%$ that x and y are positively correlated.
- If both endpoints of the confidence interval are negative, then we conclude with confidence level $100(1-\alpha)\%$ that x and y are negatively correlated.
- If one endpoint is negative and one endpoint is positive, then we conclude with confidence level $100(1-\alpha)\%$ that x and y are not linearly correlated.

8.14 CONFIDENCE INTERVAL FOR THE MEAN VALUE OF y GIVEN x

Point estimates for values of the response variable for a given value of the predictor value may be obtained by an application of the estimated regression equation $\hat{y} = b_0 + b_1x$. Unfortunately, these kinds of point estimates do not provide a probability statement regarding their accuracy. The analyst is therefore advised to provide for the end-user two types of intervals, which are as follows:

- A confidence interval for the mean value of y given x.
- A prediction interval for the value of a randomly chosen y, given x.

Both of these intervals require that the residuals be normally distributed.

THE CONFIDENCE INTERVAL FOR THE MEAN VALUE OF y FOR A GIVEN VALUE OF x

$$\hat{y}_p \pm t_{n-2}(s)\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

where

x_p is the particular value of x for which the prediction is being made,

$\hat{y}_p$ is the point estimate of y for a particular value of x,

t_{n-2} is a multiplier associated with the sample size and confidence level, and s is the standard error of the estimate.

Before we look at an example of this type of confidence interval, we are first introduced to a new type of interval, the prediction interval.

8.15 PREDICTION I CHOSEN VALUE OF y

Baseball buffs, which is easier
or the batting average of a ran
while perusing the weekly ba
(which each represent the m
team) are more tightly bunche
players themselves. This wou
would be more precise than a
the same confidence level. Th
variable than to predict a rand

For another example of
think it unusual for a randoml
quite remarkable for the class
that the variability associated
associated with an individual
deviation of the univariate ran
the sampling distribution of t
average on an exam is an easi
student.

In many situations, an
value, rather than the mean of
more interested in predicting
than predicting the mean cre
be interested in the expressio
of all similar genes.

Prediction intervals are
of y, given x. Clearly, this is a
in intervals of greater width (
with the same confidence lev

THE PREDICTION INTER y FOR A GIVEN VALUE OF

$\hat{y}_p \pm$

Note that this formula i
interval for the mean value o
the square root. This reflects t
value of y rather than the me
wider than the analogous con

8.15 PREDICTION INTERVAL FOR A RANDOMLY CHOSEN VALUE OF y GIVEN x

Baseball buffs, which is easier to predict: the mean batting average for an entire team, or the batting average of a randomly chosen player? Perhaps, you may have noticed while perusing the weekly batting average statistics that the team batting averages (which each represent the mean batting average of all the players on a particular team) are more tightly bunched together than are the batting averages of the individual players themselves. This would indicate that an estimate of the team batting average would be more precise than an estimate of a randomly chosen baseball player, given the same confidence level. Thus, in general, it is easier to predict the mean value of a variable than to predict a randomly chosen value of that variable.

For another example of this phenomenon, consider exam scores. We would not think it unusual for a randomly chosen student's grade to exceed 98, but it would be quite remarkable for the class mean to exceed 98. Recall from elementary statistics that the variability associated with the mean of a variable is smaller than the variability associated with an individual observation of that variable. For example, the standard deviation of the univariate random variable x is σ, whereas the standard deviation of the sampling distribution of the sample mean $\bar{x}$ is $\sigma/\sqrt{n}$. Hence, predicting the class average on an exam is an easier task than predicting the grade of a randomly selected student.

In many situations, analysts are more interested in predicting an individual value, rather than the mean of all the values, given x. For example, an analyst may be more interested in predicting the credit score for a particular credit applicant, rather than predicting the mean credit score of all similar applicants. Or, a geneticist may be interested in the expression of a particular gene, rather than the mean expression of all similar genes.

Prediction intervals are used to estimate the value of a randomly chosen value of y, given x. Clearly, this is a more difficult task than estimating the mean, resulting in intervals of greater width (lower precision) than confidence intervals for the mean with the same confidence level.

THE PREDICTION INTERVAL FOR A RANDOMLY CHOSEN VALUE OF *y* FOR A GIVEN VALUE OF *x*

$$\hat{y}_p \pm t_{n-2}(s)\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Note that this formula is precisely the same as the formula for the confidence interval for the mean value of y, given x, except for the presence of the "1+" inside the square root. This reflects the greater variability associated with estimating a single value of y rather than the mean; it also ensures that the prediction interval is always wider than the analogous confidence interval.

Recall the orienteering example, where the time and distance traveled was observed for 10 hikers. Suppose we are interested in estimating the distance traveled for a hiker traveling for $y_p = 5$, $x = 5$ hours. The point estimate is easily obtained using the estimated regression equation, from Table 8.6: $\hat{y} = 6 + 2(x) = 6 + 2(5) = 16$. That is, the estimated distance traveled for a hiker walking for 5 hours is 16 kilometers. Note from Figure 8.3 that this prediction $(x = 5, y = 16)$ falls directly on the regression line, as do all such predictions.

However, we must ask the question: How sure are we about the accuracy of our point estimate? That is, are we certain that this hiker will walk precisely 16 kilometers, and not 15.9 or 16.1 kilometers? As usual with point estimates, there is no measure of confidence associated with it, which limits the appicability and usefulness of the point estimate.

We would therefore like to construct a confidence interval. Recall that the regression model assumes that, at each of the x-values, the observed values of y are samples from a normally distributed population with a mean on the regression line $(E(y) = \beta_0 + \beta_1 x)$, and constant variance σ^2, as illustrated in Figure 8.9. The point estimate represents the mean of this population, as estimated by the data.

Now, in this case, of course, we have only observed a single observation with the value $x = 5$ hours. Nevertheless, the regression model assumes the existence of an entire normally distributed population of possible hikers with this value for *time*. Of all possible hikers in this distribution, 95% will travel within a certain bounded distance (the margin of error) from the point estimate of 16 kilometers. We may therefore obtain a 95% confidence interval (or whatever confidence level is desired) for the mean distance traveled by all possible hikers who walked for 5 hours. We use the formula provided above, as follows:

$$\hat{y}_p \pm t_{n-2}(s)\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

where

- $\hat{y}_p = 16$, the point estimate,
- $t_{n-2,\alpha} = t_{=8,95\%} = 2.306$,
- $s = 1.22474$, from Table 8.6,
- $n = 10$,
- $x_p = 5$, and
- $\bar{x} = 5$.

We have $\sum (x_i - \bar{x})^2 = (2-5)^2 + (2-5)^2 + (3-5)^2 + \cdots + (9-5)^2 = 54$, and we therefore calculate the 95% confidence interval as follows:

$$\hat{y}_p \pm t_{n-2}(s)\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$= 16 \pm (2.306)(1.22474)\sqrt{\frac{1}{10} + \frac{(5-5)^2}{54}}$$

$$= 16 \pm 0.893$$
$$= (15.107, 16.893)$$

We are 95% confident that the mean distance traveled by all possible 5-hour hikers lies between 15.107 and 16.893 kilometers.

However, are we sure that this mean of all possible 5-hour hikers is the quantity that we really want to estimate? Wouldn't it be more useful to estimate the distance traveled by a particular randomly selected hiker? Many analysts would agree, and would therefore prefer a prediction interval for a single hiker rather than the above confidence interval for the mean of the hikers.

The calculation of the prediction interval is quite similar to the confidence interval above, but the interpretation is quite different. We have

$$\hat{y}_p \pm t_{n-2}(s)\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$= 16 \pm (2.306)(1.22474)\sqrt{1 + \frac{1}{10} + \frac{(5-5)^2}{54}}$$
$$= 16 \pm 2.962$$
$$= (13.038, 18.962)$$

In other words, we are 95% confident that the distance traveled by a randomly chosen hiker who had walked for 5 hours lies between 13.038 and 18.962 kilometers. Note that, as mentioned earlier, the prediction interval is wider than the confidence interval, because estimating a single response is more difficult than estimating the mean response. However, also note that the interpretation of the prediction interval is probably more useful for the data miner.

We verify our calculations by providing in Table 8.13 the *Minitab* results for the regression of distance on time, with the confidence interval and prediction interval indicated at the bottom ("Predicted Values for New Observations"). The *Fit* of 16 is the point estimate, the standard error of the fit equals $(s)\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$, the 95% CI indicates the confidence interval for the mean distance of all 5-hour hikers, and the 95% PI indicates the prediction interval for the distance traveled by a randomly chosen 5-hour hiker.

8.16 TRANSFORMATIONS TO ACHIEVE LINEARITY

If the normal probability plot shows no systematic deviations from linearity, and the residuals-fits plot shows no discernible patterns, then we may conclude that there is no graphical evidence for the violation of the regression assumptions, and we may then proceed with the regression analysis. However, *what do we do if these graphs*

TABLE 8.13 Regression of distance on time, with confidence interval and prediction interval shown at the bottom

```
The regression equation is
Distance = 6.00 + 2.00 Time


Predictor     Coef   SE Coef       T      P
Constant    6.0000    0.9189    6.53  0.000
Time        2.0000    0.1667   12.00  0.000


S = 1.22474   R-Sq = 94.7%   R-Sq(adj) = 94.1%


Analysis of Variance

Source           DF       SS      MS       F      P
Regression        1   216.00  216.00  144.00  0.000
Residual Error    8    12.00    1.50
Total             9   228.00


Predicted Values for New Observations

New Obs     Fit  SE Fit         95% CI              95% PI
      1  16.000   0.387  (15.107, 16.893)  (13.038, 18.962)
```

indicate violations of the assumptions? For example, suppose our normal probability plot of the residuals looked something such as plot (c) in Figure 8.14, indicating non-constant variance? Then we may apply a transformation to the response variable y, such as the *ln* (natural log, log to the base e) transformation. We illustrate with an example drawn from the world of board games.

Have you ever played the game of Scrabble®? Scrabble is a game in which the players randomly select letters from a pool of letter tiles, and build crosswords. Each letter tile has a certain number of points associated with it. For instance, the letter "E" is worth 1 point, while the letter "Q" is worth 10 points. The point value of a letter tile is roughly related to its letter frequency, the number of times the letter appears in the pool.

Table 8.14 contains the frequency and point value of each letter in the game. Suppose we were interested in approximating the relationship between frequency and point value, using linear regression. As always when performing simple linear regression, the first thing an analyst should do is to construct a scatter plot of the response versus the predictor, in order to see if the relationship between the two variables is indeed linear. Figure 8.18 presents a scatter plot of the point value versus the frequency. Note that each dot may represent more than one letter.

TABLE 8.14 Frequency in of the letters in the alphab

Letter	Frequency in S
A	9
B	2
C	2
D	4
E	12
F	2
G	3
H	2
I	9
J	1
K	1
L	4
M	2
N	6
O	8
P	2
Q	1
R	6
S	4
T	6
U	4
V	2
W	2
X	1
Y	2
Z	1

Perusal of the sca point value and letter f *curvilinear*, in this ca relationship between p such as simple linear r and incorrect inference linearity in the relation

Frederick, Moste suggest "the bulging ru stand the bulging rule f Tukey[4]).

Compare the cur Figure 8.19. It is most

TABLE 8.14 Frequency in Scrabble®, and Scrabble® point value of the letters in the alphabet

Letter	Frequency in Scrabble®	Point Value in Scrabble®
A	9	1
B	2	3
C	2	3
D	4	2
E	12	1
F	2	4
G	3	2
H	2	4
I	9	1
J	1	8
K	1	5
L	4	1
M	2	3
N	6	1
O	8	1
P	2	3
Q	1	10
R	6	1
S	4	1
T	6	1
U	4	1
V	2	4
W	2	4
X	1	8
Y	2	4
Z	1	10

Perusal of the scatter plot indicates clearly that there is a relationship between point value and letter frequency. However, the relationship is not linear, but rather *curvilinear*, in this case quadratic. It would not be appropriate to model the relationship between point value and letter frequency using a linear approximation such as simple linear regression. Such a model would lead to erroneous estimates and incorrect inference. Instead, the analyst may apply a transformation to achieve linearity in the relationship.

Frederick, Mosteller, and Tukey, in their book *Data Analysis and Regression*[4], suggest "the bulging rule" for finding transformations to achieve linearity. To understand the bulging rule for quadratic curves, consider Figure 8.19 (after Mosteller and Tukey[4]).

Compare the curve seen in our scatter plot, Figure 8.18, to the curves shown in Figure 8.19. It is most similar to the curve in the lower left quadrant, the one labeled

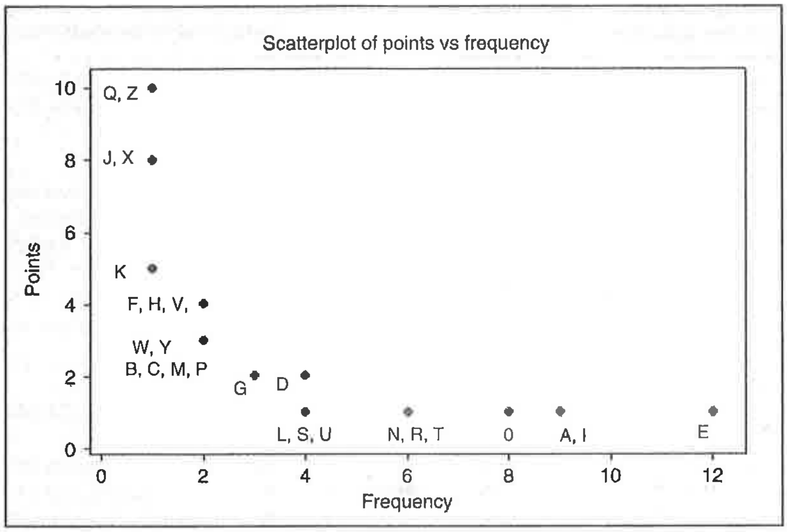


Figure 8.18 Scatter plot of *points* versus *frequency* in Scrabble®: nonlinear!

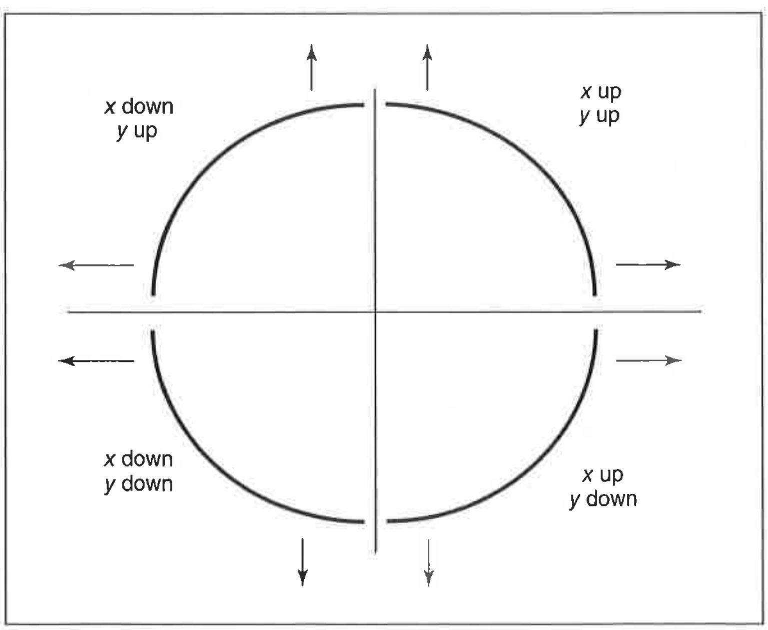


Figure 8.19 The bulging rule: a heuristic for variable transformation to achieve linearity.

"x down, y down." Mos
are essentially a set of]

LADDER OF RE-EXI

The ladder of re-expres
any continuous variable

t^{-}

For our curve, th
means that we should
from x's present positic
for y. The present posi
rule suggests that we a
transformation to both l
relationship between th

Thus, we apply t
and consider the scatter
Unfortunately, the grap
frequency is still not li
regression. Evidently, tl
in this case.

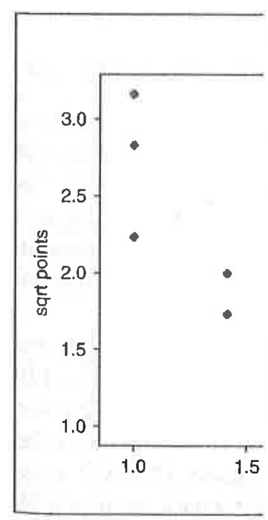


Figure 8.20 After applyi

[7]Mosteller and Tukey, *Data A*

"x down, y down." Mosteller and Tukey[7] propose a "ladder of re-expressions," which are essentially a set of power transformations, with one exception, ln(t).

LADDER OF RE-EXPRESSIONS (MOSTELLER AND TUKEY)

The ladder of re-expressions consists of the following ordered set of transformations for any continuous variable t.

$$t^{-3} \quad t^{-2} \quad t^{-1} \quad t^{-1/2} \quad \ln(t) \quad \sqrt{t} \quad t^{1} \quad t^{2} \quad t^{3}$$

For our curve, the heuristic from the bulging rule is "x down, y down." This means that we should transform the variable x, by going down one or more spots from x's present position on the ladder. Similarly, the same transformation is made for y. The present position for all untransformed variables is t^1. Thus, the bulging rule suggests that we apply either the square root transformation or the natural log transformation to both letter tile frequency and point value, in order to achieve a linear relationship between the two variables.

Thus, we apply the square root transformation to both *frequency* and *points*, and consider the scatter plot of *sqrt points* versus *sqrt frequency*, given in Figure 8.20. Unfortunately, the graph indicates that the relationship between sqrt points and sqrt frequency is still not linear, so that it would still be inappropriate to apply linear regression. Evidently, the square root transformation was too mild to effect linearity in this case.

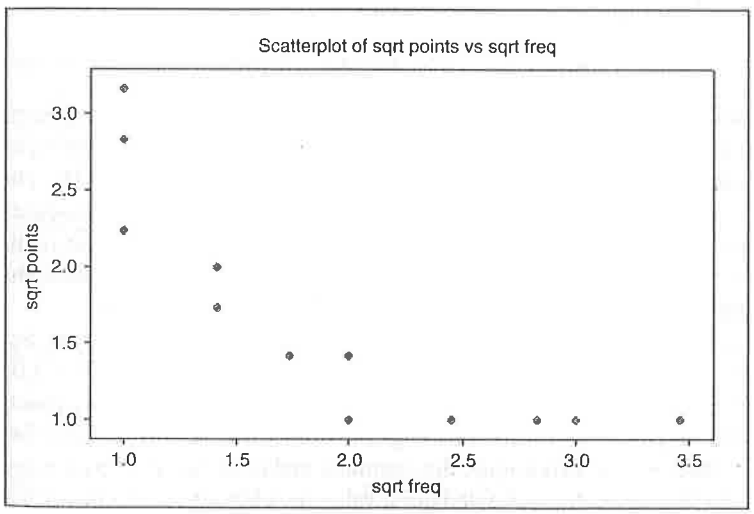


Figure 8.20 After applying square root transformation, still not linear.

[7]Mosteller and Tukey, *Data Analysis and Regression*, Addison-Wesley, Reading, MA, 1977.

We therefore move one more notch down the ladder of re-expressions, and apply the natural log transformation to each of frequency and point value, generating the transformed variables ln points and ln frequency. The scatter plot of *ln points* versus *ln frequency* is shown in Figure 8.21. This scatter plot exhibits acceptable linearity, although, as with any real-world scatter plot, the linearity is imperfect. We may therefore proceed with the regression analysis for *ln points* and *ln frequency*.

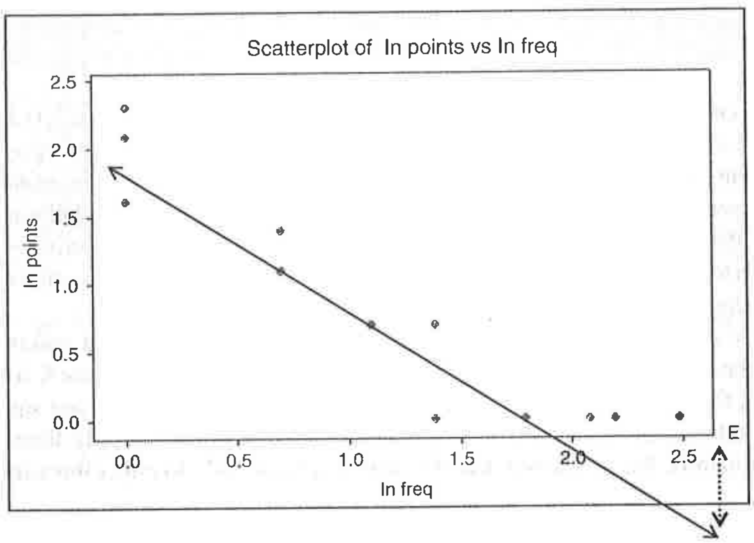


Figure 8.21 The natural log transformation has achieved acceptable linearity (single outlier, *E*, indicated).

Table 8.15 presents the results from the regression of *ln points* on *ln frequency*. Let us compare these results with the results from the inappropriate regression of points on frequency, with neither variable transformed, shown in Table 8.16. The coefficient of determination for the untransformed case is only 45.5%, as compared to 87.6% for the transformed case, meaning that, the transformed predictor accounts for nearly twice as much of the variability in the transformed response than in the case for the untransformed variables.

We can also compare the predicted point value for a given frequency, say frequency = 4 tiles. For the proper regression, the estimated *ln points* equals 1.94 − 1.01 (*ln freq*) = 1.94 − 1.01 (1.386) = 0.5401, giving us an estimated $e^{0.5401} = 1.72$ points for a letter with frequency 4. As the actual point values for letters with this frequency are all either one or two points, this estimate makes sense. However, using the untransformed variables, the estimated point value for a letter with frequency 4 is 5.73 − 0.633 (frequency) = 5.73 − 0.633 (4) = 3.198, which is much larger than any of the actual point values for letter with frequency 4. This exemplifies the danger of applying predictions from inappropriate models.

TABLE 8.15 Regression of

```
The regression equa
ln points = 1.94 -

Predictor      Coe
Constant     1.9403
ln freq     -1.0053

S = 0.293745   R-S

Analysis of Varian

Source            DF
Regression         1
Residual Error    24
Total             25

Unusual Observatio

Obs  ln freq  ln p
  5     2.48     0

R denotes an obser
```

TABLE 8.16 Inappropriate

```
The regression equ
Points = 5.73 - 0.

Predictor     Coef
Constant    5.7322
Frequency  -0.6330

S = 2.10827   R-Sq

Analysis of Varian

Source           DF
Regression        1
Residual Error   24
Total            25
```

TABLE 8.15 **Regression of *ln points* on *ln frequency***

```
The regression equation is
ln points = 1.94 - 1.01 ln freq


Predictor       Coef  SE Coef       T      P
Constant     1.94031  0.09916   19.57  0.000
ln freq     -1.00537  0.07710  -13.04  0.000


S = 0.293745   R-Sq = 87.6%   R-Sq(adj) = 87.1%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  14.671  14.671  170.03  0.000
Residual Error  24   2.071   0.086
Total           25  16.742


Unusual Observations

Obs  ln freq  ln points      Fit  SE Fit  Residual  St Resid
  5     2.48     0.0000  -0.5579  0.1250    0.5579     2.10R

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

TABLE 8.16 **Inappropriate regression of *points* on *frequency***

```
The regression equation is
Points = 5.73 - 0.633 Frequency


Predictor     Coef  SE Coef      T      P
Constant    5.7322   0.6743   8.50  0.000
Frequency  -0.6330   0.1413  -4.48  0.000


S = 2.10827   R-Sq = 45.5%   R-Sq(adj) = 43.3%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       1   89.209  89.209  20.07  0.000
Residual Error  24  106.676   4.445
Total           25  195.885
```

In Figure 8.21 and Table 8.15, there is a single outlier, the letter "E." As the standardized residual is positive, this indicates that the point value for E is higher than expected, given its frequency, which is the highest in the bunch, 12. The residual of 0.5579 is indicated by the dashed vertical line in Figure 8.21. The letter "E" is also the only "influential" observation, with a Cook's distance of 0.5081 (not shown), which just exceeds the 50th percentile of the $F_{1,25}$ distribution.

8.17 BOX–COX TRANSFORMATIONS

Generalizing from the idea of a ladder of transformations, to admit powers of any continuous value, we may apply a Box–Cox transformation.[8] A Box–Cox transformation is of the form:

$$W = \begin{cases} \left(y^{\lambda} - 1\right)/\lambda, & \text{for } \lambda \neq 0, \\ \ln y, & \text{for } \lambda = 0 \end{cases}$$

For example, we could have $\lambda = 0.75$, giving us the following transformation, $W = (y^{0.75} - 1)/0.75$. Draper and Smith[9] provide a method of using maximum likelihood to choose the optimal value of λ. This method involves first choosing a set of candidate values for λ, and finding SSE for regressions performed using each value of λ. Then, plotting SSE_{λ} versus λ, find the lowest point of a curve through the points in the plot. This represents the maximum-likelihood estimate of λ.

THE R ZONE

Read in and prepare Cereals data

```
cereal <- read.csv(file = "C:/ … /cereals.txt",
    stringsAsFactors=TRUE, header=TRUE, sep="\t")
# Save Rating and Sugar as new variables
sugars <- cereal$Sugars; rating <- cereal$Rating
which(is.na(sugars)) # Record 58 is missing
sugars <- na.omit(sugars) # Delete missing value
rating <- rating[-58] # Delete Record 58 from Rating to match
```

[8] Box and Cox, An Analysis of Transformations, *Journal of the Royal Statistical Society, Series B*, Volume **26**, pages 2211—243, 1964. (This formula above is valid only for $y > 0$.)

[9] Draper and Smith, *Applied Regression Analysis*, 3rd edition, Wiley Publishers, Hoboken, New Jersey, 1998.

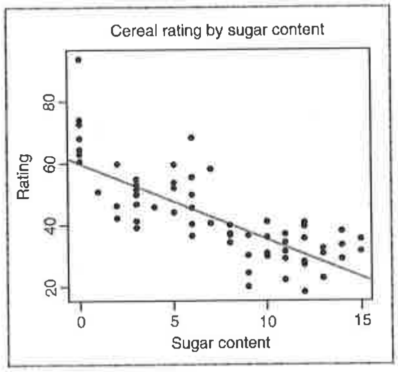
Run regression analysis

```
lm1 <-
    lm(rating~sugars)
# Display summaries
summary(lm1)
anova(lm1)
```

```
> summary(lm1)

Call:
lm(formula = rating ~ sugars)

Residuals:
    Min      1Q  Median      3Q     Max
-17.877  -5.612  -1.285   4.689  33.852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.8530     1.9975   29.96  < 2e-16 ***
sugars       -2.4614     0.2417  -10.18 1.01e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.166 on 74 degrees of freedom
Multiple R-squared: 0.5835, Adjusted R-squared: 0.5779
F-statistic: 103.7 on 1 and 74 DF,  p-value: 1.006e-15

> anova(lm1)
Analysis of Variance Table

Response: rating
          Df Sum Sq Mean Sq F value    Pr(>F)
sugars     1 8711.9  8711.9  103.69 1.006e-15 ***
Residuals 74 6217.4    84.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Plot data with regression line

```
plot(sugars, rating,
    main = "Cereal Rating by Sugar Content",
    xlab = "Sugar Content", ylab = "Rating",
    pch = 16, col = "blue")
abline(lm1, col = "red")
```



Residuals, r^2, standardized residuals, leverage

```
lm1$residuals # All residuals
lm1$residuals[12] # Residual of Cheerios, Record 12
a1 <- anova(lm1)
# Calculate r^2
r2.1 <- a1$"Sum Sq"[1] / (a1$"Sum Sq"[1] +
    a1$"Sum Sq"[2])
std.res1 <- rstandard(lm1) # Standardized residuals
lev <- hatvalues(lm1) # Leverage
```

```
> lm1$residuals[12]
       12
-6.626598
> r2.1
[1] 0.5835462
```

Orienteering example

```
# Input the data
x <- c(2, … ,9)
y <- c(10, … , 25)
o.data <- data.frame(cbind(
    "Time" = x,
    "Distance" = y))
lm2 <- lm(Distance ~
    Time, data = o.data)
a2 <- anova(lm2)
# Directly calculate r^2
r2.2 <- a2$"Sum Sq"[1] /
    (a2$"Sum Sq"[1] +
    a2$"Sum Sq"[2])
# MSE
mse <- a2$"Mean Sq"[2]
s <- sqrt(mse) # s
# Std dev of Y
sd(o.data$Distance)
r <- sign(lm2$coefficients[2])* sqrt(r2.2) # r
```

```
> summary(lm2)

Call:
lm(formula = Distance ~ Time, data = o.data)

Residuals:
   Min     1Q Median     3Q    Max
 -2.00  -0.75   0.00   0.75   2.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.0000     0.9189   6.529 0.000182 ***
Time          2.0000     0.1667  12.000 2.14e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 1.225 on 8 degrees of freedom
Multiple R-squared: 0.9474, Adjusted R-squared: 0.9408
F-statistic:   144 on 1 and 8 DF,  p-value: 2.144e-06

> a2
Analysis of Variance Table

Response: Distance
          Df Sum Sq Mean Sq F value    Pr(>F)
Time       1    216   216.0     144 2.144e-06 ***
Residuals  8     12     1.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
> sd(o.data$Distance)
[1] 5.033223
> r2.2
[1] 0.9473684
> mse
[1] 1.5
> s
[1] 1.224745
> r
     Time
0.9733285
```

Regression using other hikers

```
# Hard-core hiker
hardcore <- cbind("Time" = 16,
    "Distance" = 39)
o.data <- rbind(o.data, hardcore)
lm3 <- lm(Distance ~ Time,
    data = o.data)
summary(lm3); anova(lm3)
hatvalues(lm3)
# Leverage
rstandard(lm3)
# Standardized residual
cooks.distance(lm3)
# Cook's Distance
# 5-hour, 20-km hiker
o.data[11,] <- cbind("Time" = 5, "Distance" = 20)
lm4 <- lm(Distance ~ Time, data = o.data)
summary(lm4); anova(lm4); rstandard(lm4) ;
hatvalues(lm4) ; cooks.distance(lm4)
# 10-hour, 23-km hiker
o.data[11,] <- cbind("Time" = 10, "Distance" = 23)
lm5 <- lm(Distance ~ Time, data = o.data)
summary(lm5); anova(lm5); hatvalues(lm5);
rstandard(lm5); cooks.distance(lm5)
```

L

```
> summary(lm3)

Call:
lm(formula = Distance ~ Time, data = o.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1786 -0.4286  0.1421  0.3044  1.8931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.67666    0.57317   9.904 1.74e-06
Time         2.07171    0.06951  29.806 4.23e-11


> anova(lm3)
Analysis of Variance Table

Response: Distance
          Df  Sum Sq Mean Sq F value    Pr(>F)
Time       1 1097.31 1097.31  888.37 4.225e-11
Residuals 10   12.35    1.24

> hatvalues(lm3)
         1          2          3          4
0.17470665 0.17470665 0.14080834 0.11473272
         5          6          7          8
0.11473272 0.09647979 0.08604954 0.08344198
         9         10         11         12
0.08865711 0.10169492 0.41199478 0.41199478
> rstandard(lm3)
          1           2           3           4
 0.17820117  1.16863808  0.10504338 -0.92138866
          5           6           7           8
 0.03491053 -0.97991227  1.78172600 -2.04753860
          9          10          11          12
-0.23593694  0.64361631  0.20652794  0.20652794
> cooks.distance(lm3)
           1            2            3
3.361183e-03 1.445543e-01 9.041609e-04
           4            5            6
5.501342e-02 7.897612e-05 5.126759e-02
           7            8            9
1.494437e-01 1.908354e-01 2.707657e-03
          10           11           12
2.344766e-02 1.494301e-02 1.494301e-02
```

Verify the assumpt

```
par(mfrow=c(2,2)); plot(lm2)
# Normal probability plot: top
# Residuals vs Fitted: top-left
# Square root of absolute valu
# of standardized residuals:
# bottom-left
# Reset the plot space
par(mfrow=c(1,1))
```

Plot Standardized

```
plot(lm2$fitted.values, rstanda
    pch = 16, col = "red",
    main = "Standardized
        Residuals by Fitted V
    ylab = "Standardized Re
    xlab = "Fitted Values")
abline(0,0)
```

Check residuals ar

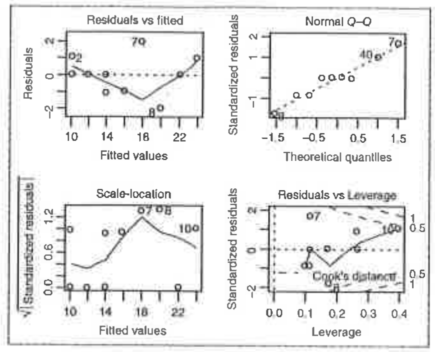
```
# Normal Q-Q Plot
qqnorm(lm1$residuals, datax =
qqline(lm1$residuals, datax =
# Anderson-Darling test
# Requires "nortest" package
library("nortest")
ad.test(lm1$residuals)
```

```
)istance ~ Time, data = o.data)

Median      3Q     Max
  0.00    0.75    2.00

stimate Std. Error t value Pr(>|t|)
 6.0000     0.9189   6.529 0.000182 ***
 2.0000     0.1667  12.000 2.14e-06 ***

:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

dard error: 1.225 on 8 degrees of freedom
uared: 0.9474, Adjusted R-squared: 0.9408
 144 on 1 and 8 DF,  p-value: 2.144e-06

ariance Table

tance
Sum Sq Mean Sq F value    Pr(>F)
   216   216.0     144 2.144e-06 ***
    12     1.5

:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

istance)
```

```
(lm3)

 = Distance ~ Time, data = o.data)

   1Q  Median      3Q     Max
).4286  0.1421  0.3044  1.8931

nts:
   Estimate Std. Error t value Pr(>|t|)
:)  5.67666    0.57317   9.904 1.74e-06
    2.07171    0.06951  29.806 4.23e-11

lm3)
 of Variance Table

: Distance
  Df  Sum Sq Mean Sq F value    Pr(>F)
   1 1097.31 1097.31  888.37 4.225e-11
; 10   12.35    1.24
ies(lm3)
1            2            3            4
i5 0.17470665 0.14080834 0.11473272
5            6            7            8
'2 0.09647979 0.08604954 0.08344198
9           10           11           12
.1 0.10169492 0.41199478 0.41199478
ird(lm3)
 1             2            3            4
.17  1.16863808  0.10504338 -0.92138866
 5             6            7            8
)53 -0.97991227  1.78172600 -2.04753860
 9            10           11           12
594  0.64361631  0.20652794  0.20652794
distance(lm3)
  1            2            3
e-03 1.445543e-01 9.041609e-04
  4            5            6
e-02 7.897612e-05 5.126759e-02
  7            8            9
e-01 1.908354e-01 2.707657e-03
 10           11           12
e-02 1.494301e-02 1.494301e-02
```

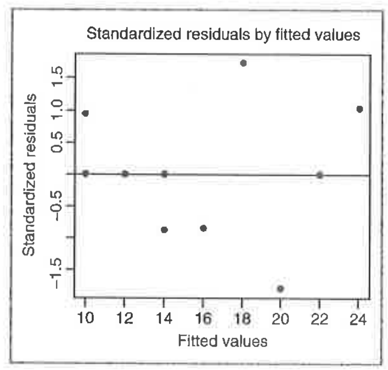

Verify the assumptions

```
par(mfrow=c(2,2)); plot(lm2)
# Normal probability plot: top-right
# Residuals vs Fitted: top-left
# Square root of absolute value
# of standardized residuals:
# bottom-left
# Reset the plot space
par(mfrow=c(1,1))
```



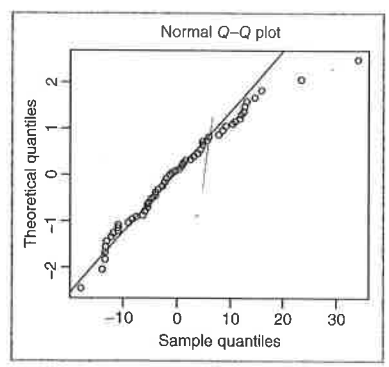
Plot Standardized residuals by fitted values

```
plot(lm2$fitted.values, rstandard(lm2),
     pch = 16, col = "red",
     main = "Standardized
          Residuals by Fitted Values",
     ylab  = "Standardized Residuals",
     xlab = "Fitted Values")
abline(0,0)
```



Check residuals are Normally distributed

```
# Normal Q-Q Plot
qqnorm(lm1$residuals, datax = TRUE)
qqline(lm1$residuals, datax = TRUE)
# Anderson-Darling test
# Requires "nortest" package
library("nortest")
ad.test(lm1$residuals)
```



t-test

```
summary(lm1)
# t-test is in the 'sugars' row
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.8530     1.9975   29.96  < 2e-16
sugars       -2.4614     0.2417  -10.18 1.01e-15
```

CI for Beta coefficients

```
confint(lm1, level = 0.95)
```

```
> confint(lm1, level = 0.95)
                 2.5 %    97.5 %
(Intercept) 55.872858 63.833176
sugars      -2.943061 -1.979779
```

Regression for Carbohydrates and Natural Log of Rating

```
carbs <- cereal$"Carbo"[-58]
lrating <- log(rating)
ad.test(lrating); ad.test(carbs)
lm6 <- lm(lrating~carbs)
summary(lm6)
a6 <- anova(lm6); a6
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.504260   0.146539  23.913   <2e-16
carbs       0.013137   0.009576   1.372    0.174

  Analysis of Variance Table

  Response: lrating
            Df Sum Sq Mean Sq F value Pr(>F)
  carbs      1 0.1976 0.19761  1.8821 0.1742
  Residuals 74 7.7697 0.10500
```

CI for r

```
alpha <- 0.05
n <- length(lrating)
r2.6 <- a6$"Sum Sq"[1] / (a6$"Sum Sq"[1] +
     a6$"Sum Sq"[2])
r <- sign(lm6$coefficients[2])*sqrt(r2.6)
sr <- sqrt((1-r^2)/(n-2))
lb <- r - qt(p=alpha/2, df = n-2, lower.tail = FALSE)*sr
ub <- r + qt(p=alpha/2, df = n-2, lower.tail = FALSE)*sr
lb;ub
```

```
> lb;ub
      carbs
-0.07124931
      carbs
0.3862266
```

Confidence and Predict

```
newdata <- data.frame(cbind(Distanc
conf.int <- predict(lm2, newdata,
     interval = "confidence")
pred.int <- predict(lm2, newdata,
     interval = "prediction")
conf.int; pred.int
```

Assess Normality in Scr

```
# Scrabble data
s.freq <- c(9, … 1); s.point <- c(1,
scrabble <- data.frame("Frequency"
     "Points" = s.point)
plot(scrabble,
     main = "Scrabble Points vs Freq
     xlab = "Frequency", ylab = "Poi
     col = "red", pch = 16,
     xlim = c(0, 13), ylim = c(0,10))
sq.scrabble <- sqrt(scrabble)
plot(sq.scrabble,
     main = "Square Root of Scrabbl
     vs Frequency",
     xlab = "Sqrt Frequency", ylab =
     Points", col = "red", pch = 16)
ln.scrabble <- log(scrabble)
plot(ln.scrabble, main = "Natural Lo
          Scrabble Points vs Frequenc
     xlab = "Ln Frequency", ylab = '
     Points", col = "red", pch = 16)
```

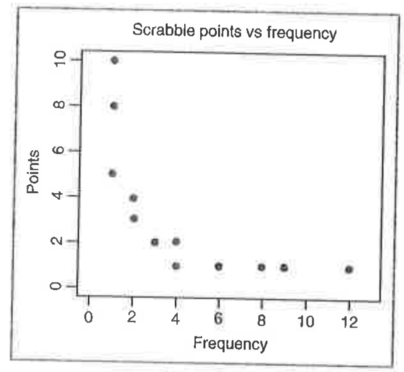
Confidence and Prediction Intervals

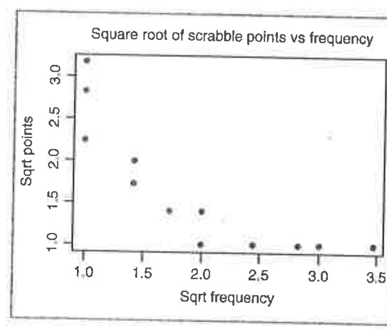
```
newdata <- data.frame(cbind(Distance = 5, Time = 5))
conf.int <- predict(lm2, newdata,
    interval = "confidence")
pred.int <- predict(lm2, newdata,
    interval = "prediction")
conf.int; pred.int
```

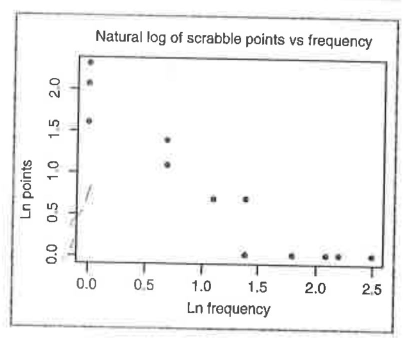
```
> conf.int
  fit      lwr      upr
1  16 15.10689 16.89311
> pred.int
  fit      lwr      upr
1  16 13.03788 18.96212
```

Assess Normality in Scrabble example

```
# Scrabble data
s.freq <- c(9, … 1); s.point <- c(1, … 10)
scrabble <- data.frame("Frequency" = s.freq,
    "Points" = s.point)
plot(scrabble,
    main = "Scrabble Points vs Frequency",
    xlab = "Frequency", ylab = "Points",
    col = "red", pch = 16,
    xlim = c(0, 13), ylim = c(0,10))
sq.scrabble <- sqrt(scrabble)
plot(sq.scrabble,
    main = "Square Root of Scrabble Points
    vs Frequency",
    xlab = "Sqrt Frequency", ylab = "Sqrt
    Points", col = "red", pch = 16)
ln.scrabble <- log(scrabble)
plot(ln.scrabble, main = "Natural Log of
        Scrabble Points vs Frequency",
    xlab = "Ln Frequency", ylab = "Ln
    Points", col = "red", pch = 16)
```







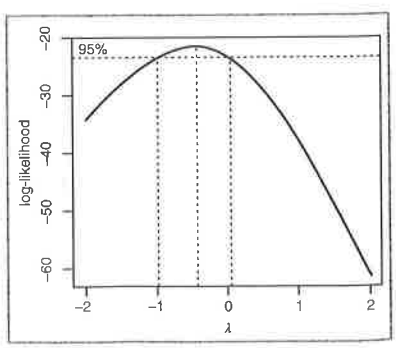
Run regression on Scrabble data, transformed and untransformed

```
lm7 <- lm(Points ~
    Frequency,
    data = ln.scrabble)
summary(lm7)
anova(lm7)
rstandard(lm7)
lm8 <- lm(Points ~
    Frequency,
    data = scrabble)
summary(lm8)
anova(lm8)
```

```
> summary(lm7)

Call:
lm(formula = Points ~ Frequency, data = ln.scrabble)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5466 -0.1448  0.1391  0.1457  0.5579

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.94031    0.09916   19.57 2.94e-16 ***
Frequency   -1.00537    0.07710  -13.04 2.20e-12 ***

> anova(lm7)
Analysis of Variance Table

Response: Points
          Df  Sum Sq Mean Sq F value    Pr(>F)
Frequency  1 14.6711 14.6711  170.03 2.197e-12 ***
Residuals 24  2.0709  0.0863

> summary(lm8)

Call:
lm(formula = Points ~ Frequency, data = scrabble)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2001 -1.4661 -0.4661  0.8068  4.9008

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.7322     0.6743   8.502 1.06e-08 ***
Frequency    -0.6330     0.1413  -4.480 0.000156 ***

> anova(lm8)
Analysis of Variance Table

Response: Points
          Df  Sum Sq Mean Sq F value    Pr(>F)
Frequency  1  89.209  89.209   20.07 0.0001558 ***
Residuals 24 106.676   4.445
```

Box-Cox Transformation

```
# Requires MASS package
library(MASS)
bc <- boxcox(lm8)
```



R REFERENCES

Juergen Gross and bug fixes
version 1.0-2. http://CRAN
R Core Team. *R: A Langua*
tria: R Foundation for Sta
.R-project.org/.
Venables WN, Ripley BD. *M*
2002. ISBN: 0-387-95457-

EXERCISES

CLARIFYING THE C

1. Indicate whether the follo
make it true.
 a. The least-squares line
 b. If all the residuals equ
 c. If the value of the co
 are negatively correla
 d. The value of the corre
 e. Outliers are influentia
 f. If the residual for an o
 than the regression es
 g. An observation may b
 age point.
 h. The best way of dete
 its Cook's distance ex
 i. If one is interested in
 no inference and no
 assumption validation
 j. In a normality plot, if
 on a straight line.
 k. The chi-square distrib
 l. Small *p*-values for t
 right-skewed.
 m. A funnel pattern in th
 pendence assumption.
2. Describe the difference be
3. Calculate the estimated r
in Table 8.3. Use either th
4. Where would a data point

R REFERENCES

Juergen Gross and bug fixes by Uwe Ligges. 2012. nortest: Tests for normality. R package version 1.0-2. http://CRAN.R-project.org/package=nortest.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2012. ISBN: 3-900051-07-0, http://www.R-project.org/.

Venables WN, Ripley BD. *Modern Applied Statistics with S*. Fourth ed. New York: Springer; 2002. ISBN: 0-387-95457-0.

EXERCISES

CLARIFYING THE CONCEPTS

1. Indicate whether the following statements are true or false. If false, alter the statement to make it true.
   a. The least-squares line is that line that minimizes the sum of the residuals.
   b. If all the residuals equal zero, then SST = SSR.
   c. If the value of the correlation coefficient is negative, this indicates that the variables are negatively correlated.
   d. The value of the correlation coefficient can be calculated, given the value of r^2 alone.
   e. Outliers are influential observations.
   f. If the residual for an outlier is positive, we may say that the observed y-value is higher than the regression estimated, given the x-value.
   g. An observation may be influential even though it is neither an outlier nor a high leverage point.
   h. The best way of determining whether an observation is influential is to see whether its Cook's distance exceeds 1.0.
   i. If one is interested in using regression analysis in a strictly descriptive manner, with no inference and no model building, then one need not worry quite so much about assumption validation.
   j. In a normality plot, if the distribution is normal, then the bulk of the points should fall on a straight line.
   k. The chi-square distribution is left-skewed.
   l. Small p-values for the Anderson–Darling test statistic indicate that the data are right-skewed.
   m. A funnel pattern in the plot of residuals versus fits indicates a violation of the independence assumption.
2. Describe the difference between the estimated regression line and the true regression line.
3. Calculate the estimated regression equation for the orienteering example, using the data in Table 8.3. Use either the formulas or software of your choice.
4. Where would a data point be situated that has the smallest possible leverage?

5. Calculate the values for leverage, standardized residual, and Cook's distance for the hard-core hiker example in the text.

6. Calculate the values for leverage, standardized residual, and Cook's distance for the 11th hiker who had hiked for 10 hours and traveled 23 kilometers. Show that, while it is neither an outlier nor of high leverage, it is nevertheless influential.

7. Match each of the following regression terms with its definition.

Regression Term	Definition
a. Influential observation	Measures the typical difference between the predicted response value and the actual response value.
b. SSE	Represents the total variability in the values of the response variable alone, without reference to the predictor.
c. r^2	An observation that has a very large standardized residual in absolute value.
d. Residual	Measures the strength of the linear relationship between two quantitative variables, with values ranging from -1 to 1.
e. s	An observation that significantly alters the regression parameters based on its presence or absence in the data set.
f. High leverage point	Measures the level of influence of an observation, by taking into account both the size of the residual and the amount of leverage for that observation.
g. r	Represents an overall measure of the error in prediction resulting from the use of the estimated regression equation.
h. SST	An observation that is extreme in the predictor space, without reference to the response variable.
i. Outlier	Measures the overall improvement in prediction accuracy when using the regression as opposed to ignoring the predictor information.
j. SSR	The vertical distance between the predicted response and the actual response.
k. Cook's distance	The proportion of the variability in the response that is explained by the linear relationship between the predictor and response variables.

8. Explain in your own words the implications of the regression assumptions for the behavior of the response variable y.

9. Explain what statistics from Table 8.11 indicate to us that there may indeed be a linear relationship between x and y in this example, even though the value for r^2 is less than 1%.

10. Which values of the slope parameter indicate that no linear relationship exist between the predictor and response variables? Explain how this works.

11. Explain what informati
mate.

12. Describe the criterion
for hypothesis testing.
situation (one p-value a
to two different conclu

13. (a) Explain why an ar
Describe how a confide

14. Explain the difference
interval is always wide
more useful to the data

15. Clearly explain the cor
of the residuals versus

16. What recourse do we ha
have been violated? De
will help us.

17. A colleague would like
make a purchase, base
colleague?

WORKING WITH

For Exercises 18–23, refer
percentage of the home tea

18. Describe any correlatio

19. Estimate as best you ca

20. Will the p-value for th
the variables be small

21. Will the confidence int

22. Will the value of s be

23. Is there an observation

For Exercises 24 and 25, us

24. Is it appropriate to per

25. What type of transforr

For Exercises 26–30, use
(from the *Churn* data set) i

26. Is there evidence of a l
ber of voice mail mess
Explain.

27. Use the data in the AN

11. Explain what information is conveyed by the value of the standard error of the slope estimate.

12. Describe the criterion for rejecting the null hypothesis when using the p-value method for hypothesis testing. Who chooses the value of the level of significance, α? Make up a situation (one p-value and two different values of α) where the very same data could lead to two different conclusions of the hypothesis test. Comment.

13. (a) Explain why an analyst may prefer a confidence interval to a hypothesis test. (b) Describe how a confidence interval may be used to assess significance.

14. Explain the difference between a confidence interval and a prediction interval. Which interval is always wider? Why? Which interval is probably, depending on the situation, more useful to the data miner? Why?

15. Clearly explain the correspondence between an original scatter plot of the data and a plot of the residuals versus fitted values.

16. What recourse do we have if the residual analysis indicates that the regression assumptions have been violated? Describe three different rules, heuristics, or family of functions that will help us.

17. A colleague would like to use linear regression to predict whether or not customers will make a purchase, based on some predictor variable. What would you explain to your colleague?

WORKING WITH THE DATA

For Exercises 18–23, refer to the scatterplot of attendance at football games versus winning percentage of the home team in Figure 8.22.

18. Describe any correlation between the variables. Interpret this correlation.

19. Estimate as best you can the values of the regression coefficients b_0 and b_1.

20. Will the p-value for the hypothesis test for the existence of a linear relationship between the variables be small or large? Explain.

21. Will the confidence interval for the slope parameter include zero or not? Explain.

22. Will the value of s be closer to 10, 100, 1000, or 10,000? Why?

23. Is there an observation that may look as though it is an outlier? Explain.

For Exercises 24 and 25, use the scatter plot in Figure 8.23 to answer the questions.

24. Is it appropriate to perform linear regression? Why or why not?

25. What type of transformation or transformations is called for? Use the bulging rule.

For Exercises 26–30, use the output from the regression of *z mail messages* on *z day calls* (from the *Churn* data set) in Table 8.17 to answer the questions.

26. Is there evidence of a linear relationship between *z vmail messages* (z-scores of the number of voice mail messages) and *z day calls* (z-scores of the number of day calls made)? Explain.

27. Use the data in the ANOVA table to find or calculate the following quantities:

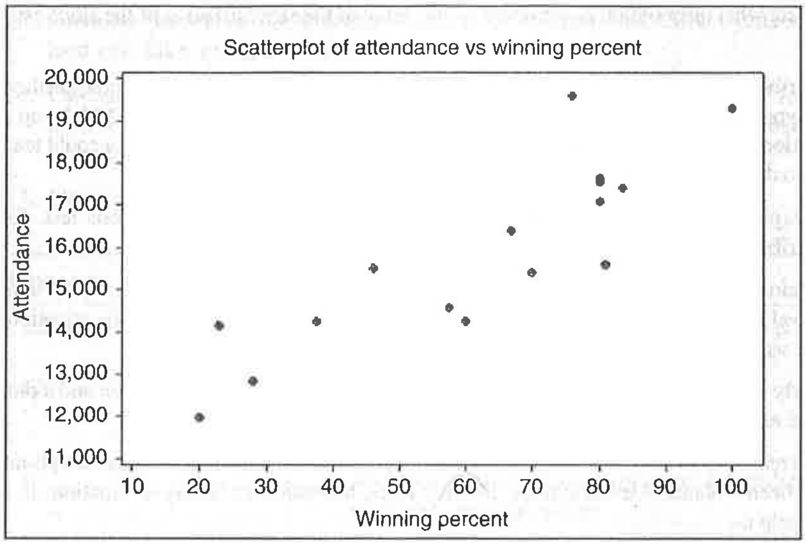


Figure 8.22 Scatter plot of attendance versus winning percentage.

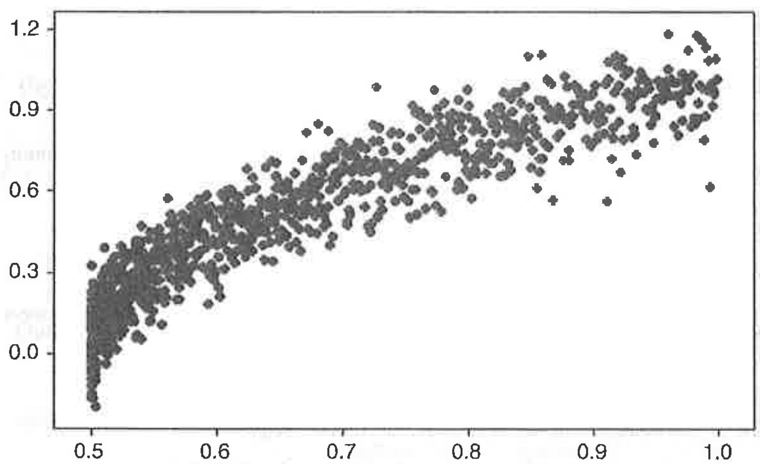


Figure 8.23 Scatter plot.

a. SSE, SSR, and SST.

b. Coefficient of determination, using the quantities in (a). Compare with the number reported by Minitab.

c. Correlation coefficient r.

d. Use SSE and the residual error degrees of freedom to calculate s, the standard error of the estimate. Interpret this value.

28. Assuming normality, construct and interpret a 95% confidence interval for the population correlation coefficient.

TABLE 8.17 Regression of *z vma*

```
The regression equation
zvmail messages = 0.0000

Predictor        Coef   S
Constant      0.00000   0
z day calls  -0.00955   0

S = 1.00010   R-Sq = 0.0

Analysis of Variance

Source            DF
Regression         1
Residual Error  3331   33
Total           3332   33
```

TABLE 8.18 Regression of an un

```
The regression equation
Y = 0.783 + 0.0559 X

Predictor      Coef   SE C
Constant    0.78262   0.03
Y           0.05594   0.03

S = 0.983986   R-Sq = 0.
```

29. Discuss the usefulness of the

30. As it has been standardized, 1.0. What would be the typic sample mean response and n is the typical error in predicti

For Exercises 31–38, use the outp x in Table 8.18 to answer the ques

31. Carefully state the regressior

32. Interpret the value of the y-ir

33. Interpret the value of the slo

34. Interpret the value of the star

TABLE 8.17 Regression of *z vmail messages* on *z day calls*

The regression equation is
zvmail messages = 0.0000 - 0.0095 z day calls

Predictor	Coef	SECoef	T	P
Constant	0.00000	0.01732	0.00	1.000
z day calls	-0.00955	0.01733	-0.55	0.582

S = 1.00010 R-Sq = 0.0% R-Sq(adj) = 0.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.304	0.304	0.30	0.582
Residual Error	3331	3331.693	1.000		
Total	3332	3331.997			

TABLE 8.18 Regression of an unspecified *y* on an unspecified *x*

The regression equation is
Y = 0.783 + 0.0559 X

Predictor	Coef	SE Coef	T	P
Constant	0.78262	0.03791	20.64	0.000
Y	0.05594	0.03056	1.83	0.067

S = 0.983986 R-Sq = 0.3% R-Sq(adj) = 0.2%

29. Discuss the usefulness of the regression of *z mail messages* on *z day calls*.

30. As it has been standardized, the response *z vmail messages* has a standard deviation of 1.0. What would be the typical error in predicting *z vmail messages* if we simply used the sample mean response and no information about day calls? Now, from the printout, what is the typical error in predicting *z vmail messages*, given *z day calls*? Comment.

For Exercises 31–38, use the output from the regression of an unspecified y on an unspecified x in Table 8.18 to answer the questions.

31. Carefully state the regression equation, using words and numbers.

32. Interpret the value of the y-intercept b_0.

33. Interpret the value of the slope b_1.

34. Interpret the value of the standard error of the estimate, s.

35. Suppose we let $\alpha = 0.10$. Perform the hypothesis test to determine if a linear relationship exists between x and y. Assume the assumptions are met.

36. Calculate the correlation coefficient r.

37. Assume normality. Construct a 90% confidence interval for the population correlation coefficient. Interpret the result.

38. Compare your results for the hypothesis test and the confidence interval. Comment.

HANDS-ON ANALYSIS

Open the *Baseball* data set, a collection of batting statistics for 331 baseball players who played in the American League in 2002, available on the book website, www.DataMiningConsultant.com. Suppose we are interested in whether there is a relationship between batting average and the number of home runs a player hits. Some fans might argue, for example, that those who hit lots of home runs also tend to make a lot of strike outs, so that their batting average is lower. Let us check it out, using a regression of the number of home runs against the player's batting average (hits divided by at bats). Because baseball batting averages tend to be highly variable for low numbers of at bats, we restrict our data set to those players who had at least 100 at bats for the 2002 season. This leaves us with 209 players. Use this data set for Exercises 39–61.

39. Construct a scatter plot of *home runs* versus *batting average*.

40. Informally, is there evidence of a relationship between the variables?

41. What would you say about the variability of the number of home runs, for those with higher batting averages?

42. Refer to the previous exercise. Which regression assumption might this presage difficulty for?

43. Perform a regression of *home runs* on *batting average*. Obtain a normal probability plot of the standardized residuals from this regression. Does the normal probability plot indicate acceptable normality, or is there skewness? If skewed, what type of skewness?

44. Construct a plot of the residuals versus the fitted values (fitted values refers to the $\hat{y}$'s). What pattern do you see? What does this indicate regarding the regression assumptions?

45. Take the natural log of *home runs*, and perform a regression of *ln home runs* on *batting average*. Obtain a normal probability plot of the standardized residuals from this regression. Does the normal probability plot indicate acceptable normality?

46. Construct a plot of the residuals versus the fitted values. Do you see strong evidence that the constant variance assumption has been violated? (Remember to avoid the Rorschach effect.) Therefore conclude that the assumptions are validated.

47. Write the population regression equation for our model. Interpret the meaning of β_0 and β_1.

48. State the regression equation (from the regression results) in words and numbers.

49. Interpret the value of the y-intercept b_0.

50. Interpret the value of the slope b_1.

51. Estimate the number
of 0.300.

52. What is the size of th
player's *batting avera*

53. What percentage of th

54. Perform the hypothes
the variables.

55. Construct and interpre
sion line.

56. Calculate the correlati
tion correlation coeffi

57. Construct and interpr
all players who had a

58. Construct and interpr
0.300 batting average

59. List the outliers. Wh
explain why he is an

60. List the high leverage
Williams a high lever

61. List the influential ob

Next, subset the *Baseball*
bats. Use this data set for I

62. We are interested in i
ber of times a player
has. Construct a scatt
relationship?

63. On the basis of the sc
not?

64. Perform the regressio
the number of stolen

65. Find and interpret the

66. What is the typical e
given his number of s

67. Interpret the y-interce

68. Inferentially, is there
this?

69. Calculate and interpre

70. Clearly interpret the r

51. Estimate the number of *home runs* (not *ln home runs*) for a player with a *batting average* of 0.300.

52. What is the size of the typical error in predicting the number of *home runs*, based on the player's *batting average*?

53. What percentage of the variability in the *ln home runs* does *batting average* account for?

54. Perform the hypothesis test for determining whether a linear relationship exists between the variables.

55. Construct and interpret a 95% confidence interval for the unknown true slope of the regression line.

56. Calculate the correlation coefficient. Construct a 95% confidence interval for the population correlation coefficient. Interpret the result.

57. Construct and interpret a 95% confidence interval for the mean number of home runs for all players who had a batting average of 0.300.

58. Construct and interpret a 95% prediction interval for a randomly chosen player with a 0.300 batting average. Is this prediction interval useful?

59. List the outliers. What do all these outliers have in common? For Orlando Palmeiro, explain why he is an outlier.

60. List the high leverage points. Why is Greg Vaughn a high leverage point? Why is Bernie Williams a high leverage point?

61. List the influential observations, according to Cook's distance and the F criterion.

Next, subset the *Baseball* data set so that we are working with batters who have at least 100 at bats. Use this data set for Exercises 62–71.

62. We are interested in investigating whether there is a linear relationship between the number of times a player has been caught stealing and the number of stolen bases the player has. Construct a scatter plot, with "caught" as the response. Is there evidence of a linear relationship?

63. On the basis of the scatter plot, is a transformation to linearity called for? Why or why not?

64. Perform the regression of the number of times a player has been caught stealing versus the number of stolen bases the player has.

65. Find and interpret the statistic that tells you how well the data fit the model.

66. What is the typical error in predicting the number of times a player is caught stealing, given his number of stolen bases?

67. Interpret the y-intercept. Does this make any sense? Why or why not?

68. Inferentially, is there a significant relationship between the two variables? What tells you this?

69. Calculate and interpret the correlation coefficient.

70. Clearly interpret the meaning of the slope coefficient.

71. Suppose someone said that knowing the number of stolen bases a player has explains most of the variability in the number of times the player gets caught stealing. What would you say?

For Exercises 72–85, use the *Cereals* data set.

72. We are interested in predicting nutrition rating based on sodium content. Construct the appropriate scatter plot. Note that there is an outlier. Identify this outlier. Explain why this cereal is an outlier.

73. Perform the appropriate regression.

74. Omit the outlier. Perform the same regression. Compare the values of the slope and y-intercept for the two regressions.

75. Using the scatter plot, explain why the y-intercept changed more than the slope when the outlier was omitted.

76. Obtain the Cook's distance value for the outlier. Is it influential?

77. Put the outlier back in the data set for the rest of the analysis. On the basis of the scatter plot, is there evidence of a linear relationship between the variables? Discuss. Characterize their relationship, if any.

78. Construct the graphics for evaluating the regression assumptions. Are they validated?

79. What is the typical error in predicting rating based on sodium content?

80. Interpret the y-intercept. Does this make any sense? Why or why not?

81. Inferentially, is there a significant relationship between the two variables? What tells you this?

82. Calculate and interpret the correlation coefficient.

83. Clearly interpret the meaning of the slope coefficient.

84. Construct and interpret a 95% confidence interval for the true nutrition rating for all cereals with a sodium content of 100.

85. Construct and interpret a 95% confidence interval for the nutrition rating for a randomly chosen cereal with sodium content of 100.

Open the *California* data set (Source: US Census Bureau, www.census.gov, and available on the book website, www.DataMiningConsultant.com), which consists of some census information for 858 towns and cities in California. This example will give us a chance to investigate handling outliers and high leverage points as well as transformations of both the predictor and the response. We are interested in approximating the relationship, if any, between the percentage of townspeople who are senior citizens and the total population of the town. That is, do the towns with higher proportions of senior citizens (over 64 years of age) tend to be larger towns or smaller towns? Use the *California* data set for Exercises 86–92.

86. Construct a scatter plot of *percentage over 64* versus *popn*. Is this graph very helpful in describing the relationship between the variables?

87. Identify the four cities that appear larger than the bulk of the data in the scatter plot.

88. Apply the *ln* transformation to the predictor, giving us the transformed predictor variable *ln popn*. Note that the application of this transformation is due solely to the skewness inherent in the variable itself (shown by the scatter plot), and is not the result of any

regression diagnostics. Pe
obtain the regression diagr

89. Describe the pattern in the

90. Describe the pattern in the
mean? Are the assumption

91. Perform the regression of
regression diagnostics. Ex
residuals versus fitted valu

92. Identify the set of outliers i
we uncovered a natural gr
the graph.

regression diagnostics. Perform the regression of *percentage over 64* on *ln popn*, and obtain the regression diagnostics.

89. Describe the pattern in the normal probability plot of the residuals. What does this mean?

90. Describe the pattern in the plot of the residuals versus the fitted values. What does this mean? Are the assumptions validated?

91. Perform the regression of *ln pct* (*ln* of *percentage over 64*) on *ln popn*, and obtain the regression diagnostics. Explain how taking the *ln* of *percentage over 64* has tamed the residuals versus fitted values plot.

92. Identify the set of outliers in the lower right of the residuals versus fitted values plot. Have we uncovered a natural grouping? Explain how this group would end up in this place in the graph.