**AMS 315**
**Data Analysis**
Chapter Eleven Study Guide
Linear Regression and Correlation
Spring 2023

## Context

The procedures in this chapter provide tools to deal with the association of one continuous dependent variable and an independent variable.

## Chapter Eleven

### 11.1. Introduction and Abstract of Research Study

This section contains a statement of the linear regression model and key definitions. It is an important section.

### 11.2 Estimating Model Parameters

Definition 11.3, the formula for the mean square residual, and the decomposition of the total sums of squares are fundamentally important.

### 11.3 Inferences about the Regression Parameters

Tests are typically applied to the null hypothesis that the slope of the regression line is zero. Study the box on the t-test that the slope is zero, the F-test that the slope is zero, and the confidence interval for the slope.

### 11.4 Predicting New *y* Values Using Regression

Make sure that you understand the difference between the confidence interval for $E(Y_{n+1})$ and the prediction interval for $Y_{n+1}$.

### 11.5 Examining Lack of Fit in Linear Regression

This is a useful procedure. The point is that good study design allows one to test objectively whether a model appears to describe the data collected. Specifically, by including multiple settings of selected values of the independent variable, a research team can test whether the assumptions make in the regression analysis appear reasonable.

## 11.6. Correlation

In a study of *n* units of observation $(x_i, y_i)$, the Pearson product moment correlation coefficient is defined to be $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$. Your text defines it through the coefficient of determination, $r_{xy}^2$, which is the fraction of variance explained by the regression of *y* on *x*. That is, when $SS(Total) = \sum_{i=1}^{n}(y_i - \bar{y})^2$, then $SS(\text{Re}\,gression) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = r_{xy}^2 SS(Total)$. Then $SS(Error) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (1 - r_{xy}^2)SS(Total)$. Most of my students find these formulas are more easily used. That is, the F-test for the slope reduces to $F = \frac{(n-2)r_{xy}^2}{(1 - r_{xy}^2)}$ with 1 numerator and $n-2$ denominator degrees of freedom.

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 calculated using *n* observations $(X_i, Y_i)$ from a bivariate normal random variable with correlation coefficient $\rho = corr(X,Y)$. The result is that $Z = \frac{1}{2}\ln(\frac{1+R_{xy}}{1-R_{xy}})$ is approximately distributed as $N(\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}), \frac{1}{n-3})$.

Example Examination Question

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 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 is approximately $N(\frac{1}{2}\ln(\frac{1+0}{1-0}), \frac{1}{n-3})$, which is $N(0, \frac{1}{n-3})$. 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 $\rho$. 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 $|z_\alpha| = 2.576$, $|z_\beta| = 2.326$, $\sigma_0 = \sigma_1 = 1$, $E_1 = 0.327$, and $E_0 = 0$. That is, $\sqrt{n-3} \geq \frac{2.576 + 2.326}{0.327 - 0} = 14.991$. That is, $n \geq 228$. That is, 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.

**11.7. Research Study: Two Methods for Detecting E.coli**

There will be no formal questions on this section. It does describe well the practical issues occurring in a regression study.

**11.8. Summary and Key Formulas**

This is a valuable summary.

**Supplemental Information on Variance-covariance Calculations**

One simple formula from basic probability theory and its generalizations are crucial. The simple formula is

$$\operatorname{var}(aX + bY) = a^2 \operatorname{var} X + b^2 \operatorname{var} Y + 2ab \operatorname{cov}(X, Y).$$

More complex calculations are often necessary. The easiest way to get answers 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 respective components are the 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,

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

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

The first use of this result is to find the variance of a linear combination of values from **Y**, an $n\times 1$ vector of random variables. Let **a** be an $n\times 1$ vector of constants. Then

$$\text{var}(a^T Y) = a^T vcv(Y)a.$$

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 $vcv(W) = M\Sigma M^T$ , as derived in class.

## Example Past Examination Questions

1. A research team collected data on $n = 450$ students in a statistics course. Their dependent variable was the student's score on the final examination, which ranged from 200 to 800 points. The observed average final examination score was 524, with an observed standard deviation of 127.6 (the divisor in the estimated variance was $n-1$). Their independent variable was the score on the first examination in the course, which also ranged from 200 to 800. The average 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 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 a student who scored 550 on the first examination. Give the 99% confidence interval for the expected final examination score of this student.

Answers:

a. Analysis of Variance Table

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

Reject the null hypothesis that the slope of the regression line is zero at the 0.01 level of significance.

b. $\hat{Y}(x) = 192.9 + 0.834x$ , and the 99% confidence interval for the slope is (0.71,0.96)

c. The 99% confidence interval for $E(Y \mid x = 550)$ is (628.9, 674.3) .

2. A research team collected data on $n = 250$ students in a statistics course. Their dependent variable was the student's score on the final examination, which ranged from 200 to 800 points. The observed average final examination score was 467, with an observed standard deviation of 107.2 (the divisor in the estimated variance was $n-1$). Their independent variable was the score on the first

examination in the course, which also ranged from 200 to 800. The average was 424, with an observed standard deviation of 81.7. The correlation coefficient between the first examination score and the final examination score was 0.58.

a. Report the analysis of variance table and the test of the null hypothesis that the slope of the regression line 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 predict the final examination score of a student who scored 550 on the first examination. Give a 99% prediction interval for this student's final examination score.

Answers:

a. Analysis of Variance Table

| | DF | SS | MS | F |
|---|---|---|---|---|
| Reg. | 1 | 962,597.9 | 962,597.9 | 125.7 |
| Res. | 248 | 1,898,870.3 | 7656.7 | |
| Total | 249 | 2,861,468.2 | | |

Reject the null hypothesis that the slope of the regression line is zero at the 0.01 level of significance.

b. $\hat{Y}(x) = 144.3 + 0.761x$, and the 99% confidence interval for the slope is (0.59,0.94)

c. The 99% prediction interval for $Y_F(550)$ is (336, 790) .

3. A research team wishes to test the null hypothesis $H_0 : \rho = 0$ at $\alpha = 0.025$ against the alternative $H_1 : \rho > 0$ using Fisher's transformation of the Pearson product moment correlation coefficient as the test statistic. They have asked their consulting statistician for a sample size *n* such that $\beta = 0.05$ when $\rho = 0.10$ (that is, $\rho^2 = 0.01$). What is this value?

   Answer: $n \geq 1294$.

4. 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 as the test statistic. They have asked their consulting statistician for a sample size *n* such that $\beta = 0.01$ when $\rho = 0.25$ (that is, $\rho^2 = 0.0625$). What is this value?

   Answer: $n \geq 372$.

5. A research team exposed sixty four animals to a range of dosages of a supplemental diet and observed the response *Y*. Sixteen animals were exposed to one unit of dosage with observed average and sample variance (unbiased estimate) $y_{1\bullet} = 53.4$ and $s_1^2 = 18.1$; sixteen were exposed to two units of dosage with $y_{2\bullet} = 68.9$ and $s_2^2 = 23.6$; sixteen were exposed to three units of dosage with

$y_{3\bullet} = 86.3$ and $s_3^2 = 16.4$ and sixteen were exposed to four units of dosage with $y_{4\bullet} = 98.3$ and $s_4^2 = 19.7$.

a. Complete the analysis of variance table for the linear regression of the dependent variable on the dosage. Use the sum of squares for the linear contrast as the regression sum of squares. The coefficients of the linear contrast are $-3,-1,1,3$. Test the null hypothesis that the average response is not associated with the dosage given. Use the 0.10, 0.05, and 0.01 levels of significance.

b. Complete the analysis of variance table including the sum of squares for lack of fit. What is your conclusion? Use the 0.10, 0.05, and 0.01 levels of significance. With regard to the coefficients of the orthogonal polynomials, the coefficients of the quadratic contrast are $1,-1,-1,1$; and the coefficients of the cubic contrast are $-1,3,-3,1$. The total value of this problem is 80 points.

Answers:

a. Analysis of Variance Table Part a.

| | DF | SS | MS | F |
|---|---|---|---|---|
| Linear dosage | 1 | 18507.5 | 18507.5 | 911.7 |
| Error. | 62 | 1258.7 | 20.3 | |
| Total | 63 | 19766.2 | | |

Reject the null hypothesis that the slope of the regression line is zero at the 0.01 level of significance.

b. Analysis of Variance Table for Lack of Fit Test

| | DF | SS | MS | F |
|---|---|---|---|---|
| Linear dosage | 1 | 18507.5 | 18507.5 | |
| Lack of fit | 2 | 91.6 | 45.8 | 2.36 |
| (Pure) Error. | 60 | 1167.0 | 19.45 | |
| Total | 63 | 19766.2 | | |

Accept the null hypothesis that the regression model that is linear in dose appears to fit, as the lack of fit test was 2.36 with 2 numerator and 60 denominator degrees of freedom at the 0.10 level of significance (which has critical value 2.39).

6. The correlation matrix of the random variables $Y_1, Y_2, Y_3, Y_4$ is

$$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}, \ 0 < \rho < 1, \text{ and each random variable has variance } \sigma^2.$$

Find the variance of $\overline{Y}_4$.

Answer: $\operatorname{var}(\overline{Y}_4) = (\frac{1}{4} + \frac{3\rho}{8} + \frac{\rho^2}{4} + \frac{\rho^3}{8})\sigma^2$.

7. 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 $\sigma^2$. 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 $(W_1, W_2)$.

Answer: $vcv\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \begin{bmatrix} 3+6\rho & 2+7\rho \\ 2+7\rho & 3+6\rho \end{bmatrix}\sigma^2$.

8. The correlation matrix of the random variables $Y_1, Y_2, Y_3, Y_4$ is $\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$, $0 < \rho < 1$, and each random variable has variance $\sigma^2$. Let $W_1 = Y_1 - Y_2$, $W_2 = Y_2 - Y_3$, and $W_3 = Y_3 - Y_4$. Find the variance covariance matrix of $(W_1, W_2, W_3)$.

Answer:

$$vcv\begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix} = \begin{bmatrix} 2-2\rho & -1+2\rho-\rho^2 & -\rho+2\rho^2-\rho^3 \\ -1+2\rho-\rho^2 & 2-2\rho & -1+2\rho-\rho^2 \\ -\rho+2\rho^2-\rho^3 & -1+2\rho-\rho^2 & 2-2\rho \end{bmatrix}\sigma^2.$$