Chapter 2: Inference in Regression Analysis

Review: Normal error regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i: observed response in the ith trial
- x_i: a known constant, the level of the predictor variable in the ith trial
- β_0 and β_1: parameters
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2, \cdots, n$

Inference about β_1

- $b_1 \sim N(\beta_1; \text{var}(b_1))$

  where $\text{var}(b_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

- $\widehat{\text{var}(b_1)} = s^2(b_1) = \dfrac{\text{MSE}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

- $t = \dfrac{b_1 - \beta_1}{s(b_1)} \sim t(n-2)$

  Note: The degrees of freedom (df) here is $n-2$ instead of $n-1$, because besides estimating β_1, we are also estimating β_0.

Confidence Interval for β_1

$$b_1 \pm t\left(1 - \frac{\alpha}{2}, n-2\right) s(b_1)$$

- $t\left(1 - \frac{\alpha}{2}, n-2\right)$ is the upper tail $100(1-\alpha/2)$ percentile of the t-distribution with $n-2$ degrees of freedom
- $1-\alpha$ is the confidence level

Significance Tests for β_1

$$H_0 : \beta_1 = \beta_{10}$$

Test statistic: $t^* = \dfrac{b_1 - \beta_{10}}{s(b_1)}$ = # of standard deviation of b_1 separating b_1 and β_{10}

	Reject H_0, if
$H_\alpha : \beta_1 > \beta_{10}$	$t^* > t(1-\alpha; n-2)$
$H_\alpha : \beta_1 < \beta_{10}$	$t^* < -t(1-\alpha; n-2)$
$H_\alpha : \beta_1 \neq \beta_{10}$	$\lvert t^* \rvert > t(1-\alpha/2; n-2)$

or Reject H_0 when p value is less than the significance level α.

Recall: p value is the probability of observing a more extreme value of the test statistics than was actually observed

Inference about β_0

$$b_0 \sim N(\beta_0; \text{var}(b_0))$$

where

$$\text{var}(b_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\widehat{\text{var}(b_0)} = s^2(b_0) = \frac{\text{MSE} \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$t = \frac{b_0 - \beta_0}{s(b_0)} \sim t(n-2)$$

Note: The df here is $n-2$ instead of $n-1$, because besides estimating β_0, we are also estimating β_1.

Confidence Interval for β_0

$$b_0 \pm t(1 - \frac{\alpha}{2}, n - 2)s(b_0)$$

- $t(1 - \frac{\alpha}{2}, n - 2)$ is the upper tail $100(1 - \alpha/2)$ percentile of the t-distribution with $n - 2$ degrees of freedom
- $1 - \alpha$ is the confidence level

Significance Tests for β_0

$$H_0 : \beta_0 = \beta_{00}$$

Test statistic: $t^* = \dfrac{b_0 - \beta_{00}}{s(b_0)}$ = # of standard deviation of b_0 separating b_0 and β_{00}

	Reject H_0, if
$H_\alpha : \beta_0 > \beta_{00}$	$t^* > t(1-\alpha; n-2)$
$H_\alpha : \beta_0 < \beta_{00}$	$t^* < -t(1-\alpha; n-2)$
$H_\alpha : \beta_0 \neq \beta_{00}$	$\lvert t^* \rvert > t(1-\alpha/2; n-2)$

or Reject H_0 when p value is less than the significance level α.

Comments:

- The normality of b_0 and b_1 follows from the fact that each is a linear combination of the y_i's and y_i's are independent and normally distributed
- Often the CI and significant test for β_0 is not of interest.
- If the ε are not normal but are approximately normal, then the CI's and significant tests are generally reasonable approximations.

Sec 2.3 Power of Test

The power of a significant test is the probability that the decision rule will lead to conclusion H_α when H_α in fact holds. This probability depends on the particular value of the parameter in the alternative space.

- $H_0 : \beta_1 = \beta_{10} \quad v.s \quad H_\alpha : \beta_1 \neq \beta_{10}$
- $t^* = \dfrac{b_1 - \beta_{10}}{s(b_1)}$
  $|t^*| \leq t(1-\alpha/2; n-2)$, conclude H_0
  $|t^*| > t(1-\alpha/2; n-2)$, conclude H_α
-

$$\text{Power} = P\{|t^*| > t(1-\alpha/2; n-2)|\delta\}$$

where δ is the noncentrality measure—a measure of how far the true value of β_1 is from β_{10}, $\delta = \dfrac{|\beta_1 - \beta_{10}|}{\sigma\{b_1\}}$.

Example: Test

$$H_0 : \beta_1 = \beta_{10} = 0$$

v.s

$$H_\alpha : \beta_1 \neq \beta_{10} = 0$$

suppose we wish to know the power of the test when $\beta_1 = 1.5$. Also assume that $\sigma\{b_1\} = .3553$ and degrees of freedom is 23. Then $\delta = |1.5 - 0|/.3553 = 4.22$. Use Table B.5 for $\alpha = .05$ and 23 degrees of freedom and interpolate linearly between $\delta = 4.00$ and $\delta = 5.00$. We obtain:

$$.97 + \frac{4.22 - 4.00}{5.00 - 4.00}(1.00 - .97) = .9766$$

Thus, if $\beta_1 = 1.5$, the probability would be about .98 that we would be led to conclude $H_\alpha(\beta_1 \neq 0)$.

Hypothesis testing and Confidence Interval in R

R Handout 3 example 1

Section 2.4 Interval Estimation of $E(Y_h)$ (a point on the line $y = \beta_0 + \beta_1 x$)

The mean value of Y for $X = X_h$ is

$$E(Y_h) = u_h = \beta_0 + \beta_1 X_h$$

$$\widehat{E(Y_h)} = \hat{Y}_h = b_0 + b_1 X_h$$

Theory for Estimation of $E\{Y_h\}$

- $\hat{Y}_h \sim N(\beta_0 + \beta_1 X_h, \sigma^2\{\hat{Y}_h\})$ where

$$\sigma^2\{\hat{Y}_h\} = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right].$$

—Normality is a consequence of the fact that $\hat{Y}_h = b_0 + b_1 X_h$ is a linear combination of Y_i's.

—The estimation is more accurate when X_h is near $\bar{X}$.

- We estimate $\sigma^2\{\hat{Y}_h\}$ by

$$s^2\{\hat{Y}_h\} = \text{MSE}\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right].$$

- It follows that the appropriate distribution for inferences about the point $\beta_0 + \beta_1 X_h$ is

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 X_h) - (\beta_0 + \beta_1 X_h)}{\text{SE}(\hat{\beta}_0 + \hat{\beta}_1 X_h)} = \frac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)} \sim t_{n-2}.$$

- 95% CI for $E(Y_h)$ is

$$\hat{Y}_h \pm t_{n-2}(1 - \alpha/2)s\{\hat{Y}_h\}$$

Sec 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.

- X_h: the level of X for the new trial
- $Y_{h(\text{new})}$: new observation on Y
- Assume that the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation
- Want to construct an interval into which we predict the next observation $Y_{h(\text{new})}$ (for a given X_h) will fall.

Prediction limits for a new observation $Y_{h(\text{new})}$ at a given level X_h are obtained by means of the following theorem:

$$\frac{Y_{h(\text{new})} - \hat{Y}_h}{s\{\text{pred}\}} \sim t(n-2)$$

- Use point estimator $\hat{Y}_h$ in the numerator rather than the true mean $E(Y_h)$, since the true mean is unknown and canot be used in making a prediction
- Numerator represents how far the new observation $Y_{h(\text{new})}$ will deviate from the estimated mean $\hat{Y}_h$ based on the original n cases in the study. This difference may be viewed as the prediction error, with $\hat{Y}_h$ serving as the best point estimate of the value of the new observation $Y_{h(\text{new})}$

$$\begin{aligned}
\sigma^2\{\text{pred}\} &= \sigma^2\{Y_{h(\text{new})} - \hat{Y}_h\} \\
&= \sigma^2\{Y_{h(\text{new})}\} + \sigma^2\{\hat{Y}_h\} \\
&= \sigma^2 + \sigma^2\{\hat{Y}_h\}
\end{aligned}$$

- $\sigma^2\{\text{pred}\}$ has two components

  —-(1) The variance of the distribution of Y at $X = X_h$, namely σ^2

  —-(2) The variance of the sampling distribution of $\hat{Y}_h$, namely $\sigma^2\{\hat{Y}_h\}$

$$\begin{aligned} s^2\{\text{pred}\} &= MSE + s^2\{\hat{Y}_h\} \\ &= MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \end{aligned}$$

- The prediction interval for $\hat{Y}_{h(new)}$ is wider than the confidence interval for $\hat{Y}_h$ because it has a larger variance.

- When the variance σ^2 is large, it is not always possible to get good prediction intervals.

R Handout 3

Sec 2.7 Analysis of Variance (ANOVA) Approach to Regression Analysis

Variance Decomposition

- Total variation in y is measured by

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

this is a measure of uncertainty (variability) in predicting y without considering x.

- If we fit a regression line to the data and obtain $\hat{y}_i = b_0 + b_1 x_i$, the evaluation in y's about regression line is measured by

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

SSE is a measure of uncertainty (variability) in predicting y when we use information about x (in a linear regression model)

–If there is a linear relationship between y and x, then SSE is small compared to SSTO.

–If there is no linear relationship between y and x, then $\hat{y}_i = b_0 = \bar{y}$ will be close to $\bar{y}$ and SSE will be close to SSTO. ($b_0 = \bar{y} - b_1 \bar{x}$)

- We use SSR = SSTO-SSE to measure the strength of linear association between y and x or a measure of the effect of x in reducing variation in predicting y.

- Partition SSTO into two parts

  Model (explained by regression)

  Error (unexplained /residual)

by normal equations

$\sum_{i=1}^{n} e_i = 0$

$\sum_{i=1}^{n} e_i x_i = 0$

$$
\begin{aligned}
\text{SSTO} &= \sum_{i=1}^{i=n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{i=n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= \text{SSE} + \text{SSR} + 2 \sum e_i \hat{y}_i - 2 \sum e_i \bar{y} \\
&= \text{SSE} + \text{SSR} + 2 \sum e_i (b_0 + b_1 x_i) \\
&= \text{SSE} + \text{SSR} + 2 b_1 \sum e_i x_i \\
&= \text{SSE} + \text{SSR}
\end{aligned}
$$

$$\text{SSTO} = \sum_{i=1}^{i=n}(y_i - \bar{y})^2 = SSE + SSR$$

—SSR is the sum of squares due to regression model. If $\hat{y}_i = b_0 + b_1 x_i$ does a good job predicting y_i for given x_i, then $\hat{y}_i \approx y_i$ and

$$SSTO = \sum(y_i - \bar{y})^2 \approx \sum(\hat{y}_i - \bar{y})^2 = SSR,$$

so SSE will be small

—on the other hand, if there is no relationship between y and x, $b_1 \approx 0$, and $\hat{y}_i = \bar{y}$,

$$SSTO = \sum(y_i - \bar{y})^2 \approx \sum(y_i - \hat{y}_i)^2 = SSE,$$

and SSR will be small.

—Linear relationship

Yes $\rightarrow$ SSTO $\approx$ SSR, therefore SSE small

No $\rightarrow$ SSTO $\approx$ SSE, therefore SSR small

Degrees of Freedom Decomposition

- SSTO, degrees of freedom is $n - 1$, since we are estimating $\bar{y}$, lose one df
- SSE, df is $n - 2$, since we are estimating β_0 and β_1.
- SSR, df is 1, although there are n deviations $\hat{y}_i - \bar{y}$, all fitted value $\hat{y}_i$ are calculated from the same estimated regression line. Two degrees of freedom are associated with a regression line. One of them is lost because $\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = 0$.

Mean Squares

Mean squares are computed by dividing sum of squares by its degrees of freedom.

Analysis of Variance (ANOVA) Table (From Page 143)

Source of Variation	SS	df	MS	F
Regression	SSR=$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	MSR = $\frac{SSR}{1}$	$\frac{\text{MSR}}{\text{MSE}}$
Error	SSE=$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	n-2	MSE = $\frac{SSE}{n-2}$	σ^2
Total	SSTO=$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

- Rhandout 3

Section 2.7

- Review ANOVA Table
- F test of $H_0 : \beta_1 = 0$ v.s $H_\alpha : \beta_1 \neq 0$
- $E(MSR) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2$ & $E(MSE) = \sigma^2$
- If $\beta_1 = 0$, MSR & MSE both estimate σ^2 and they should be approximately equal
- If $\beta_1 \neq 0$, $E(MSR)$ is larger than $E(MSE)$. MSR should be larger than MSE
- $F^* = \dfrac{MSR}{MSE}$, reject $H_0 : \beta_1 = 0$, if F^* is too big

- If $H_0 : \beta_1 = 0$ is true and we have the normal error model, F^* has an distribution with degrees of freedom 1 and $n-2$

- At level of significance α

  Reject $H_0 : \beta_1 = 0$ against $H_\alpha : \beta_1 \neq 0$,

  if $F^* = \dfrac{MSR}{MSE} > F(1-\alpha; 1, n-2)$

- The F^* Test of $H_0 : \beta_1 = 0$ v.s $H_\alpha : \beta_1 \neq 0$ is equivalent to the t^* test of H_0 v.s H_α with $(t^*)^2 = F^*$.

Recall

Inference about β_1

- $b_1 \sim N(\beta_1; \text{var}(b_1))$

  where $\text{var}(b_1) = \dfrac{\sigma^2}{\sum(x_i - \bar{x})^2}$

- $\widehat{\text{var}(b_1)} = s^2(b_1) = \dfrac{\text{MSE}}{\sum(x_i - \bar{x})^2}$

- $t = \dfrac{b_1 - \beta_1}{s(b_1)} \sim t(n-2)$

  Note: The degrees of freedom (df) here is $n-2$ instead of $n-1$, because besides estimating β_1, we are also estimating β_0.

$$
\begin{aligned}
|t^*| &= \left|\frac{b_1}{s(b_1)}\right| \\
&= \left|\frac{b_1}{\sqrt{MSE/\sum(x_i-\bar{x})^2}}\right| \\
&= \left|\frac{b_1\sqrt{\sum(x_i-\bar{x})^2}}{\sqrt{MSE}}\right| \\
&= \sqrt{\frac{b_1^2\sum(x_i-\bar{x})^2}{MSE}} \\
&= \sqrt{\frac{MSR}{MSE}} \\
&= \sqrt{F^*}
\end{aligned}
$$

Prove $MSR = b_1^2 \sum (x_i - \bar{x})^2$

$$
\begin{aligned}
MSR &= SSR/1 \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (b_0 + b_1 x_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (\bar{y} - b_1 \bar{x} + b_1 x_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} b_1^2 (x_i - \bar{x})^2
\end{aligned}
$$

- The t^* and F^* statistics are functionally related

  Reject $H_0 : \beta_1 = 0$ v.s $H_\alpha : \beta_1 \neq 0$ when $|t^*| > t(1-\frac{\alpha}{2}; n-2)$

  $\Leftrightarrow$ Reject H_0 when $(t^*)^2 > [t(1-\frac{\alpha}{2}; n-2)]^2$

  $\Leftrightarrow$ Reject H_0 when $F^* > F(1-\alpha; 1, n-2) = [t(1-\frac{\alpha}{2}; n-2)]^2$

Section 2.8 General Linear Test

- A different view of testing $\beta_1 = 0$
- Want to use error sum of squares to compare two models:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{(full model)}$$

$$Y_i = \beta_0 + \varepsilon_i \quad \text{(reduced model)}$$

- Define $SSE(F)$ be the sum of squares error of the full model

  $SSE(R)$ be the sum of squares error of the reduced model

- Idea: compare $SSE(F)$ and $SSE(R)$
  If $SSE(F)$ is a lot smaller than $SSE(R)$, use the full model (thus $\beta_1 \neq 0$)
  otherwise, if they are close, use the reduced model (thus, $\beta_1 = 0$)

- Define

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{dfE(R) - dfE(F)}}{\frac{SSE(F)}{dfE(F)}}$$

- Reject $H_0 : \beta_1 = 0$ if $F^* > F(1 - \alpha; df_R - df_F, df_F)$

- For simple linear regression, it is the same test as before

R handout 3

General approach to testing in regression

- Fit the full model and obtain $SSE(F)$
- Fit the reduced model (the model with conditions placed by hypotheses) and get $SSE(R)$
- Find F^* and test the hypothesis specified by the reduced model

Pearson Correlation

ρ is the usual correlation coefficient (estimated by r).

- It is a number between -1 and $+1$ that measures the strength of the linear relationship between two variables

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

since

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

we have

$$r = b_1 \sqrt{\frac{\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2}}$$

R^2 and r^2

- $R^2 = \dfrac{SSR}{SSTO}$ is the ratio of explained and total variation, or proportion of variation in Y that is accounted for by the regression relationship with X
- r^2 is the square of the correlation between X and Y
- Note, in simple linear regression, R^2 is the same as r^2. $r = \pm\sqrt{R^2}$. But this is not true in Multiple Linear Regression. There will be a different r for each X variable, but only one R^2.

Adjusted R^2 takes into account the number of predictor variables and the sample size, i.e., it is adjusted based on the df. Adjusted R^2 becomes more relevant as a diagnostic tool when used in multiple regression.

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

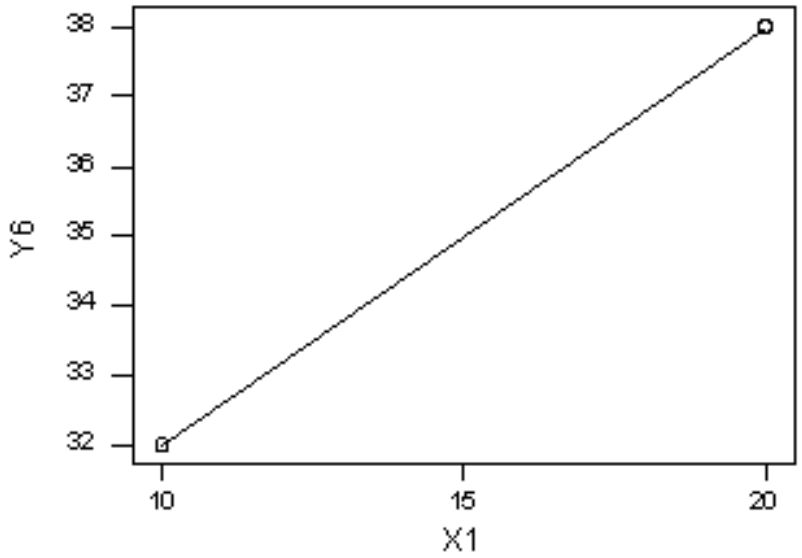
—n: number of observations in the sample

—k: number of predictor variables

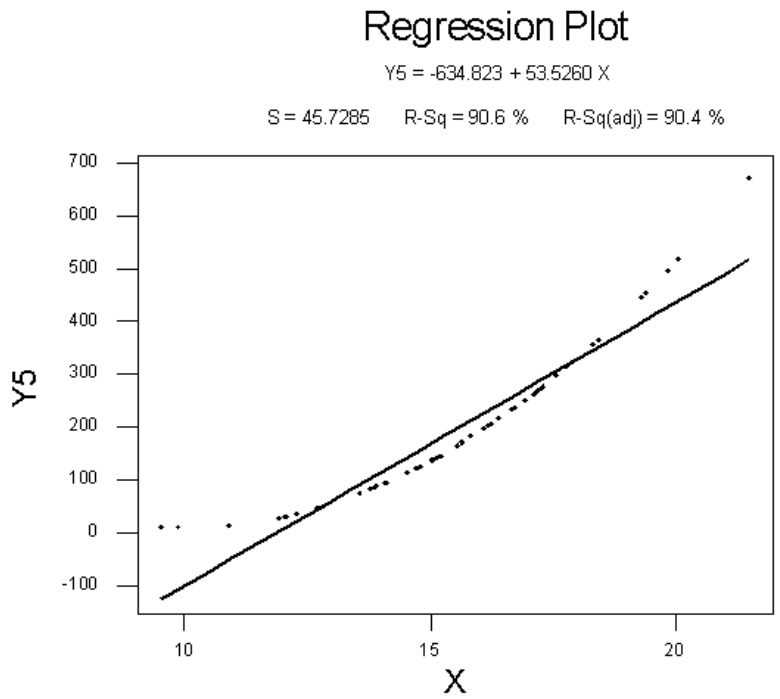
From this formula, you can see that when the number of observations is small and the number of predictors is large, there will be a much greater difference between R-square and adjusted R-square. By contrast, when the number of observations is very large compared to the number of predictors, the value of R-square and adjusted R-square will be much closer because the ratio of $(n-1)/(n-k-1)$ will approach 1.

Limitations of R^2

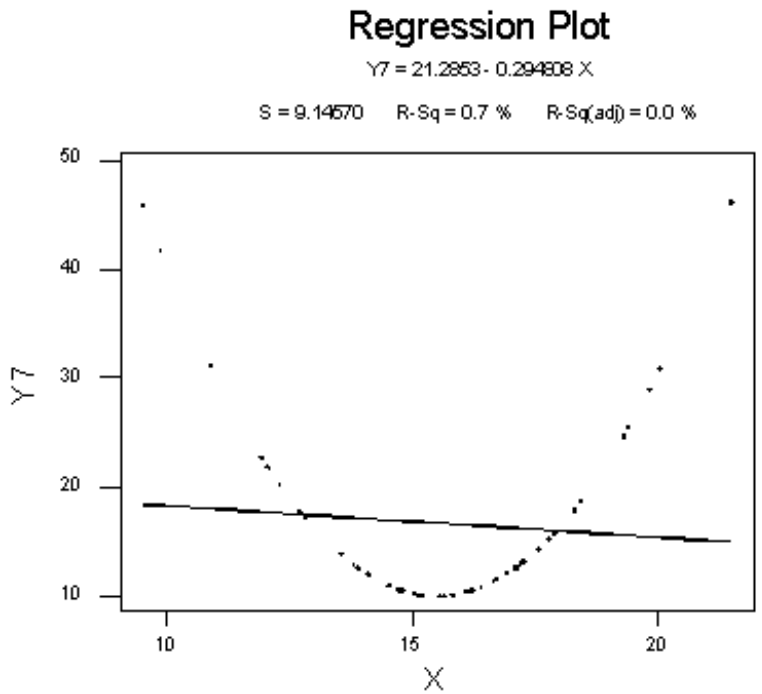
- Misunderstanding 1: A high coefficient of determination indicates that useful predictions can be made.



- Misunderstanding 2: A high coefficient of determination indicates that the estimated regression line is a good fit.



- Misunderstanding 3: A low coefficient of determination indicates that x and y are not related.



Comments:

- Misunderstanding 1 arises because R^2 measures only a relative reduction from SSTO and provides no information about absolute precision for estimating a mean response or predicting a new observation
- Misunderstanding 2 and 3 arise because R^2 measures the degree of linear association between x and y, whereas the actual regression relation may be curvilinear.

Inverse Prediction

The inverse prediction problem is also known as a calibration problem since it is applicable when inexpensive, quick, and approximate measurements (Y) are related to precise, often expensive, and time-consuming measurements (X) based on n observations. The resulting regression model is then used to estimate the precise measurement $X_{h(new)}$ for a new approximate measurement $Y_{h(new)}$. Given Y_h, predict the corresponding value of X, solve the fitted equation for X_h

$$\hat{X}_h = \frac{Y_h - b_0}{b_1}, b_1 \neq 0$$

Example: A medical researcher studied a new, quick method for measuring low concentration of sugar in the blood. Twelve samples were used in the study containing known concentrations (X), with three samples at each of four different levels. The measured concentration (Y) was then observed for each sample. The researcher now wishes to use the regression relation to ascertain the actual concentration $X_{h(new)}$ for a new patient for whom the quick response procedure yielded a measured concentration of $Y_{h(new)} = 6.52$.