

Ma 322: Biostatistics

Solutions to Homework Assignment 10

Prof. Wickerhauser

Due Friday, April 13th, 2012

Read Chapter 15, “ANOVA and Regression,” pages 263–287 of our text.

NOTE: It should be possible to cut and paste the data from this document into a text file, or into an R variable by use of the `scan()` function.

1. The number of chirps per minute emitted by three crickets is tabulated at several temperatures:

Temperature (°C)	Cricket 1 (chirps/min)	Cricket 2 (chirps/min)	Cricket 3 (chirps/min)
20	110	115	112
21	118	121	118
22	128	127	123
23	133	135	131
24	140	140	140
25	144	147	150
27	165	161	160
28	169	166	171
29	176	179	173
30	182	187	182
31	191	196	190
33	201	206	207
35	216	222	221

- (a) Compute the linear regression equation, namely find a, b , estimating chirp frequency f as a function of temperature T by $f = a + bT$.
- (a') Compute the linear regression equation estimating temperature as a function of chirp frequency: $T = a' + b'f$.
- (b) Calculate the standard error of estimate $S_{f.T}$ of the regression in part (a).
- (b') Calculate the standard error of estimate $S_{T.f}$ of the regression in part (a').
- (c) Test, by t test at the $\alpha = 0.05$ level, the hypothesis $H_0 : \beta = 0$ for the regression in part a.
- (c') Test, by analysis of variance at the $\alpha = 0.05$ level, the hypothesis $H_0 : \beta' = 0$ for the regression in part a'.
- (d) Calculate the coefficient of determination of the regression $f = a + bT$.
- (d') Calculate the coefficient of determination of the regression $T = a' + b'f$.
- (e) Test H_0 : the population regression of f as a function of T is linear.
- (f) Suppose one of the crickets is heard chirping 180 times in one minute. Estimate the temperature find its 50% confidence interval.

Solution: See the file `hw10R.txt` for the R commands that produced these results.

(a) $b = 7.197$, $a = -32.324$, in $f = a + bT$.

(a') $b' = 0.1383$, $a' = 4.5948$, in $T = a' + b'f$.

(b) Read the output of `summary(lm(f~T))`: $S_{f.T} = 2.294$, noting that this is the “Residual standard error” in R’s terminology.

(b') Do the same as in part (b), but for `summary(lm(T~f))`. This gives $S_{T.f} = 0.318$, the “Residual standard error” in R’s terminology.

(c) The t -statistic for b is 89.02 and has a two-tailed p -value less than 10^{-15} . Hence we **reject** the hypothesis $H_0 : \beta = 0$ at the 0.05 significance level.

(c') The F -statistic for b' is 7924 and has a p -value less than 10^{-15} . Hence we **reject** the hypothesis $H_0 : \beta' = 0$ at the 0.05 level.

(d) $r^2 = 0.9954$

(d') $r^2 = 0.9954$, same as (d) of course.

(e) Test for linearity as in the example on the class website. This gives a deviation from linear SS of 12.67, with 12-1=11 degrees of freedom, and a residual SS of 182.00 with 26 degrees of freedom. The ratio of mean squares gives $F = 0.1645 < 1$, so **do not reject** H_0 : the relationship is linear.

(f) From the computations in part (a'), get the parameters $b' = 0.1383$ and $a = 4.5948$ in $T = a' + b'f$. Evaluate this at $f = 180$ with `predict()` to get the estimate $\hat{T} = 29.49$, with a 50% confidence interval $\hat{T} \pm \delta = [29.45, 29.53]$. \square

2. Given the following data:

Y	X_1	X_2	X_3	X_4
51.4	0.2	17.8	24.6	18.9
72.0	1.9	29.4	20.7	8.0
53.2	0.2	17.0	18.5	22.6
83.2	10.7	30.2	10.6	7.1
57.4	6.8	15.3	8.9	27.3
66.5	10.6	17.6	11.1	20.8
98.3	9.6	35.6	10.6	5.6
74.8	6.3	28.2	8.8	13.1
92.2	10.8	34.7	11.9	5.9
97.9	9.6	35.8	10.8	5.5
88.1	10.5	29.6	11.7	7.8
94.8	20.5	26.3	6.7	10.0
62.8	0.4	22.3	26.5	14.3
58.4	6.6	15.7	8.7	26.3
81.6	2.3	37.9	20.0	0.5

(a) Fit the multiple regression $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ to the data, computing the sample partial regression coefficients and Y intercept.

(b) Test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ by ANOVA at the 0.05 level.

(c) Compute the standard error of each partial regression coefficient and test $H_0 : \beta_i = 0$, at the $\alpha = 0.05$ level, individually for each $i = 1, 2, 3, 4$.

(d) Calculate the standard error of estimate and the coefficient of determination.

(e) What is the predicted mean population value \hat{Y} at $X_1 = 5.4$, $X_2 = 20.3$, $X_3 = 18.7$, $X_4 = 11.2$?

(f) What is the 95% confidence interval for \hat{Y} in part (e)?

Solution: See `hw10R.txt` for the R commands used to compute the results.

(a) $(b_1, b_2, b_3, b_4) = (2.0735, 2.5798, 0.6407, 1.1014)$; $a = -30.1423$.

(b) $F = 109.1550$ with 4 numerator and 10 denominator degrees of freedom. This has a one-tailed p -value of 3.313456×10^{-8} , so we **reject** H_0 at the 0.05 level.

(c) H_0 is rejected at the 0.05 significance level in part b. For all parameters, $\nu = 10 = n - m - 1$, giving the following t statistics and their p -values:

$$\begin{pmatrix} S_{b_1} \\ S_{b_2} \\ S_{b_3} \\ S_{b_4} \end{pmatrix} = \begin{pmatrix} 0.609 \\ 0.2904 \\ 0.6525 \\ 0.3114 \end{pmatrix} \Rightarrow \begin{pmatrix} t = b_1/S_{b_1} \\ t = b_2/S_{b_2} \\ t = b_3/S_{b_3} \\ t = b_4/S_{b_4} \end{pmatrix} = \begin{pmatrix} 3.336 \\ 6.154 \\ -1.845 \\ -5.26 \end{pmatrix} \Rightarrow P \approx \begin{pmatrix} 0.00536 \\ 0.00003 \\ 0.08790 \\ 0.00015 \end{pmatrix}$$

Hence, at the 0.05 level we **reject** the null hypotheses $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_4 = 0$, but we **do not reject** the null hypotheses $\beta_3 = 0$.

(d) Coefficient of multiple determination: $R^2 = 0.9776$. Adjusted coefficient of multiple determination: $R_a^2 = 0.9687$. Standard error of estimate: $S_{Y \cdot 1, \dots, M} = \sqrt{\text{Residual MS}} = 2.948$.

(e) Substitute the values into the multiple regression equation $\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$ to get $\hat{Y} = 57.74$.

(f) Start with the X values used in part (e) to compute \hat{Y} with 95% confidence interval endpoints: $[47.99, 67.49]$. \square

3. Perform a stepwise regression analysis of the data in Problem 2.

Solution: See `hw10out.txt` for the R commands used to compute the results. The first step is already done in Problem 1, part (c); variable X_3 has the greatest P value for $H_0 : \beta = 0$, so remove it from the regression.

In Step 2, repeat parts (a,b,d,c) of Problem 1 on the remaining data set (Y, X_1, X_2, X_4) . This gives $(b_1, b_2, b_4) = (1.4740, 1.7031, 0.1720)$, $a = 18.1039$. For all parameters, $\nu = 11 = n - m - 1$, giving the following t values and their likelihoods:

$$\begin{pmatrix} S_{b_1} \\ S_{b_2} \\ S_{b_4} \end{pmatrix} = \begin{pmatrix} 0.1540 \\ 0.3968 \\ 0.3792 \end{pmatrix} \Rightarrow \begin{pmatrix} t = b_1/S_{b_1} \\ t = b_2/S_{b_2} \\ t = b_4/S_{b_4} \end{pmatrix} = \begin{pmatrix} 9.569 \\ 4.292 \\ 0.454 \end{pmatrix} \Rightarrow P \approx \begin{pmatrix} 1 \times 10^{-6} \\ 0.00127 \\ 0.65884 \end{pmatrix}$$

Hence, at the 0.05 level we **reject** the null hypotheses $\beta_1 = 0$ and $\beta_2 = 0$, but we **do not reject** the null hypotheses $\beta_4 = 0$.

In Step 3, repeat parts (a,b,d,c) of Problem 1 on the remaining data set (Y, X_1, X_2) . This gives $(b_1, b_2) = (1.4752, 1.5297)$, and $a = 24.8652$. For all parameters, $\nu = 12 = n - m - 1$, giving the following t values and their likelihoods:

$$\begin{pmatrix} S_{b_1} \\ S_{b_2} \end{pmatrix} = \begin{pmatrix} 0.1488 \\ 0.1030 \end{pmatrix} \Rightarrow \begin{pmatrix} t = b_1/S_{b_1} \\ t = b_2/S_{b_2} \end{pmatrix} = \begin{pmatrix} 9.912 \\ 14.846 \end{pmatrix} \Rightarrow P \approx \begin{pmatrix} 4 \times 10^{-7} \\ 4 \times 10^{-9} \end{pmatrix}$$

Hence, at the 0.05 level we **reject** the null hypotheses $\beta_1 = 0$ and $\beta_2 = 0$, concluding that $Y = \alpha + \beta_1X_1 + \beta_2X_2$ with $\alpha \approx 24.8652$, $\beta_1 \approx 1.4752$, and $\beta_2 \approx 1.5297$. \square

4. Analyze the five variables in Problem 2 as a multiple correlation.

(a) Compute the simple correlation coefficient for each pair of variables.

(b) Compute the multiple correlation coefficient R for each variable in terms of the other 4, and test $H_0 : R = 0$ at the 0.05 level in each case.

(c) Compute the partial correlation coefficients for the five variables.

Solution: See `hw10R.txt` for the R commands used to compute the results.

(a) With the Y variable in the first column (as X_0 , so to speak), the simple correlation matrix is:

$$r = \frac{1}{\sqrt{\text{diag } SSQP}} SSQP \frac{1}{\sqrt{\text{diag } SSQP}}$$

$$= \begin{pmatrix} 1.0000000 & 0.6791849 & 0.86282623 & -0.45558037 & -0.82482122 \\ 0.6791849 & 1.0000000 & 0.25194242 & -0.80577226 & -0.23873447 \\ 0.8628262 & 0.2519424 & 1.00000000 & -0.09089558 & -0.96534792 \\ 0.4555804 & -0.8057723 & -0.09089558 & 1.00000000 & -0.04742297 \\ 0.8248212 & -0.2387345 & -0.96534792 & -0.04742297 & 1.00000000 \end{pmatrix}$$

where $SSQP$ is the “sum of squares and cross products” matrix.

(b) Read the F statistic for these tests from the R output:

$$F = \left(\frac{R^2}{1 - R^2} \right) \left(\frac{\text{Residual DF}}{\text{Regression DF}} \right)$$

For the 5 linear models, we have:

$$\begin{pmatrix} R^2(Y \sim X_1 + X_2 + X_3 + X_4) \\ R^2(X_1 \sim Y + X_2 + X_3 + X_4) \\ R^2(X_2 \sim X_1 + Y + X_3 + X_4) \\ R^2(X_3 \sim X_1 + X_2 + Y + X_4) \\ R^2(X_4 \sim X_1 + X_2 + X_3 + Y) \end{pmatrix} = \begin{pmatrix} 0.9776 \\ 0.9674 \\ 0.9917 \\ 0.9314 \\ 0.9863 \end{pmatrix} \Rightarrow F = \begin{pmatrix} 109.2 \\ 74.25 \\ 298.9 \\ 33.97 \\ 180 \end{pmatrix} \Rightarrow P \approx \begin{pmatrix} 3 \times 10^{-8} \\ 2 \times 10^{-7} \\ 2 \times 10^{-10} \\ 9 \times 10^{-6} \\ 3 \times 10^{-9} \end{pmatrix}.$$

Each F statistic has 4 and 10 degrees of freedom in the numerator and denominator, respectively. Thus in all cases, **reject** the null hypothesis.

(c) The pairwise partial correlation coefficient matrix is given by the off-diagonal terms (the diagonals are all -1 with the simplified formula that we use):

$$p \stackrel{\text{def}}{=} -\frac{1}{\sqrt{\text{diag } r^{-1}}} r^{-1} \frac{1}{\sqrt{\text{diag } r^{-1}}}$$

$$= \begin{pmatrix} -1.0000000 & 0.8360126 & 0.7581710 & 0.4242227 & 0.4338247 \\ 0.8360126 & -1.0000000 & -0.9121198 & -0.8219524 & -0.7644270 \\ 0.7581710 & -0.9121198 & -1.0000000 & -0.8196893 & -0.9100163 \\ 0.4242227 & -0.8219524 & -0.8196893 & -1.0000000 & -0.8916415 \\ 0.4338247 & -0.7644270 & -0.9100163 & -0.8916415 & -1.0000000 \end{pmatrix}$$

Notice how this uncovers the weak dependence of Y , the first column variable, on X_3 and X_4 in rows 4 and 5. The simple correlation matrix R shows a too-strong correlation, in positions (1,4) and (1,5), between Y and X_3 and between Y and X_4 , respectively. \square

5. Each of five research papers was read by each of six reviewers. Each reviewer then marked the quality of the five papers as follows:

Reviewer	Paper				
	1	2	3	4	5
A	5	4	3	1	2
B	4	5	3	2	1
C	5	4	1	2	3
D	5	3	2	4	1
E	4	5	2	3	1
F	5	4	1	3	2

- (a) Calculate the Kendall coefficient of concordance.
 (b) Test, at the $\alpha = 0.01$ significance level, whether the rankings by the six reviewers are in agreement.

Solution: See `hw10R.txt` for the R commands used to compute the results.

- (a) Compute the rank sums and the Kendall concordance coefficient for $m = 6$ judges and $n = 5$ ranked items:

$$\begin{pmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{pmatrix} = \begin{pmatrix} 28 \\ 25 \\ 12 \\ 15 \\ 10 \end{pmatrix}; \quad W = \frac{\sum_{i=1}^n R_i^2 - \frac{1}{n} [\sum_{i=1}^n R_i]^2}{m^2(n^3 - n)/12} = 0.7167$$

- (b) For $W = 0.7167$, the Friedman chi-squared value is $\chi_r^2 = m(n-1)W = 17.2$. This has a significant $p < 0.01$, so we **reject** the null hypothesis H_0 : the six reviewers agree, in favor of H_A : the six reviewers are not in agreement.

□