Data Analysis
Spring Semester, 2023
March 9, 2023
Lecture 13

Second Midterm on Thursday, March 30. It will focus on Chapters 11 and 12.

Chapter 11
Linear Regression and Correlation

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

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

An equivalent formula for the OLS slope is

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

The Pearson product moment correlation is a dimensionless measure of association. The formula is

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

Another formula for the OLS estimate of the slope is

$$\hat{\beta}_1 = \frac{s_Y}{s_X} \bullet r(x,y).$$

A final formula is $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x}_n) y_i}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$

Fisher's Decomposition of the Total Sum of Squares

Fisher's decomposition is a fundamental tool for the analysis of the linear model. 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	$n-1$	$TSS=(n-1)s_{DV}^2$		

Variance Calculations

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

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

Examples

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

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

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

Then,

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

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

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

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

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 not known, the $(1-\alpha)$% confidence interval for β_1 is

$$\hat{\beta}_1 \pm |\, t_{\alpha/2,n-2} \,| \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}} \, .$$

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,

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

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

In summary, $\hat{Y}(x)$ has the normal distribution $N\left(\beta_0 + \beta_1 x, \sigma^2_{Y|x}\left(\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}\right)\right)$.

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

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

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

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

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

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

$$\hat{Y}(550) \pm t_{2.576, n-2}\sqrt{MSE(\frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2})} = 651.6 \pm 2.587\sqrt{9841.4(\frac{1}{450} + \frac{(550-397)^2}{449 \bullet (96.4)^2})}$$

.

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 $\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = [\begin{matrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{matrix}] \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix}$. That is,

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

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

Then

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

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

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

Chapter 12
Multiple Regression and the General Linear Model

The research context is that two or more independent variables and one dependent variable have been observed for each of *n* participants. Here, I will discuss two independent variables x_{1i} and $x_{2i}, i = 1, \ldots, n$ and one dependent variable $y_i, i = 1, \ldots, n.$ The mathematics and analysis for more independent variables generalize routinely. The research team then has a spreadsheet with *n* vectors of observations $(x_{1i}, x_{2i}, y_i), i = 1, \ldots, n$. As in Chapter 11, one of the variables (here *y*) is the outcome variable or dependent variable. This is the variable hypothesized to be affected by the other variables in scientific research. The other variables (here x_{1i} and $x_{2i}, i = 1, \ldots, n$) are the independent variables. They may be hypothesized to predict the outcome variable or to cause a change in the outcome variable. The research task is to document the association between independent and dependent variables.

As before, a recommended first step is to create the scatterplots of observations, with the vertical axis representing the dependent variable and the horizontal axis representing one of the independent variables. The "pencil test" can be used again. If the plot passes this test, then it is reasonable to assume that a ***linear model*** (such as $\beta_0 + \beta_1 x_1 + \beta_2 x_2$) describes the data. The linear model is reasonable for many data sets in observational studies. Specifically, the model for Chapter 12 is $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_{Y\bullet 12} Z_i$. The parameters $(\beta_0, \beta_1, \beta_2)$ are fixed but unknown. The parameter $\sigma_{Y\bullet 12}$ is the unknown conditional standard deviation of Y_i controlling for x_{1i} and $x_{2i}, i = 1, \ldots, n$. The standard deviation of Y_i is assumed to be equal for each observation. The random errors Z_i are assumed to be independent. The independence of the random errors (and hence independence of Y_i) is important. The assumption of a linear regression function (that is, $E(Y_i \mid x_{1i}, x_{2i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$) is also important. As in Chapter 11, this is equivalent to the joint distribution of the dependent variable values being $NID(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \sigma^2_{Y\bullet 12})$.

Estimating the Linear Model Parameters

Again, OLS (ordinary least squares) is the most commonly used method to estimate the parameters of the linear model. A linear model with arbitrary arguments $b_0 + b_1 x_1 + b_2 x_2$ is used as a *fit* for the dependent variable values. The method uses the *residual* $y_i - b_0 - b_1 x_{1i} - b_2 x_{2i}$. As in Chapter 11, the fitting model is judged by how small the set of residuals is. Here OLS minimizes the sum of squares function $SS(b_0, b_1, b_2) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2$. The OLS method is to find the arguments $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ that make $SS(b_0, b_1, b_2)$ as small as possible. This minimization is a standard calculus problem. Step 1 is to calculate the partial derivatives of $SS(b_0, b_1, b_2)$ with respect to each argument. First, the partial with respect to b_0:

$$\frac{\partial SS(b_0, b_1, b_2)}{\partial b_0} = \frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2 = \sum_{i=1}^{n} \frac{\partial}{\partial b_0} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2$$

$$= \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i}) \frac{\partial (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})}{\partial b_0}.$$

Simplifying, $\frac{\partial SS(b_0, b_1, b_2)}{\partial b_0} = \sum_{i=1}^{n} (-2)(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})$.

Second, the partial with respect to b_1:

$$\frac{\partial SS(b_0, b_1, b_2)}{\partial b_1} = \frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2 = \sum_{i=1}^{n} \frac{\partial}{\partial b_1} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2$$

$$= \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i}) \frac{\partial (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})}{\partial b_1}.$$

Then, $\frac{\partial SS(b_0, b_1, b_2)}{\partial b_1} = \sum_{i=1}^{n} (-2)(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i}) x_{1i}$

Third, the partial with respect to b_2 is parallel, with the result that

$$\frac{\partial SS(b_0, b_1, b_2)}{\partial b_2} = \sum_{i=1}^{n} (-2)(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i}) x_{2i}$$

Step 2 is to find the arguments $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ that make the three partial derivatives simultaneously zero. The resulting equations are still called the *normal equations*:

$$\sum_{i=1}^{n} (-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) = 0,$$

$$\sum_{i=1}^{n} (-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) x_{1i} = 0, \text{ and}$$

$$\sum_{i=1}^{n} (-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) x_{2i} = 0, .$$

These equations still have a very important mathematical interpretation. Let $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}, i = 1, \ldots, n$. The first normal equation is equivalent to $\sum_{i=1}^{n} r_i = 0$; the second is $\sum_{i=1}^{n} r_i x_{1i} = 0$; and the third is $\sum_{i=1}^{n} r_i x_{2i} = 0$ That is, there are three constraints on the *n* residuals. The OLS residuals must sum to zero, and the OLS residuals are orthogonal to the two independent variable values. The *n* residuals then have $n-3$ degrees of freedom.

Step 3 is to solve this three linear equation system in three unknowns. There is a more general approach to solving systems like this. The first equation is $\sum_{i=1}^{n}(-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) = 0,$ which can be written $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})$ and $\sum_{i=1}^{n}(1 \times y_i) = [\sum_{i=1}^{n}(1 \times 1)]\hat{\beta}_0 + [\sum_{i=1}^{n}(1 \times x_{1i})]\hat{\beta}_1 + [\sum_{i=1}^{n}(1 \times x_{2i})]\hat{\beta}_2$. Similarly, the second normal equation can be written $\sum_{i=1}^{n}(x_{1i} \times y_i) = [\sum_{i=1}^{n}(1 \times x_{1i})]\hat{\beta}_0 + [\sum_{i=1}^{n}(x_{1i} \times x_{1i})]\hat{\beta}_1 + [\sum_{i=1}^{n}(x_{1i} \times x_{2i})]\hat{\beta}_2$; and the third $\sum_{i=1}^{n}(x_{2i} \times y_i) = [\sum_{i=1}^{n}(1 \times x_{2i})]\hat{\beta}_0 + [\sum_{i=1}^{n}(x_{2i} \times x_{1i})]\hat{\beta}_1 + [\sum_{i=1}^{n}(x_{2i} \times x_{2i})]\hat{\beta}_2$.

While these look like complicated equations, matrix algebra leads to a simpler expression. Let $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ be the $n \times 1$ column vector of dependent variable values, and let $X = [\begin{matrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{matrix}]$ be the $n \times 3$ matrix of coefficients of the parameters

$(\beta_0, \beta_1, \beta_2)$. From matrix algebra, $X^T X = \begin{bmatrix} \sum_{i=1}^{n}(1\times 1) & \sum_{i=1}^{n}(1\times x_{1i}) & \sum_{i=1}^{n}(1\times x_{2i}) \\ \sum_{i=1}^{n}(1\times x_{1i}) & \sum_{i=1}^{n}(x_{1i}\times x_{1i}) & \sum_{i=1}^{n}(x_{1i}\times x_{2i}) \\ \sum_{i=1}^{n}(1\times x_{2i}) & \sum_{i=1}^{n}(x_{1i}\times x_{21}) & \sum_{i=1}^{n}(x_{2i}\times x_{2i}) \end{bmatrix}$, and

$$X^T Y = \begin{bmatrix} \sum_{i=1}^{n}(1\times y_i) \\ \sum_{i=1}^{n}(x_{1i}\times y_i) \\ \sum_{i=1}^{n}(x_{2i}\times y_i) \end{bmatrix}.$$

Recall the three normal equations above:

$$\sum_{i=1}^{n}(1\times y_i) = [\sum_{i=1}^{n}(1\times 1)]\hat{\beta}_0 + [\sum_{i=1}^{n}(1\times x_{1i})]\hat{\beta}_1 + [\sum_{i=1}^{n}(1\times x_{2i})]\hat{\beta}_2$$

$$\sum_{i=1}^{n}(x_{1i}\times y_i) = [\sum_{i=1}^{n}(1\times x_{1i})]\hat{\beta}_0 + [\sum_{i=1}^{n}(x_{1i}\times x_{1i})]\hat{\beta}_1 + [\sum_{i=1}^{n}(x_{1i}\times x_{2i})]\hat{\beta}_2$$

$$\sum_{i=1}^{n}(x_{2i}\times y_i) = [\sum_{i=1}^{n}(1\times x_{2i})]\hat{\beta}_0 + [\sum_{i=1}^{n}(x_{1i}\times x_{2i})]\hat{\beta}_1 + [\sum_{i=1}^{n}(x_{2i}\times x_{2i})]\hat{\beta}_2$$

The left-hand side terms are the same as the terms of $X^T Y$, and the coefficients of the OLS estimators match with the terms of $X^T X$. For this problem, then, the normal equations can be written in matrix form as $(X^T X)\hat{\beta} = X^T Y$. This result also holds for three or more independent variables. The proof is exactly the same as for the two independent variable case.

If $(X^T X)^{-1}$ exists, then $\hat{\beta} = (X^T X)^{-1} X^T Y$. The existence of $(X^T X)^{-1}$ is the usual case in observational studies using multiple regression. If $(X^T X)^{-1}$ does not exist, then the OLS estimators exist but are not unique.

Distribution of $\hat{\beta} = (X^T X)^{-1} X^T Y$

Let $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$ be the vector of the random outcome variables. That is, the data will be collected in the future as opposed to having the data in hand as we assumed in our OLS estimator derivation. The probabilistic model for the data can be written in matrix form $Y = X\beta + \sigma_{Y\bullet 12} Z$, where Z is the column vector of random errors Z_i that are assumed to be independent. From now on, we consider the general case with

$p-1$independent variables. The vector β of parameters in the regression function is then $p\times 1$, remembering that there is an intercept term in our model. The matrix X of coefficients of the parameters of the regression function is now $n\times p,\ p<n,$ with rank p. Then $E(Y)=E(X\beta+\sigma_{Y\bullet 12}Z)=E(X\beta)+E(\sigma_{Y\bullet 12}Z)=X\beta+\sigma_{Y\bullet 12}E(Z)=X\beta$, and $\mathrm{vcv}(Y)=\sigma^2_{Y\bullet 12}I_{n\times n}$. An equivalent description is to say that Y is multivariate normal with dimension n; that is, Y has the distribution $MVN_n(X\beta,\sigma^2_{Y\bullet 12}I_{n\times n})$.

When the matrix X has rank p, $(X^TX)^{-1}$exists. Then the vector of OLS estimators is $\hat{\beta}=(X^TX)^{-1}X^TY$. The expected value is given by
$E(\hat{\beta})=E[(X^TX)^{-1}X^TY]=(X^TX)^{-1}X^TE(Y)=(X^TX)^{-1}X^T(X\beta)=[(X^TX)^{-1}(X^TX)]\beta=I_{p\times p}\beta=\beta.$
The variance-covariance matrix of $\hat{\beta}=(X^TX)^{-1}X^TY$ is calculated by
$\mathrm{vcv}(\hat{\beta})=\mathrm{vcv}[(X^TX)^{-1}X^TY]=(X^TX)^{-1}X^T\mathrm{vcv}(Y)[(X^TX)^{-1}X^T]^T$. This can be simplified using the matrix algebra result that $(AB)^T=B^TA^T$ so that
$[(X^TX)^{-1}X^T]^T=(X^T)^T[(X^TX)^{-1}]^T=X(X^TX)^{-1}$.

Recall that the transpose of the transpose of a matrix is just the matrix so that $(X^T)^T=X$. Further, a matrix is symmetric if its transpose is the matrix itself. That is, $(X^TX)^T=X^T(X^T)^T=X^TX$. The inverse of a symmetric matrix is symmetric so that $[(X^TX)^{-1}]^T=(X^TX)^{-1}$. Using these results in
$\mathrm{vcv}(\hat{\beta})=(X^TX)^{-1}X^T\mathrm{vcv}(Y)[(X^TX)^{-1}X^T]^T=(X^TX)^{-1}X^T\sigma^2_{\mathrm{Y}\bullet\mathrm{x}}I_{n\times n}X(X^TX)^{-1}$,
$\mathrm{vcv}(\hat{\beta})=(X^TX)^{-1}X^T\sigma^2_{\mathrm{Y}\bullet\mathrm{x}}I_{n\times n}X(X^TX)^{-1}=\sigma^2_{\mathrm{Y}\bullet\mathrm{x}}\{(X^TX)^{-1}[X^TX]\}(X^TX)^{-1}=\sigma^2_{\mathrm{Y}\bullet\mathrm{x}}\{I_{p\times p}\}(X^TX)^{-1}=\sigma^2_{\mathrm{Y}\bullet\mathrm{x}}(X^TX)^{-1}.$
The distribution of $\hat{\beta}=(X^TX)^{-1}X^TY$ is $MVN_p(\beta,\sigma^2_{\mathrm{Y}\bullet\mathrm{x}}(X^TX)^{-1})$.