Data Analysis
Spring Semester, 2023
March 7, 2023
Lecture 12

Chapter 11
Linear Regression and Correlation

The research context is that two variables have been observed for each of *n* participants. The research team then has a spreadsheet with *n* pairs of observations $(x_i, y_i), i = 1,\ldots,n$. A recommended first step is to create the scatterplot of observations, with the vertical axis representing the dependent variable and the horizontal axis representing the independent variable. The "pencil test" is to hold up a pencil to the scatterplot and examine whether that describes the data well. If so, then it is reasonable to assume that a ***linear model*** (such as $\beta_0 + \beta_1 x$) describes the data. The linear model is reasonable for many data sets in observational studies. A more objective procedure is to use a "nonlinear smoother" such as LOWESS to estimate the association. If the LOWESS curve is not well approximated by a line, then the assumption of linearity is not reasonable.

The OLS estimate of the intercept is $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$, and the estimate of the slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)x_i}{\sum_{i=1}^{n}(x_i - \bar{x}_n)x_i}.$$

There are several modifications of this formula that are helpful. The first results from noting that $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)x_i - \sum_{i=1}^{n}(x_i - \bar{x}_n)\bar{x}_n = \sum_{i=1}^{n}(x_i - \bar{x}_n)x_i$ and $\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n) = \sum_{i=1}^{n}(y_i - \bar{y}_n)x_i - \sum_{i=1}^{n}(y_i - \bar{y}_n)\bar{x}_n = \sum_{i=1}^{n}(y_i - \bar{y}_n)x_i$. The OLS solution is then

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}.$$

The second shows the relation of $\hat{\beta}_1$ and the Pearson product moment correlation. The Pearson product moment correlation is a dimensionless measure of association. The formula is

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

Using the correlation coefficient,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \bullet \sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}} \bullet \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}} = r(x,y) \bullet \frac{\sqrt{(n-1)s_Y^2}}{\sqrt{(n-1)s_X^2}} = \frac{s_Y}{s_X} \bullet r(x,y).$$

The second formula is then $\hat{\beta}_1 = \frac{s_Y}{s_X} \bullet r(x,y)$.

The next variation will be used in calculating the distributional properties of $\hat{\beta}_1$ and uses the identity that

$$\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n) = \sum_{i=1}^{n}[(x_i - \bar{x}_n)y_i] - \sum_{i=1}^{n}[(x_i - \bar{x}_n)\bar{y}_n] = \sum_{i=1}^{n}(x_i - \bar{x}_n)y_i.$$

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

Fisher's Decomposition of the Total Sum of Squares

The total sum of squares of the dependent variable is defined to be $TSS = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$ with $n-1$ degrees of freedom. The *i*th residual was defined above to be $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, i = 1,\ldots,n$. After substituting for $\hat{\beta}_0$, $r_i = y_i - \bar{y}_n - \hat{\beta}_1(x_i - \bar{x}_n), i = 1,\ldots,n$.

Fisher's decomposition is a fundamental tool for the analysis of the linear model. It starts with

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \sum_{i=1}^{n}[y_i - \bar{y}_n - \hat{\beta}_1(x_i - \bar{x}_n) + \hat{\beta}_1(x_i - \bar{x}_n)]^2 = \sum_{i=1}^{n}[r_i + \hat{\beta}_1(x_i - \bar{x}_n)]^2.$$

Next $TSS = \sum_{i=1}^{n}[r_i + \hat{\beta}_1(x_i - \bar{x}_n)]^2 = \sum_{i=1}^{n}[r_i^2 + \hat{\beta}_1^2(x_i - \bar{x}_n)^2 + 2\hat{\beta}_1 r_i(x_i - \bar{x}_n)]$, and

$TSS = \sum_{i=1}^{n} r_i^2 + \sum_{i=1}^{n} \hat{\beta}_1^2(x_i - \bar{x}_n)^2 + 2\hat{\beta}_1 \sum_{i=1}^{n} r_i(x_i - \bar{x}_n)$.

The first sum $\sum_{i=1}^{n} r_i^2 = SSE$, the sum of squared errors, and has $n-2$ degrees of freedom. The second sum $\sum_{i=1}^{n} \hat{\beta}_1^2(x_i - \bar{x}_n)^2$ is called the regression sum of squares and has 1 degree of freedom. It can be simplified:

$$\text{RegSS} = \sum_{i=1}^{n} \hat{\beta}_1^2(x_i - \bar{x}_n)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = [r(x,y)]^2 \left[\frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}} \right]^2 \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \text{ and}$$

$$\text{RegSS} = [r(x,y)]^2 \left[\frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}} \right]^2 \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = [r(x,y)]^2 \sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = [r(x,y)]^2 TSS .$$

Finally, the third sum $2\hat{\beta}_1 \sum_{i=1}^{n} r_i(x_i - \bar{x}_n) = 2\hat{\beta}_1(\sum_{i=1}^{n} r_i x_i - \sum_{i=1}^{n} r_i \bar{x}_n) = 2\hat{\beta}_1(0-0) = 0$.

This is conventionally displayed in an Analysis of Variance Table as below:

Analysis of Variance Table
One Predictor Linear Regression

Source	DF	Sum of Squares	Mean Square	F
Regression	1	$\sum_{i=1}^{n} \hat{\beta}_1^2(x_i - \bar{x}_n)^2 = [r(x,y)]^2 TSS$	$[r(x,y)]^2 TSS$	$\frac{(n-2)[r(x,y)]^2}{1-[r(x,y)]^2}$
Error	$n-2$	$\sum_{i=1}^{n} r_i^2 = \{1-[r(x,y)]^2\}TSS$	$\frac{\{1-[r(x,y)]^2\}TSS}{(n-2)}$	
Total	$n-1$	$TSS = (n-1)s_{DV}^2$		

11.3 Inferences

The model for one predictor linear regression is $Y_i = \beta_0 + \beta_1 x_i + \sigma_{Y|x} Z_i$. The outcome or dependent (random) variables $Y_i, i = 1, \ldots, n$ are each assumed to be the sum of the linear regression expected value $\beta_0 + \beta_1 x_i$ and a random error term $\sigma_{Y|x} Z_i$. The

random variables $Z_i, i = 1,\ldots,n$ are assumed to be independent standard normal random variables. The parameter β_0 is the intercept parameter and is fixed but unknown. The parameter β_1 is the slope parameter and is also fixed but unknown. This parameter is the focus of the statistical analysis. The parameter $\sigma_{Y|x}$ is also fixed but unknown. Another description of this model is that $Y_i, i = 1,\ldots,n$ are independent normally distributed random variables with Y_i having the distribution $N(\beta_0 + \beta_1 x_i, \sigma^2_{Y|x})$. That is, $E(Y_i \mid X = x_i) = \beta_0 + \beta_1 x_i$, and $\text{var}(Y_i \mid X = x_i) = \sigma^2_{Y|x}$. The assumption that $\text{var}(Y_i \mid X = x_i) = \sigma^2_{Y|x}$ is called the *homoscedasticity* assumption.

There are four assumptions. There are two important assumptions: the outcome variables $Y_i, i = 1,\ldots,n$ are independent and $E(Y_i \mid X = x_i) = \beta_0 + \beta_1 x_i$ for $i = 1,\ldots,n$. Homoscedasticity is less important. The assumption that $Y_i, i = 1,\ldots,n$ are normally distributed random variables is least important.

Variance Calculations

More complex calculations are required for the variance-covariance matrix of the OLS estimates. The easiest way is to use the variance-covariance matrix of a random vector. Let Y be an $n \times 1$ vector of random variables $(Y_1, Y_2, \ldots Y_n)^T$. That is, each component of the vector is a random variable. Then the expected value of vector Y is the $n \times 1$ vector whose components are the respective means of the random variables; that is, $E(Y) = (EY_1, EY_2, \ldots EY_n)^T$.

The variance-covariance matrix of the random vector Y is the $n \times n$ matrix whose diagonal entries are the respective variances of the random variables and whose off-diagonal elements are the covariances of the random variables. That is,

$$\text{vcv}(Y) = \begin{bmatrix} \text{var}(Y_1) & \text{cov}(Y_1, Y_2) & \cdots & \text{cov}(Y_1, Y_n) \\ \text{cov}(Y_2, Y_1) & \text{var}(Y_2) & \cdots & \text{cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(Y_n, Y_1) & \text{cov}(Y_n, Y_2) & \cdots & \text{var}(Y_n) \end{bmatrix}.$$

In terms of expectation operator calculations, $\text{vcv}(Y) = E[(Y - EY)(Y - EY)^T] = \Sigma$.

Variance of a Set of Linear Combinations

Let W be the $m \times 1$ random vector of linear combinations of Y given by $W = MY$, where M is a matrix of constants having *m* rows and *n* columns.

Then $E(W) = E(MY) = ME(Y)$,

The definition of the variance-covariance matrix of *W* is

$vcv(W) = E[(W - EW)(W - EW)^T] = E[(MY - MEY)(MY - MEY)^T]$, and
$vcv(W) = E[(MY - MEY)(MY - MEY)^T] = E\{M(Y - EY)[M(Y - EY)]^T\}$

From matrix algebra, when *A* is an $n \times m$ matrix and *B* is an $m \times p$ matrix, then $(AB)^T = B^T A^T$.

Then $[M(Y - EY)]^T = (Y - EY)^T M^T$, and
$\text{vcv}(W) = \text{vcv}(MY) = E\{M(Y - EY)(Y - EY)^T M^T\} = M\{E[(Y - EY)(Y - EY)^T]\}M^T$ from the linear operator property of *E*.

Since $\text{vcv}(Y) = E[(Y - EY)(Y - EY)^T] = \Sigma$, $\text{vcv}(W) = \text{vcv}(MY) = M \times \text{vcv}(Y) \times M^T = M\Sigma M^T$

Examples

The first use of this result is to find the variance of a linear combination of values from *Y*, an *n×1* vector of random variables. Let ***a*** be an *n×1* vector of constants, and let $W = a^T Y$. Then $\text{var}(a^T Y) = a^T \times \text{vcv}(Y) \times (a^T)^T = a^T \times \text{vcv(Y)} \times \text{a}$. This is the completely general form of $\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab \text{cov}(X, Y)$.

The second example is fundamental to this chapter. The OLS estimates of the parameters are always the same functions of the observed data: $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$ and $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x}_n) y_i}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$. It is then reasonable to study the random variables

$$\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{1}{n} Y_1 + \frac{1}{n} Y_2 + \cdots + \frac{1}{n} Y_n$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} = \frac{(x_1 - \bar{x}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} Y_1 + \frac{(x_2 - \bar{x}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} Y_2 + \cdots + \frac{(x_n - \bar{x}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} Y_n .$$

Let $w_i = \dfrac{(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}, i = 1,\ldots,n$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)Y_i}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = w_1Y_1 + w_2Y_2 + \cdots + w_nY_n$$

Then .

Let $\begin{pmatrix} \bar{Y}_n \\ \hat{\beta}_1 \end{pmatrix} = \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ w_1 & w_2 & \cdots & w_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, which has the form MY, where

$M = \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ w_1 & w_2 & \cdots & w_n \end{bmatrix}$.

In the model $Y_i = \beta_0 + \beta_1 x_i + \sigma_{Y|x} Z_i$, $\text{vcv}(Y) = \sigma^2_{Y|x} I_{n\times n}$. Then

$$\text{vcv}\begin{pmatrix} \bar{Y}_n \\ \hat{\beta}_1 \end{pmatrix} = \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ w_1 & w_2 & \cdots & w_n \end{bmatrix} \times \sigma^2_{Y|x} I_{n\times n} \times \begin{bmatrix} 1/n & w_1 \\ 1/n & w_2 \\ \vdots & \vdots \\ 1/n & w_n \end{bmatrix} = \sigma^2_{Y|x} \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ w_1 & w_2 & \cdots & w_n \end{bmatrix} \times \begin{bmatrix} 1/n & w_1 \\ 1/n & w_2 \\ \vdots & \vdots \\ 1/n & w_n \end{bmatrix}.$$

Then,

$$\text{vcv}\begin{pmatrix} \bar{Y}_n \\ \hat{\beta}_1 \end{pmatrix} = \sigma^2_{Y|x} \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ w_1 & w_2 & \cdots & w_n \end{bmatrix} \times \begin{bmatrix} 1/n & w_1 \\ 1/n & w_2 \\ \vdots & \vdots \\ 1/n & w_n \end{bmatrix} = \sigma^2_{Y|x} \begin{bmatrix} \sum_{i=1}^{n}\frac{1}{n^2} & \sum_{i=1}^{n}\frac{w_i}{n} \\ \sum_{i=1}^{n}\frac{w_i}{n} & \sum_{i=1}^{n} w_i^2 \end{bmatrix}.$$

Now $\sum_{i=1}^{n}\frac{w_i}{n} = \sum_{i=1}^{n}\frac{1}{n}\frac{(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\sum_{i=1}^{n}(x_i - \bar{x}_n) = 0$, and

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n}\left[\frac{(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right]^2 = \frac{1}{[\sum_{i=1}^{n}(x_i - \bar{x}_n)^2]^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}.$$

The result is that $\text{vcv}\begin{pmatrix}\bar{Y}_n \\ \hat{\beta}_1\end{pmatrix} = \sigma^2_{Y|x}\begin{bmatrix}\sum_{i=1}^n \frac{1}{n^2} & \sum_{i=1}^n \frac{w_i}{n} \\ \sum_{i=1}^n \frac{w_i}{n} & \sum_{i=1}^n w_i^2\end{bmatrix} = \begin{bmatrix}\frac{\sigma^2_{Y|x}}{n} & 0 \\ 0 & \frac{\sigma^2_{Y|x}}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\end{bmatrix}$

To summarize this result, $\text{var}(\bar{Y}_n) = \frac{\sigma^2_{Y|x}}{n}$, $\text{var}(\hat{\beta}_1) = \frac{\sigma^2_{Y|x}}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$, and

$\text{cov}(\bar{Y}_n, \hat{\beta}_1) = 0$.

Testing a null hypothesis about β_1

The last detail before deriving tests and confidence intervals for the slope of the regression function is to find $E(\hat{\beta}_1)$. Then

$$E(\hat{\beta}_1) = E(\sum_{i=1}^n w_i Y_i) = \sum_{i=1}^n w_i E(Y_i) = \sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i) = \beta_0(\sum_{i=1}^n w_i) + \beta_1(\sum_{i=1}^n w_i x_i).$$

Now, from above, $\sum_{i=1}^n w_i = \frac{1}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \sum_{i=1}^n (x_i - \bar{x}_n) = 0.$

The second sum is

$$\sum_{i=1}^n w_i x_i = \sum_{i=1}^n \frac{(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} x_i = \frac{\sum_{i=1}^n (x_i - \bar{x}_n) x_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = 1.$$

Then $\boldsymbol{E(\hat{\beta}_1) = \beta_1}$.

Under the data model, the distribution of $\hat{\beta}_1$ is $N(\beta_1, \frac{\sigma^2_{Y|x}}{\sum_{i=1}^n (x_i - \bar{x}_n)^2})$.

The key null hypothesis is $H_0 : \beta_1 = 0$, and the alternative hypothesis is $H_1 : \beta_1 \neq 0$.

The test statistic is $\hat{\beta}_1$, and the null distribution is $N(0, \frac{\sigma^2_{Y|x}}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2})$. The standard score form of the statistic is $Z = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\sigma^2_{Y|x}}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}}$. When the level of significance is α and $\sigma^2_{Y|x}$ is known, then $H_0 : \beta_1 = 0$ is rejected when $|Z| \geq |z_{\alpha/2}|$. When $\sigma^2_{Y|x}$ is not known, it is estimated by $\hat{\sigma}^2_{Y|x} = MSE$. This requires the use of the Student's t distribution.

The studentized form of the statistic is $T_{n-2} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}}$. Then, $H_0 : \beta_1 = 0$ is rejected when $|T_{n-2}| \geq |t_{\alpha/2,n-2}|$. An equivalent approach is to use $TS = \frac{MS\ REG}{MSE} = F$. Under $H_0 : \beta_1 = 0$, the null distribution of *F* is a central F with 1 numerator and $n-2$ denominator degrees of freedom.

Confidence interval for β_1

When $\sigma^2_{Y|x}$ is known, the $(1-\alpha)$% confidence interval for β_1 is $\hat{\beta}_1 \pm |z_{\alpha/2}| \sqrt{\frac{\sigma^2_{Y|x}}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}$. When $\sigma^2_{Y|x}$ is not known, the $(1-\alpha)$% confidence interval for β_1 is $\hat{\beta}_1 \pm |t_{\alpha/2,n-2}| \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}$.

Distribution of the Estimated Intercept

Since $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$, $\hat{\beta}_0$ is normally distributed with

$E(\hat{\beta}_0) = E(\bar{Y}_n - \hat{\beta}_1 \bar{x}_n) = E(\bar{Y}_n) - E(\hat{\beta}_1 \bar{x}_n) = E(\bar{Y}_n) - \bar{x}_n E(\hat{\beta}_1) = E(\bar{Y}_n) - \beta_1 \bar{x}_n$ because $E(\hat{\beta}_1) = \beta_1$.

Now $E(\bar{Y}_n) = E(\frac{Y_1 + Y_2 + \cdots + Y_n}{n}) = \frac{E(Y_1) + E(Y_2) + \cdots + E(Y_n)}{n} = \frac{(\beta_0 + \beta_1 x_1) + \cdots + (\beta_0 + \beta_1 x_n)}{n}$,

with $E(\bar{Y}_n) = \frac{(\beta_0 + \beta_1 x_1) + \cdots + (\beta_0 + \beta_1 x_n)}{n} = \frac{n\beta_0 + \beta_1 \sum x_i}{n} = \beta_0 + \beta_1 \bar{x}_n$. Then $E(\hat{\beta}_0) = E(\bar{Y}_n) - \beta_1 \bar{x}_n = \beta_0 + \beta_1 \bar{x}_n - \beta_1 \bar{x}_n = \beta_0$.

Finally,

$$\text{var}(\hat{\beta}_0) = \text{var}(\bar{Y}_n - \hat{\beta}_1 \bar{x}_n) = \text{var}(\bar{Y}_n) + (\bar{x}_n)^2 \text{var}(\hat{\beta}_1) - 2\bar{x}_n \text{cov}(\bar{Y}_n, \hat{\beta}_1) = \frac{\sigma^2_{Y|x}}{n} + \frac{(\bar{x}_n)^2 \sigma^2_{Y|x}}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} - 2 \bullet 0 .$$

In summary, $\hat{\beta}_0$ is $N(\beta_0, \frac{\sigma^2_{Y|x}}{n} + \frac{(\bar{x}_n)^2 \sigma^2_{Y|x}}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2})$

Confidence Interval for $E[\hat{Y}(x)] = \beta_0 + \beta_1 x$

Since $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$,

$$E[\hat{Y}(x)] = E(\hat{\beta}_0 + \hat{\beta}_1 x) = E(\hat{\beta}_0) + xE(\hat{\beta}_1) = \beta_0 + \beta_1 x .$$

For its variance,

$\text{var}[\hat{Y}(x)] = \text{var}[\bar{Y}_n + \hat{\beta}_1 (x - \bar{x}_n)] = \text{var}(\bar{Y}_n) + (x - \bar{x}_n)^2 \text{var}(\hat{\beta}_1) + 2(x - \bar{x}_n)\text{cov}(\bar{Y}_n, \hat{\beta}_1)$, with

$$\text{var}[\hat{Y}(x)] = \frac{\sigma^2_{Y|x}}{n} + \frac{(x - \bar{x}_n)^2 \sigma^2_{Y|x}}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} + 2(x - \bar{x}_n) \bullet 0 = \sigma^2_{Y|x} (\frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}) .$$

In summary, $\hat{Y}(x)$ has the normal distribution $N(\beta_0 + \beta_1 x, \sigma^2_{Y|x} (\frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}))$. When $\sigma^2_{Y|x}$ is known, the 95% confidence interval for $E[\hat{Y}(x)] = \beta_0 + \beta_1 x$ is

$$\hat{Y}(x) \pm 1.960 \sqrt{\sigma^2_{Y|x} (\frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2})} .$$

When $\sigma^2_{Y|x}$ is not known, an estimate of $\sigma^2_{Y|x}$ is used, and the t-percentile is used rather than the z-percentile, here 1.960. If the four assumptions are met, then $E(MSE) = \sigma^2_{Y|x}$.

The 95% confidence interval for $E[\hat{Y}(x)] = \beta_0 + \beta_1 x$ is then

$\hat{Y}(x) \pm t_{1.960,n-2} \sqrt{MSE(\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})}$.

Prediction Interval for a Future Observation $Y_F(x)$

Let $Y_F(x)$ be the future value observed with the independent variable value set to *x*. That is, $Y_F(x)$ is $N(\beta_0 + \beta_1 x, \sigma^2_{Y|x})$. Its distribution is independent of $Y_i, i = 1,\ldots,n$. At a time before $Y_i, i = 1,\ldots,n$ have been observed, $\hat{Y}(x)$ has the normal distribution $N(\beta_0 + \beta_1 x, \sigma^2_{Y|x}(\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}))$. The distribution of $Y_F(x) - \hat{Y}(x)$ is normal. The expected value of $E[Y_F(x) - \hat{Y}(x)] = E[Y_F(x)] - E[\hat{Y}(x)] = \beta_0 + \beta_1 x - (\beta_0 + \beta_1 x) = 0$, and

$\text{var}[Y_F(x) - \hat{Y}(x)] = \text{var}[Y_F(x)] + \text{var}[\hat{Y}(x)] - 2\,\text{cov}[Y_F(x), \hat{Y}(x)] = \sigma^2_{Y|x} + \sigma^2_{Y|x}(\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}) - 2 \bullet 0.$

In summary, $Y_F(x) - \hat{Y}(x)$ is $N(0, \sigma^2_{Y|x}(1 + \frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}))$.

To set up a 99% prediction interval, one starts with

$\Pr\{0 - 2.576\sqrt{\sigma^2_{Y|x}(1 + \frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})} \le Y_F(x) - \hat{Y}(x) \le 0 + 2.576\sqrt{\sigma^2_{Y|x}(1 + \frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})}\} = 0.99.$

Then,

$\Pr\{\hat{Y}(x) - 2.576\sigma_{Y|x}\sqrt{(1 + \frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})} \le Y_F(x) \le \hat{Y}(x) + 2.576\sigma_{Y|x}\sqrt{(1 + \frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})}\} = 0.99.$

Assuming $\sigma^2_{Y|x}$ known, a 99% prediction interval for $Y_F(x)$ is the interval between $\hat{Y}(x) - 2.576\sigma_{Y|x}\sqrt{(1 + \frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})}$ and $\hat{Y}(x) + 2.576\sigma_{Y|x}\sqrt{(1 + \frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})}$.

There are two problems with this interval. The first is that $\hat{Y}(x)$ is a random variable; in other words, it will be observed in the future. The resolution to this is to set the time of the prediction to be after the collection of the regression data but

before the future observation is made. Then the 99% prediction interval is the interval between $\hat{y}(x) - 2.576\sigma_{Y|x}\sqrt{(1+\frac{1}{n}+\frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})}$ and $\hat{y}(x) + 2.576\sigma_{Y|x}\sqrt{(1+\frac{1}{n}+\frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})}$, where $\hat{y}(x)$ is the fitted value based on the regression data using the independent variable setting *x*. The second problem is that $\sigma^2_{Y|x}$ is not known. As usual we estimate $\sigma^2_{Y|x}$ using *MSE* and stretch 2.576 by the t-distribution with $n-2$ degrees of freedom. The 99% prediction interval is the interval between $\hat{y}(x) - t_{2.576,n-2}\sqrt{MSE(1+\frac{1}{n}+\frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})}$ and $\hat{y}(x) + t_{2.576,n-2}\sqrt{MSE(1+\frac{1}{n}+\frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2})}$.

Problem 1 from Chapter 11 Study Guide (revised)

A research team collected data on $n = 450$ students in a statistics course. The observed average final examination score was 524, with an observed standard deviation of 127.6 (the divisor in the estimated variance was $n-1$). The average first examination score was 397, with an observed standard deviation of 96.4. The correlation coefficient between the first examination score and the final examination score was 0.63.

a. Report the analysis of variance table and result of the test of the null hypothesis that the slope of the regression line of final exam score on first exam score is zero against the alternative that it is not. Use the 0.10, 0.05, and 0.01 levels of significance.
b. Determine the least-squares fitted equation and give the 99% confidence interval for the slope of the regression of final examination score on first examination score.
c. Use the least-squares prediction equation to estimate the final examination score of students who scored 550 on the first examination. Give the 99% confidence interval for the expected final examination score of these students.
d. Use the least-squares prediction equation to predict the final examination score of a student who scored 550 on the first examination. Give the 99% prediction interval for the final examination score of this student.

Solution:

For part a, the first task is to identify which variable is dependent and which independent. The question asks for the "regression line of final exam score on first exam score." This phrasing identifies the first exam score as the independent variable and the final exam score as the dependent variable. This also matches the logic of regression analysis. Then, $TSS = (n-1)s_{DV}^2 = 449 \bullet 127.6^2 = 7310510.2$, and $REGSS = [r(DV, IV)]^2 \bullet TSS = (0.63)^2 \bullet 7310510.2 = 2901541.5$. One can obtain *SSE* by subtraction or $SSE = \{1-[r(DV, IV)]^2\} \bullet TSS = [1-(0.63)^2] \bullet 7310510.2 = 4408968.7$. The degrees of freedom for error is $n-2$, and $MSE = \frac{4408968.7}{448} = 9841.4$. Then $F = \frac{MS\ REG}{MSE} = \frac{2901541.5}{9841.4} = 294.8$ with (1, 448) degrees of freedom. These values are conventionally displayed in the Analysis of Variance Table below:

Analysis of Variance Table

Problem 1

	DF	SS	MS	F
Reg.	1	2901541.5	2901541.5	294.8
Res.	448	4408968.7	9841.4	
Total	449	7310510.2		

For $\alpha = 0.10$, the critical value of an F distribution with (1, 448) degrees of freedom is 2.717; for $\alpha = 0.05$ the critical value is 3.862; and for $\alpha = 0.01$ the critical value is 6.692. Reject the null hypothesis that the slope of the regression line is zero at the 0.01 level of significance (and also at the 0.05 and 0.10 levels).

For part b, $\hat{\beta}_1 = \frac{s_Y}{s_X} \bullet r(x, y) = \frac{127.6}{96.4} \bullet 0.63 = 0.834.$ The intercept is $\hat{\beta}_0 = 524 - 0.834 \bullet 397 = 192.9$, so that $\hat{Y}(x) = 192.9 + 0.834x$. The 99% confidence interval for the slope is $\hat{\beta}_1 \pm |t_{\alpha/2, n-2}| \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}} = 0.834 \pm 2.587 \sqrt{\frac{9841.4}{(n-1)s_{IV}^2}} = 0.834 \pm 2.587 \sqrt{\frac{9841.4}{449 \bullet (96.4)^2}}$. This is the interval (0.71,0.96).

For part c, the 99% confidence interval for $E(Y \mid x = 550)$ is centered on $\hat{Y}(550) = \hat{\beta}_0 + \hat{\beta}_1 550 = 192.9 + 0.834 \bullet 550 = 651.6$. The 99% confidence interval is $\hat{Y}(550) \pm t_{2.576, n-2} \sqrt{MSE\left(\frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right)} = 651.6 \pm 2.587 \sqrt{9841.4\left(\frac{1}{450} + \frac{(550-397)^2}{449 \bullet (96.4)^2}\right)}$.

This is

$$651.6 \pm 2.587\sqrt{9841.4(0.002222 + \frac{23409}{4172539.04})} = 651.6 \pm 2.587\sqrt{9841.4(0.002222 + 0.005610)},$$

which reduces to $651.6 \pm 22.7,$ which is the interval (628.9, 674.3) .

Part d specifies the prediction interval for the final exam score of a student whose first exam score was 550. The center of the prediction interval is still $\hat{y}(550) = 651.6.$

The prediction interval is

$$\hat{y}(x) \pm t_{2.576,n-2}\sqrt{MSE(1 + \frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2})}$$

This is

$$\hat{y}(x) \pm t_{2.576,n-2}\sqrt{MSE(1 + \frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2})} = 651.6 \pm 2.587\sqrt{9841.4(1 + 0.002222 + 0.005610}$$

.

This reduces to

$$651.6 \pm 2.587 \bullet 99.20\sqrt{(1 + 0.002222 + 0.005610)} = 651.6 \pm 2.587 \bullet 99.20\sqrt{1.00783} = 651.6 \pm 257.6$$

The 99% prediction interval is (394.0, 909.23).

Problem 7, Study Guide for Chapter 11

The correlation matrix of the random variables Y_1, Y_2, Y_3, Y_4 is $\begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$, $0 < \rho < 1$, and each random variable has variance σ^2. Let $W_1 = Y_1 + Y_2 + Y_3$, and let $W_2 = Y_2 + Y_3 + Y_4$. Find the variance covariance matrix of (W_1, W_2).

Solution: The solution requires the application of the result that

$\text{vcv}(W) = \text{vcv}(MY) = M \times \text{vcv}(Y) \times M^T$ where $\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix}$. That is,

$M_{2\times 4} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$, with

$$M \times \text{vcv}(Y) = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \times \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1+2\rho & 1+2\rho & 1+2\rho & 3\rho \\ 3\rho & 1+2\rho & 1+2\rho & 1+2\rho \end{bmatrix}.$$

Then

$$M \times \text{vcv}(Y) \times M^T == \sigma^2 \begin{bmatrix} 1+2\rho & 1+2\rho & 1+2\rho & 3\rho \\ 3\rho & 1+2\rho & 1+2\rho & 1+2\rho \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3+6\rho & 2+7\rho \\ 2+7\rho & 3+6\rho \end{bmatrix} \sigma^2.$$

That is, $\text{var}(W_1) = \text{var}(W_2) = (3+6\rho)\sigma^2$, and $\text{cov}(W_1, W_2) = (2+7\rho)\sigma^2$.

Fisher's Transformation of the Correlation Coefficient (Section 11.6)

Your text uses Fisher's transformation of the correlation coefficient to get a confidence interval for a correlation coefficient. It is more useful in calculating Type II error rates and sample size calculations. The transformation is applied to the Pearson product moment correlation coefficient R_{xy} calculated using *n* observations (X_i, Y_i) from a bivariate normal random variable with population correlation coefficient $\rho = \text{corr}(X, Y)$. Fisher's result is that $F(R_{xy}) = \frac{1}{2}\ln(\frac{1+R_{xy}}{1-R_{xy}})$ is approximately distributed as $N(\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}), \frac{1}{n-3})$.

Confidence Interval for a Correlation Coefficient

The 99% confidence interval for $F(\rho) = \frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$ is $F(R_{xy}) \pm 2.576\sqrt{\frac{1}{n-3}}$. Readers, however, want to know the confidence interval for $\rho = \text{corr}(X, Y)$.

This requires solving for R_{xy} as a function of $F(R_{xy}) = \frac{1}{2}\ln(\frac{1+R_{xy}}{1-R_{xy}})$.

The solution requires some algebra: $2F(R_{xy}) = \ln(\frac{1+R_{xy}}{1-R_{xy}})$ so that

$\exp[2F(R_{xy})] = \exp[\ln(\frac{1+R_{xy}}{1-R_{xy}})] = \frac{1+R_{xy}}{1-R_{xy}}$.

Using the first and third parts of the equation,

$(1-R_{xy})\exp[2F(R_{xy})] = 1+R_{xy}$.

Putting R_{xy} on one side of the equation yields

$\exp[2F(R_{xy})] - 1 = (1+\exp[2F(R_{xy})])R_{xy}.$

Solving for R_{xy},

$$R_{xy} = \frac{\exp[2F(R_{xy})]-1}{\exp[2F(R_{xy})]+1}.$$

Modification of problem 1 above:

A research team collected data on $n = 450$ students in a statistics course. The correlation coefficient between the first examination score and the final examination score was 0.63. Find the 99% confidence interval for the population correlation of the first examination score and the final examination score.

Solution: Fisher's transformation of the observed correlation is

$F(0.63) = \frac{1}{2}\ln(\frac{1+0.63}{1-0.63}) = \frac{1}{2}\ln(\frac{1.63}{0.37}) = \frac{1}{2}\ln(4.405) = \frac{1}{2}(1.483) = 0.741.$

Since the sampling margin of error is

$$2.576\sqrt{\frac{1}{n-3}} = 2.576\sqrt{\frac{1}{450-3}} = 0.122,$$

the

99% confidence interval for $F(\rho) = \frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$ is

$0.741 \pm 0.122,$

which is the interval from 0.619 to 0.863.

Using the inversion formula $R_{xy} = \frac{\exp[2F(R_{xy})]-1}{\exp[2F(R_{xy})]+1}$,

the left endpoint of the confidence interval is

$$\frac{\exp[2\times 0.619)]-1}{\exp[2\times 0.619]+1}=\frac{3.449-1}{3.449+1}=\frac{2.449}{4.449}=0.55.$$

The right endpoint is

$$\frac{\exp[2\times 0.863)]-1}{\exp[2\times 0.863]+1}=\frac{5.618-1}{5.618+1}=\frac{4.618}{6.618}=0.70.$$

Even with 450 observations, the 99% confidence interval for

$\rho = \text{corr}(X, Y)$

is rather wide: from 0.55 to 0.70.

Example Sample Size Calculations

A research team wishes to test the null hypothesis $H_0 : \rho = 0$ at $\alpha = 0.005$ against the alternative $H_1 : \rho > 0$ using the Fisher's transformation of the Pearson product moment correlation coefficient R_{xy} as the test statistic. They have asked their consulting statistician for a sample size *n* such that $\beta = 0.01$ when $\rho = 0.316$ (that is, $\rho^2 = 0.10$).

Solution: The null distribution of Fisher transformation of the Pearson product moment correlation R_{xy} is approximately $N(\frac{1}{2}\ln(\frac{1+0}{1-0}), \frac{1}{n-3})$, which is $N(0, \frac{1}{n-3})$. The null hypothesis $H_0 : \rho = 0$ is rejected at $\alpha = 0.005$ against the alternative $H_1 : \rho > 0$ when $F(R_{xy}) \geq 0 + 2.576\sqrt{\frac{1}{n-3}}$. For the alternative specified, the test statistic (Fisher's transformation of the Pearson product moment correlation) is approximately $N(\frac{1}{2}\ln(\frac{1+0.316}{1-0.316}), \frac{1}{n-3})$. Since $\frac{1}{2}\ln(\frac{1+0.316}{1-0.316}) = \frac{1}{2}\ln(1.924) = 0.327$, the approximate alternative distribution is $N(0.327, \frac{1}{n-3})$ A useful fact for checking your work is that $\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}) \cong \rho$ for small values of ρ. The probability of a Type II

error is $\beta = \Pr_1\{\text{Accept } H_0\} = \Pr_1\{F(R_{xy}) < 0 + 2.576\sqrt{\frac{1}{n-3}}\}$. Then

$$\beta = \Pr_1\{F(R_{xy}) < 0 + 2.576\sqrt{\frac{1}{n-3}}\} = \Pr\{\frac{[F(R_{xy}) - E_1(F(R_{xy})]}{\sigma_1(F(R_{xy}))} < \frac{0 + 2.576\sqrt{\frac{1}{n-3}} - 0.327}{\sqrt{\frac{1}{n-3}}}\}.$$

As before, to select *n* so that $\beta = 0.01 = \Pr\{Z \le -2.326\} = \Phi(-2.326)$ requires that $\frac{0 + 2.576\sqrt{\frac{1}{n-3}} - 0.327}{\sqrt{\frac{1}{n-3}}} = -2.326$. This is essentially the same setup for sample size calculations in the one and two sample problems so that the solution follows the same steps: $0 + 2.576\sqrt{\frac{1}{n-3}} - 0.327 = -2.326\sqrt{\frac{1}{n-3}}$,

$$2.576\sqrt{\frac{1}{n-3}} + 2.326\sqrt{\frac{1}{n-3}} = 0.327 - 0, \text{ and}$$

$$\sqrt{n-3} \ge \frac{2.576\sqrt{1} + 2.326\sqrt{1}}{|0.327 - 0|} = 14.991. \text{ That is, } n \ge 228.$$

Researchers need on the order of 230 observations to detect reliably (that is, two sided level of signficance $\alpha = 0.01$, $\beta = 0.01$)an association that explains 10% of the variation of the dependent variable.

This result is easily generalized. The fundamental design equation then gives us that $\sqrt{n-3} \ge \frac{|z_\alpha|\sigma_0 + |z_\beta|\sigma_1}{|E_1 - E_0|}$, where $|z_\alpha| = 2.576$, $|z_\beta| = 2.326$, $\sigma_0 = \sigma_1 = 1$, $E_1 = F(\rho_1)$, and $E_0 = 0$.

Examining Lack of Fit in Linear Regression (Secton 11.5)

We will not cover this until we get to Chapter 9. You are not responsible for problems like number 5 in the Chapter 11 Study Guide for the first mid-term.

Project 1

Data is available on the Class Blackboard.

- Choose your software—see file "Obtaining Statistical Computing Resources" in assignment section of class blackboard.

Choices of software for computing project:

- R is free and will increase in prevalence as time goes on, but it requires some computing sophistication.

- SAS is popular in the pharmaceutical industry and in banking. Federal regulators insist on SAS programming. It is not a forgiving package and requires some computing sophistication.
- Minitab is user friendly, menu driven, and well documented. It has many quite sophisticated techniques.
- SPSS is user friendly, menu driven, and well documented. It is popular in market research and social science related industries.
- Excel does not have a good reputation for statistical work at this point. It is very good at data processing issues. Since it is extremely popular, it is to your professional advantage to know the basics of the problem. Most students in past semesters have used Excel to merge the data files for the first part of the first project.

Data Processing

- Do your editing using computer programs rather than cut and paste or deleting rows with missing data one by one. Also, I find it helpful to add the date of your editing to the file name.
- Make sure that you understand how your statistical package will deal with missing data. That is, check whether your package has a default option of listwise deletion.
- I recommend that you set a plateau of creating a data file. Excel is popular for this. Many students have found the VLOOKUP function in Excel to be helpful in merging the data sets. Create files with a relatively small number of cases with the problems that you are dealing with and work to create software that handles the problem. Then use the perfected software on your larger files.
- Do an internet search on "missing data" and the package that you wish. Also the package may have an internet site with information about how to deal with problems. For example, Minitab has documentation on getting started. The help menu has an option for "minitab on the web." One can choose that option and enter "missing data" in the search area. It will return a selection that is titled "Remove rows with missing values." This is not an acceptable choice for dealing with missing data.
- Check your work. Make sure that the final data file is correct. The grading in part A of project 1 is largely determined by whether you have the correct number of cases.
- Do not let data processing issues stop you from starting the analysis of part B.

Analyzing data:

- Calculate summary statistics on each variable (i.e., mean, standard deviation, median, quartile points, maximum, minimum, number of cases). The histogram of the variable may also be helpful.
- Examine the scatterplot.
- Calculate the bivariate statistics (i.e., correlation coefficients). Most of the correlations that one sees in practice are minimal in size. Consequently, a small fraction of the part B data sets have dependent variables that have no association with the independent variable.
- Calculate the Chapter 11 regression statistics. Plot the residuals against the predicted variable values. One hopes for a patternless residual plot. If not, then try transformations of both the independent and dependent variables. If there are repeated values of the independent variable, one can perform a lack of fit test. We will discuss this further in our work on Chapters 8 and 9.

Writing the report

- You must have a report. If you hand in a lot of computer output with no explanation, you will get 0 points.
- The classic format for scientific reports is: 1. Statement of problem and introduction; 2. Methods; 3. Results; 4. Conclusions and discussion.
- Do not plagiarize—either from the example report or from your fellow students. It is acceptable to quote another paper or book provided that you put quotation marks around the quoted material and identify the source. Cutting and pasting material from someone else's paper that has errors in it will be obvious in grading.