Statistical Inference Theory
Lesson 45
Correlation and Regression Analysis II

45.1- Estimating y_p

45.1 - Problem 1:

Step 1:The number of degrees of freedom is d = 7 - 2 = 5.

Step 2: From the Student distribution table, $t_{0.45} = 2.015$.

Step 3: To compute $S_{x,y}$ complete the following table:

x	y	$y_s = 0.63x + 31.46$	$(y - y_s)^2$
36.0	42.8	54.14	128.6
43.7	69.1	58.99	102.19
55.9	73.0	66.68	39.98
78.4	65.7	80.85	229.58
81.0	91.2	82.49	75.86
85.8	86.1	85.51	0.34
92.0	88.8	89.42	0.38
Total = 472.7			$S^2_{x,y}$ = **Total/7** = 82.41

Step 4: $S_{x,y} = \sqrt{82.41} \approx 9.08$.

Step 5: For x = 75, $y_s = 0.63(75) + 31.46 = 78.71$

Step 6: $\bar{x} = \dfrac{472.7}{7} \approx 67.53$

Step 7: $S_x^2 = \dfrac{(36 - 67.53)^2 + (43.7 - 67.53)^2 + \ldots + (92 - 67.53)^2}{7} \approx 418.49$

$S_x = \; = \sqrt{418.49} \approx 20.46$

Step 8: Given the above confidence interval,

$$78.71 - \frac{2.015}{\sqrt{7-2}}(9.08)\sqrt{8 + \frac{7(75 - 67.53)^2}{418.49}} \leq y_p \leq 78.71 + \frac{2.015}{\sqrt{7-2}}(9.08)\sqrt{8 + \frac{7(75 - 67.53)^2}{418.49}} \; .$$

Step 9:The above simplifies to $78.71 - 24.46 \leq y_p \leq 78.1.71 + 24.46$.

$54.25 \leq y_p \leq 103.17$

45.2 - Hypotheses Testing for ρ.

45.2 - Problem 1:

➤(a).

There is either no climate correlation or a significant climate correlation . Therefore,

H_o: $\rho = 0$

H_a: $\rho \neq 0$

➤(b).

Step 1: The degree of freedom is 20 - 2 = 18.

Step 2: Since we have a two-tail test, we have $t_{.475} = 2.101$.

Step 3:$t = \dfrac{r\sqrt{N-2}}{\sqrt{1 - r^2}} = \dfrac{(0.2)\sqrt{20-2}}{\sqrt{1 - 0.2^2}} \approx 0.87.$

Step 4: Since 0.87 < 2.101, we cannot reject $H_{o.}$ Therefore, Mr. Smith could conclude that the two climates do not have a correlation significantly different than zero.

45.2 - Problem 2:

➤(a).

Here we wish to test if the new brand of feed does not perform as well.

H_o: $\rho = 0.80$

H_a: $\rho < 0.80$

➤(b).

Since H_o: $\rho = 0.80$, we follow case 2.

Step 1: $R = \frac{1}{2}\ln(\frac{1+r}{1-r}) = \frac{1}{2}\ln(\frac{1+0.77}{1-0.77}) \approx 1.02$

Step 2: $\mu = \frac{1}{2}\ln(\frac{1+\rho)}{1-\rho)} = \frac{1}{2}\ln(\frac{1+0.80)}{1-0.80)} \approx 1.10$

Step 3: $\sigma_x = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{100-3}} \approx 0.10$

Step 4: $z = \frac{x-\mu}{\sigma_x} = \frac{1.02-1.1}{0.10} = -0.80$

Step 5: Since x and z are normally distributed, and we are using a 5% significance level, the normal distribution table gives z = -1.64.

Step 6: Since z = -0.8 > -1.64 H_o would not be rejected. We could conclude that there is no significant statistical evidence that there has been a decrease in the production of eggs.

Supplementary Problems

1.
If X is normally distributed, the formula for the confidence interval is

$$X - z\sigma_{\overline{X}} \leq \mu \leq X + z\sigma_{\overline{X}} .$$

Using this formula for the distribution of $x = \frac{1}{2}\ln(\frac{1+r}{1-r})$ which is normally distributed with mean

$$\mu = \frac{1}{2}\ln(\frac{1+\rho)}{1-\rho)} \quad \text{and}$$

$$\sigma_x = \frac{1}{\sqrt{N-3}}$$

we have the following formula for a confidence interval for ρ:

$$\frac{1}{2}\ln(\frac{1 + r}{1 - r}) - \frac{z}{\sqrt{N - 3}} \le \rho \le \frac{1}{2}\ln(\frac{1 + r}{1 - r}) + \frac{z}{\sqrt{N - 3}}.$$

2.
The following is 45.2 - Example 2**:**

In studying the price movements of corn and cattle, Mrs. Jones concludes that the correlation between these two commodity prices is at least 0.50. To test this hypotheses a study of these monthly prices over a 48 month period resulted in a correlation coefficient r = 0.41.

We use the formula derived in problem 1:

$$\frac{1}{2}\ln(\frac{1 + r}{1 - r}) - \frac{z}{\sqrt{N - 3}} \le \rho \le \frac{1}{2}\ln(\frac{1 + r}{1 - r}) + \frac{z}{\sqrt{N - 3}}.$$

Since r = 0.41, N = 48, z = 1.96 (for a area of 0.475 in the normal distribution table):

$$\frac{1}{2}\ln(\frac{1 + 0.41}{1 - 0.41}) - \frac{1.96}{\sqrt{48 - 3}} \le \rho \le \frac{1}{2}\ln(\frac{1 + 0.41}{1 - 0.41}) + \frac{1.96}{\sqrt{48 - 3}}.$$

$0.18 \le \rho \le 0.73$

Assume we have two populations. From each population we take a sample and compute for each sample correlation coefficients r_1 and r_2 respectively. The distribution

$$R = \frac{1}{2}\ln(\frac{1 + r}{1 - r}),$$

$$z = \frac{R_1 - R_2 - \mu_{R_1 - R_2}}{\sigma_{R_1 - R_2}}$$ are normally distributed where

$$\mu_{R_1 - R_2} = \mu_{R_1} - \mu_{R_2},$$

$$\sigma_{R_1 - R_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}},$$

N_1 and N_2 are the sample sizes respectively.

A research institute recently did a study to see if there is a significant difference between men and women according to their respective correlations of weight and cholesterol. They sampled $N_1 = 200$ women and $N_2 = 100$ and found $r_1 = 0.45$ and $r_2 = 0.58$.

3.

$H_o: \mu_{R_2 - R_1} = \mu_{R_2} - \mu_{R_1} = 0$

$H_a: \mu_{R_2 - R_1} = \mu_{R_2} - \mu_{R_1} \neq 0$

4.

We need to use

$$z = \frac{R_2 - R_1 - \mu_{R_2 - R_1}}{\sigma_{R_2 - R_1}} \text{ where}$$

$$R_1 = \frac{1}{2} \ln\left(\frac{1 + 0.45}{1 - 0.45}\right) \approx 0.48$$

$$R_2 = \frac{1}{2} \ln\left(\frac{1 + 0.58}{1 - 0.58}\right) \approx 0.66$$

$$\sigma_{R_2 - R_1} = \sqrt{\frac{1}{100 - 3} + \frac{1}{200 - 3}} \approx 0.12.$$

We assume $\mu_{R_2 - R_1} = 0.$

Since $\alpha = 0.05$ and we have a 2 sided test, we look-up the z value for the area 0.95/2 = 0.475:

$z = 1.96.$

$$z = \frac{0.66 - 0.48 - 0}{0.12} = 1.5$$

Since $1.5 < 1.96$, there is no significant difference between the two correlations. Reject H_a..

5.

From 45.1 - example 1, we have

x	y	$y_s = 0.40x + 144.15$	$(y - y_s)^2$
175	225	214.15	117.72
197	205	222.95	322.20
210	220	228.15	66.42
210	320	228.15	8436.42
225	190	234.15	1949.22
235	172	238.15	4375.82
237	250	238.95	122.10
255	210	246.15	1306.82
278	320	255.35	4179.62
298	256	263.35	54.02
Total = 2320			$S^2_{x,y}$ = **Total/10** = 2093.04

For x = 250, $y_s = 0.40(250) + 144.15 = 244.15$.

For a 99% (0.99/2 = 0.495) confidence interval and 10 - 2 = 8 degrees of freedom, t = 3.355.

$$S_{x,y} = \sqrt{2093.04} \approx 45.75$$

$$\bar{x} = \frac{2320}{10} = 232$$

$$S_x^2 = \frac{(175 - 232)^2 + (197 - 232)^2 + \ldots + (298 - 232)^2}{10} \approx 1392.04$$

$$244.15 - \frac{3.355}{\sqrt{10 - 2}}(45.75)\sqrt{1 + \frac{(250 - 232)^2}{1392.04}} \le \mu_{xy} \le 244.15 + \frac{3.355}{\sqrt{10 - 2}}(45.75)\sqrt{1 + \frac{(250 - 232)^2}{1392.04}}$$

$$244.14 - 60.25 \le \mu_{xy} \le 244.14 + 60.25$$

$$183.77 \le \mu_{xy} \le 304.39$$