Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 2: Inferences in Regression and Correlation Analysis

Thursday 09:10-12:00, 資訊 140301

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

Overview

1. Take up inferences (_____) concerning the regression parameters β_0 and β_1.

2. Discuss interval estimation of the mean $E(Y)$ of the probability distribution of Y, for given X, prediction intervals for a new observation Y, confidence bands for the regression line, the analysis of variance approach to regression analysis, the general linear test approach, and descriptive measures of association.

3. Assume that the normal error regression model (1.24) is applicable:

   _____,

   where β_0 and β_1, are parameters, X_i are known constants, ϵ_i are independent _____.

2.1 Inferences Concerning β_1

1. Testing whether or not _____ is that, when $\beta_1 = 0$, there is no _____ between Y and X.

   $$E(Y) = \_\_\_\_\_$$

2. For normal error regression model (2.1), the condition $\beta_1 = 0$ follows that the probability distributions of Y are _____. There is no relation of any type between Y and X.

Sampling Distribution of β_1

Question ..... (p42)

For normal error regression model (2.1), show that b_1, the point estimator of β_1, is a linear combination of the observation Y_i. That is

$$b_1 = \sum k_i Y_i.$$

sol:

Question ..... (p42)

For normal error regression model (2.1), if b_1 is expressed as $b_1 = \sum k_i Y_i$, show that

$$\sum k_i = 0,\ \sum k_i X_i = 1, \text{and } \sum k_i^2 = \frac{1}{\sum (X_i - \bar{X})^2}.$$

sol:

Question ..... (p41)

For normal error regression model (2.1), show that the sampling distribution of b_1, the point estimator of β_1, is normal, with mean and variance:

$$E(b_1) = \beta_1, \quad \text{and} \quad \sigma^2(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}.$$

sol:

Question ..... (p43)

Show that b_1 has minimum variance among all unbiased linear estimator of the form:

$$\hat{\beta}_1 = \sum c_i Y_i,$$

where the c_i are arbitrary constants.

sol:

Sampling Distribution of $(b_1 - \beta_1)/s(b_1)$

1. Since b_1 is normally distributed, we know that the standardized statistic _____ is a standard normal variable.
2. We need to estimate $\sigma(b_1)$ by _____, and hence are interested in the distribution of the statistic $(b_1 - \beta_1)/s(b_1)$.
3. When a statistic is standardized but the denominator is an estimated standard deviation rather than the true standard deviation, it is called a _____.

✎ Question ..... (p44)

Show the studentized statistic $\dfrac{b_1 - \beta_1}{s(b_1)}$ is distributed as $t_{(n-2)}$ for regression model (2.1).

sol:

Confidence Interval for β_1

✎ Question ..... (p45)

Find the $(1 - \alpha)\%$ confidence interval for β_1.

sol:

Question ..... (p45)

(Toluca Company Example) Management wishes an estimate of β_1, with 95 percent confidence coefficient.

sol:

Obtain

$$s^2(b_1) = \frac{MSE}{\sum(X_i - \bar{X})^2} = \frac{2,384}{19,800} = 0.12040, \quad s(b_1) = 0.3470.$$

For a 95 percent confidence coefficient, we find $t_{(0.975;23)} = 2.069$. The 95 percent confidence interval:

$$3.5702 - 2.069(0.3470) \leq \beta_1 \leq 3.5702 + 2.069(0.3470)$$

$$\Rightarrow 2.85 \leq \beta_1 \leq 4.29$$

Thus, with confidence coefficient .95, we estimate that the mean number of work hours increases by somewhere between 2.85 and 4.29 hours for each additional unit in the lot.

FIGURE 2.2 Portion of MINITAB Regression Output—Toluca Company Example.

```
The regression equation is
Y = 62.4 + 3.57 X

Predictor       Coef      Stdev    t-ratio          p
Constant       62.37      26.18       2.38      0.026
X             3.5702     0.3470      10.29      0.000

s = 48.82      R-sq = 82.2%     R-sq(adj) = 81.4%

Analysis of Variance

SOURCE       DF          SS          MS          F          p
Regression    1      252378      252378     105.88      0.000
Error        23       54825        2384
Total        24      307203
```

Tests Concerning β_1

✎ Question ..... (p47)

Two-Sided Test A cost analyst in the Toluca Company is interested in testing, using regression model (2.1), whether or not there is a linear association between work hours and lot size, i.e., whether or not, $\beta_1 = 0$. Please conduct the Two-Sided Test for this problem and control the risk of a Type I error at $\alpha = 0.05$.

sol:

✎ Question ..... (p47)

One-Sided Test Suppose the analyst in the Toluca Company had wished to test whether or not , β_1, is positive, controlling the level of significance at $\alpha = 0.05$. Please conduct the One-Sided Test for this problem.

sol:

Comments:

1. The P-value is sometimes called the _____.
2. Many scientific publications commonly report the P-value together with the value of the test statistic. In this way, one can conduct a test at any desired level of significance a by comparing the P-value with the specified level aα.
3. Users of statistical calculators and computer packages need to be careful to ascertain whether _____ or _____ P-values are reported.
4. It is desired to test whether or not β_1 equals some specified nonzero value _____. The alternatives are:

   _____ versus _____

   and the appropriate test statistic is:

   _____

2.2 Inferences Concerning β_0

1. The point estimator b_0: _____.
2. The sampling distribution of b_0 is normal, with mean and variance:

   $$E(b_0) = \_\_\_\_\_, \quad \sigma^2(b_0) = \_\_\_\_\_$$

3. An estimator of $\sigma^2(b_0)$is obtained by replacing σ^2 by its point estimator _____:

   $$s^2(b_0) = \_\_\_\_\_$$

4. The Sampling distribution of $(b_0 - \beta_0)/s(b_0)$ is _____ for regression model (2.1)
5. The confidence intervals for β_0 is _____.

2.3 Some Considerations on Making Inferences Concerning β_0 and β_1

Effects of Departures from Normality

1. If the probability distributions of Y are not exactly normal but _____ _____, the sampling distributions of b_0 and b_1 will be approximately _____, and the use of the t distribution will provide approximately the specified confidence coefficient or level of significance.

2. Even if the distributions of Y are far from normal, the estimators b_0 and b_1 generally have the property of _____ - their distributions approach normality under very general conditions as the _____ increases.

Interpretation of Confidence Coefficient and Risks of Errors

1. Since regression model (2.1) assumes that the X_i are known constants, the, confidence coefficient and risks of errors are interpreted with respect to taking _____ in which the X observations are kept at the same levels as in the observed sample.

2. (Toluca Company Example) The meaning of a confidence interval (CI) for β_1, with confidence coefficient 0.95: if many independent samples are taken where the levels of X (the lot sizes) are the same as in the data set and a 95 percent confidence interval is constructed for each sample, _____ of the intervals will _____ the true value of β_1.

Spacing of the X levels

1. For given n and σ^2, the variances of $b_1$and b_0 are affected by the spacing of the X levels in the observed data.

2. The _____ is the spread in the X levels, the larger is the quantity _____ and the _____ is the variance of b_1.

Power of Tests

(NOTE: The power of tests on β_0 and β_1, can be obtained from Appendix Table B.5.)

1. The general test concerning β_1:

$$H_0 : \_\_\_\_\_ \quad \text{versus} \quad H_a : \_\_\_\_\_$$

2. Test statistic: $t^* =$ _____ .

3. Decision rule for level of significance α:

$$\text{If } \_\_\_\_\_, \text{ conclude } H_0.$$

$$\text{If } |t^*| > t_{(1-\alpha/2;n-2)}, \text{ conclude } H_a.$$

4. The power of this test is the probability that the decision rule will lead to conclusion H_a when H_a in fact holds:

$$\text{Power} = \_\_\_\_\_$$

where δ is the noncentrality measure - i.e., a measure of how far the true value of β_1, is from β_{10}:

$$\delta = \_\_\_\_\_$$

✎ Question ..... (p51)

In Toluca Company example, conduct the test for:

$$H_0 : \beta_1 = \beta_{10} = 0, \quad \text{versus} \quad H_a : \beta_1 \neq \beta_{10} = 0.$$

Calculate the power of the test when $\beta_1 = 1.5$.

sol:

2.4 Interval Estimation of $E(Y_h)$

1. Let _____ denote the level of X for which we wish to estimate the mean response.
2. X_h may be a value which occurred in the sample, or it may be some other value of the predictor variable within the scope of the model.
3. The mean response when $X = X_h$ is denoted by _____. The point estimator Y_h of $E(Y_h)$ is _____.

✎ Question ..... (p52)

For normal error regression model, show that the sampling distribution of Y_h is normal, with mean and variance:

$$E(Y_h) = E(Y_h) \quad \text{and} \quad \sigma^2(Y_h) = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right].$$

sol:

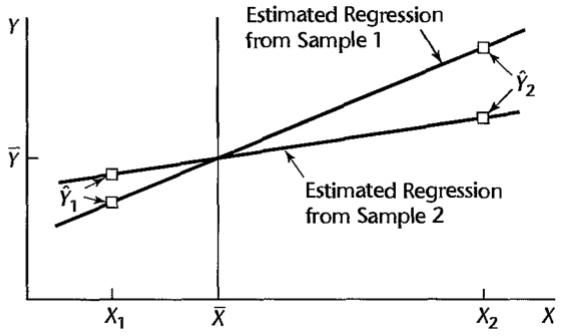


FIGURE 2.3 Effect on $\hat{Y}_h$ of Variation in b_1 from Sample to Sample in Two Samples with Same Means $\bar{Y}$ and $\bar{X}$.

Sampling Distribution of $(\hat{Y}_h - E(Y_h))/s(\hat{Y}_h)$

1. $\dfrac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)}$ is distributed as _____ for regression model (2.1).

Confidence Interval for $E(Y_h)$

1. A $(1-\alpha)\%$ confidence interval for $E(Y_h)$ is

✎ Question ..... (p54)

In the Toluca Company example, find a 90% CI for $E(Y_h)$ when the lot size is $X_h = 65$ units.

sol:

2.5 Prediction of New Observation

The new observation on Y to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based. We denote the level of X for

the new trial as _____ and the new observation on Y as _____.

Prediction Interval for $Y_{h(new)}$ when Parameters Known

In general, when the regression parameters of normal error regression model (2.1) are known, the $(1-\alpha)\%$ prediction limits for $Y_{h(new)}$ are:

$$E(Y_h) \pm z_{(1-\alpha/2)}\sigma$$

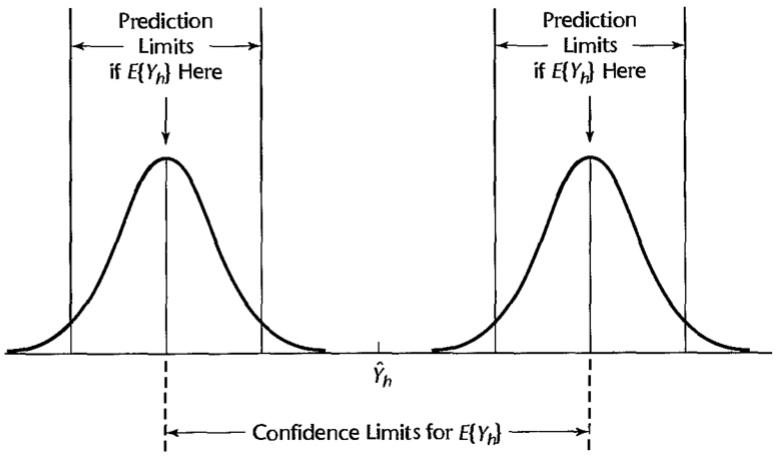
Prediction Interval for $Y_{h(new)}$ when Parameters Unknown

✎ Question ..... (p58)

As we know, $\dfrac{Y_{h(new)} - \hat{Y}_h}{s(\text{pred})}$ is distributed as $t_{(n-2)}$ for a normal error regression model. Find the prediction limits for a new observation $Y_{h(new)}$ at a given level X_h.

sol:

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



✎ Question ..... (p59)

The Toluca Company studied the relationship between lot size and work hours primarily to obtain information on the mean work hours required for different lot sizes for use in determining the optimum lot size. The company was also interested, however, to see whether the regression relationship is useful for predicting the required work hours for individual lots. Find a 90 percent prediction interval for the number of work hours for the next production runs of $X_h = 100$ units.

sol:

Prediction of Mean of m New Observations for Given X_h

1. Denote the mean of the new Y observations to be predicted as _____. The $1-\alpha$ prediction limits are, assuming that the new Y observations are independent:

   _____

   where

   $$s(\text{predmean}) = \_\_\_\_\_$$

   or equivalently:

   $$s(\text{predmean}) = \_\_\_\_\_ .$$

✎ Question ..... (p61)

In the Toluca Company example, find the 90 percent prediction interval for the mean number of work hours $\bar{Y}_{h(new)}$ in three new production runs, each for $X_h = 100$ units.

sol:

2.6 Confidence-Band for Regression Line

1. A confidence band for the entire regression line $E(Y) = \beta_0 + \beta_1 X$ enables us to see the _____ in which the entire regression line lies. It is particularly useful for determining the appropriateness of a fitted regression function.

2. The Working-Hotelling $(1-\alpha)\%$ confidence band for the regression line for regression model (2.1) has the following two boundary values at any level X_h:

   _____, where _____.

✎ Question ..... (p62)

Find the 90 percent confidence band for the regression line to determine how precisely we have been able to estimate the regression function for the Toluca Company example.

sol:

2.7 Analysis of Variance Approach to Regression Analysis

Partitioning of Total Sum of Squares

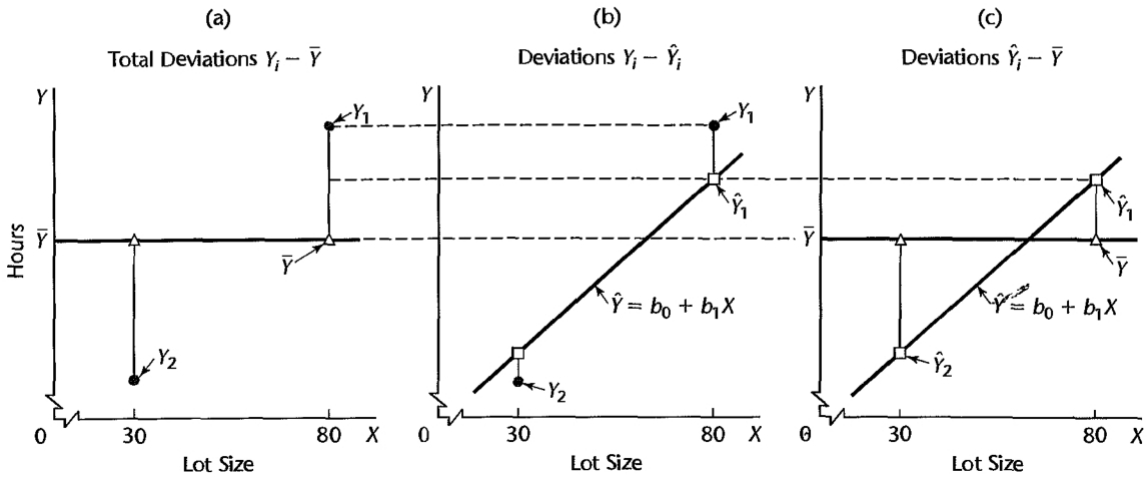
1. The variation is measured in terms of the deviations of the Y_i around their mean $\bar{Y}$: _____.

2. $SSTO$ (total sum of squares): the measure of total variation is the sum of the squared deviations: _____.

3. SSE (error sum of squares): the measure of variation in Y_i that is present when the predictor variable X is taken into account: _____.

4. SSR (regression sum of squares): _____.

FIGURE 2.7 Illustration of Partitioning of Total Deviations $Y_i - \bar{Y}$—Toluca Company Example (not drawn to scale; only observations Y_1 and Y_2 are shown).



✎ Question ..... (p65)

Show that $SSTO = SSR + SSE$. That is

$$\sum(Y_i - \bar{Y}_i)^2 = \sum(\hat{Y}_u - \bar{Y}_i)^2 + \sum(Y_i - \hat{Y}_i)^2$$

sol:

Breakdown of Degrees of Freedom

1. Corresponding to the partitioning of the total sum of squares $SSTO$, there is a partitioning of the associated degrees of freedom (df).

2. SSTO has _____ degrees of freedom associated with it. One degree of freedom is lost because the deviations _____ are subject to one constraint: they must sum to _____. Equivalently, one degree of freedom is lost because the sample mean $\bar{Y}$ is used to estimate the population mean.

3. SSE has _____ degrees of freedom associated with it. Two degrees of freedom are lost because the two parameters _____ are estimated in obtaining the fitted values $\hat{Y}_i$ ·

4. SSR has _____ degree of freedom associated with it. Although there are n deviations _____, all fitted values $\hat{Y}_i$ are calculated from the same estimated regression line.

Mean Squares

1. A sum of squares divided by its associated degrees of freedom is called a _____ (MS).
2. The regression mean square: _____.
3. The error mean square: _____.

Analysis of Variance Table

1. **Basic Table**:
   (a) The breakdowns of the total sum of squares and associated degrees of freedom are displayed in the form of an analysis of variance table _____ in Table 2.2.
   (b) The ANOVA table contains a column of _____ that will be utilized.

TABLE 2.2 ANOVA Table for Simple Linear Regression.

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

2. **Modified Table**:
   (a) The modified ANOVA table is based on the fact that the total sum of squares can be decomposed into two parts:

$$SSTO = \_\_\_\_\_ = \_\_\_\_\_$$

   (b) In the modified ANOVA table, the total _____ sum of squares, denoted by SSTOU, is defined as:

$$SSTOU = \_\_\_\_\_$$

and the correction for the mean sum of squares, denoted by SS(correction for mean), is defined as:

$$SS(\text{correction for mean}) = \_\_\_\_\_$$

TABLE 2.3 Modified ANOVA Table for Simple Linear Regression.

Source of Variation	SS	df	MS
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$	
Correction for mean	$SS(\text{correction for mean}) = n\bar{Y}^2$	1	
Total, uncorrected	$SSTOU = \sum Y_i^2$	n	

Expected Mean Squares

Question ..... (p68)

Show that

$$\begin{aligned} E(MSE) &= \sigma^2, \quad \text{and} \\ E(MSR) &= \sigma^2 + \beta^2 \sum (X_i - \bar{X})^2. \end{aligned}$$

sol:

F Test of $\beta_1 = 0$ versus $\beta_1 \neq 0$

1. The analysis of variance provides us with a test for:

   _____ versus _____.

2. **Test Statistic**: The test statistic for the analysis of variance approach is denoted by F^*:

   $$F^* = \frac{\quad}{\quad}$$

3. Large values of F^* support _____ and values of F^* near _____ support H_0.

   ✎ Question ..... (p70)

   Show that if H_0 holds, F^* follows the noncentral $F_{(1,n-2)}$ distribution.

   sol:

1. **Construction of Decision Rule**: Since the test is upper-tail and F^* is distributed as $F_{(1,n-2)}$ when H_0 holds, the decision rule is as follows when the risk of a Type I error is to be controlled at α:

   If _____, conclude H_0,

   If $F^* > F_{(1-\alpha;1,n-2)}$, conclude H_a

Question (p71)

For the Toluca Company example, conduct a F test for $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

sol:

Question (p71)

Show that for a given α level, the F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is equivalent algebraically to the two-tailed t test.

sol:

2.8 General Linear Test Approach

Full Model

1. For the simple linear regression case, the full model or unrestricted model is the normal error regression model:

   _____.

2. The error sum of squares for the full model:

   $SSE(F) =$ _____ = _____ = _____.

3. $SSE(F)$ measures the variability of the Y_i observations around the fitted regression line.

Reduced Model

1. Consider $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, the model when H_0 holds is called the reduced or restricted model:

   _____.

2. The error sum of squares for the reduced model:

   $SSE(R) =$ _____ = _____ = _____.

Test Statistic

1. It can be shown that $SSE(F)$ never is greater than $SSE(R)$:

   _____.

2. The actual test statistic is a function of $SSE(R) - SSE(F)$,

   $$F^* = \frac{\qquad\qquad\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad\qquad\qquad},$$

   which follows the F distribution when H_0 holds.

3. The decision rule therefore is:

$$\text{If } \underline{\qquad\qquad\qquad\qquad}, \text{ conclude } H_0$$

$$\text{If } F^* > F_{(l-\alpha;df_R-df_F,df_F)}, \text{ conclude } H_a$$

4. For testing whether or not $\beta_1 = 0$, we therefore have:

$$SSE(R) = SSTO, \quad SSE(F) = SSE, \quad df_R = n-1, \quad df_F = n-2,$$

so that we obtain

$$F^* = \underline{\qquad\qquad\qquad\qquad} = \underline{\qquad\qquad} = \underline{\qquad\qquad}$$

which is identical to the analysis of variance test statistic.

2.9 Descriptive Measures of Linear Association between X and Y

Coefficient of Determination

1. The coefficient of determination R^2 is defined to measure the effect of X in reducing the variation in Y. It is expressed as the reduction in variation _____ as a proportion of the total variation:

$$R^2 = \underline{\qquad\qquad} = \underline{\qquad\qquad}.$$

2. We may interpret R^2 (_____) as the proportionate reduction of total variation associated with the use of the predictor variable X.

3. The larger R^2 is, the more the total _____ of Y is reduced by introducing the predictor variable X.

4. The limiting values of R^2 may occur:

   (a) When all observations fall on the fitted regression line, then _____ and _____. The predictor variable X accounts for _____ in the observations Y_i

(b) When the fitted regression line is horizontal so that _____ and _____, then _____ and _____. There is no linear association between X and Y in the sample data.

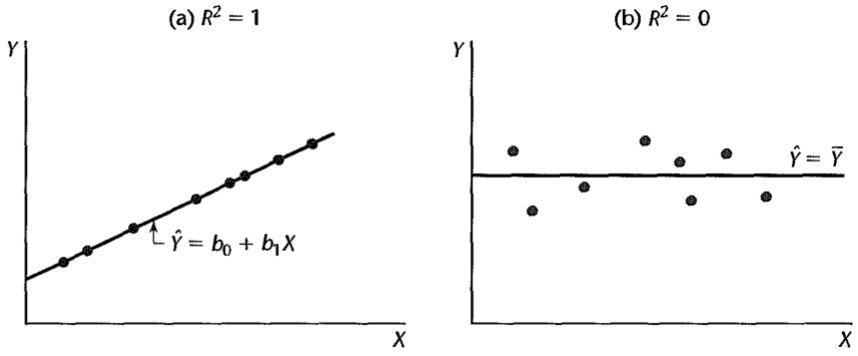


FIGURE 2.8 Scatter Plots when $R^2 = 1$ and $R^2 = 0$.

limitations of R^2: three common misunderstandings

1. **Misanderstanding 1**: A high R^2 indicates that _____ can be made. (not necessarily correct)

   (a) (Toluca Company Example) the coefficient of determination was high ($R2 = 0.82$). Yet the 90 percent prediction interval for the next lot, consisting of 100 units, was wide (332 to 507 hours) and not precise enough to permit management to schedule workers effectively.

   (b) Misunderstanding 1 arises because R^2 measures only a _____ from $SSTO$ and provides no information about absolute precision for estimating a mean response or predicting a new observation.

2. **Misunderstanding 2**: A high R^2 indicates that the estimated regression line is a _____. (not necessarily correct)

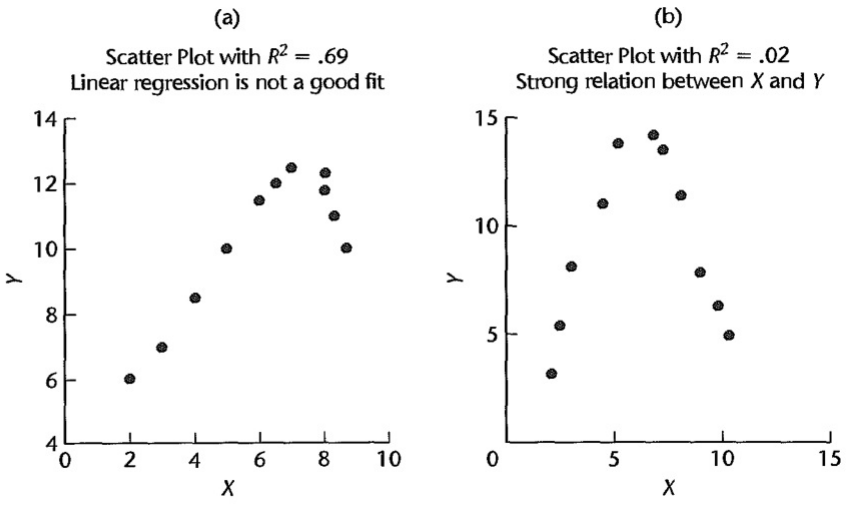
   (a) (Figure 2.9a) a scatter plot where R^2 is high ($R^2 = 0.69$). Yet a linear regression function would not be a good fit since the regression relation is curvilinear.

3. **Misunderstanding 3**: A R^2 near zero indicates that X and Y are not related. (not necessarily correct).

(a) (Figure 2.9b) a scatter plot where R^2 between X and Y is $R^2 = 0.02$. Yet X and Y are strongly related; however, the relationship between the two variables is curvilinear.

(b) Misunderstandings 2 and 3 arise because R^2 measures the degree of _____ between X and Y, whereas the actual regression relation may be curvilinear.

FIGURE 2.9 Illustrations of Two Misunderstandings about Coefficient of Determination.



Coefficient of Correlation

1. A measure of linear association between Y and X when both Y and X are random is the coefficient of correlation. This measure is the signed square root of R^2:

   _____

2. A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is _____ or _____. Thus, the range of r is: _____.

2.10 Considerations in Applying Regression Analysis*

2.11 Normal Correlation Models*

😍 TA Class

- **Problems**: 2.5, 2.8, 2.10, 2.14, 2.17, 2.24, 2.30, 2.31, 2.32
- **Exercises**: 2.50, 2.55
- **Projects**: 2.62