Unit 5 – Regression and Correlation
Practice Problems (1 of 3)

Due: Wednesday October 18, 2023

Last date to submit late for credit (-20 points): Wednesday October 18, 2023

Before you begin. Download from the course website
simplelinear.xlsx

1.

This exercise gives you practice doing a simple linear regression using **simplelinear.xlsx.** This data set has n=31 observations of boiling points (Y=boiling) and temperature (X=temp). You will be exploring the following two simple linear models:

(i) $Y = b_0 + b_1 X$; where Y=boiling and X=temp

(ii) newy = $b_0 + b_1 X$; where newy = $100*\log_{10}(y)$ and where y=boiling and X=temp

1a. Create a new variable newy = $100*\log_{10}$(boiling)

1b. For each model, obtain:

a. The fitted line estimates of $\hat{b}_0$ and $\hat{b}_1$
b. Analysis of variance table
c. R^2 = % of the variability in the outcome explained by the fitted line
d. Scatter plot with overlay of fitted line

1c. In 3-5 sentences, write a one-paragraph interpretation of your two model fits.

#2.

Note – This question does NOT require use of software (R or otherwise!)

This exercise gives you practice working with a fitted model that is provided to you. A psychiatrist wants to know whether the level of pathology (Y) in psychotic patients 6 months after treatment could be predicted with reasonable accuracy from knowledge of pretreatment symptom ratings of thinking disturbance (X_1) and hostile suspiciousness (X_2).

2a. The least squares estimation equation involving both independent variables is given by

$$Y = -0.628 + 23.639(X_1) - 7.147(X_2)$$

Using this equation, determine the predicted level of pathology (Y) for a patient with pretreatment scores of 2.80 on thinking disturbance and 7.0 on hostile suspiciousness. How does the predicted value obtained compare with the actual value of 25 observed for this patient?

2b. Using the analysis of variance tables below, carry out the *overall* F test for each of three models: i) model with X_1 alone; ii) model with X_2 alone; and iii) model with both X_1 and X_2.

Source	DF	Sum of Squares
Regression on X_1	1	1546
Residual	51	12246

Source	DF	Sum of Squares
Regression on X_2	1	160
Residual	51	13632

Source	DF	Sum of Squares
Regression on X_1 , X_2	2	2784
Residual	50	11008

2c. Based on your results in part (b), how would you rate the importance of the two variables in predicting Y?

2d. What are the R^2 values for the three regressions referred to in part (b)?

2e. Based on the above, in your opinion, which is the best model involving either one or both of the two independent variables?

#3.

Note – This question does NOT require use of software (R or otherwise!) with one exception: to obtain p-values for parts a-c. Tip – Use Art of Stat if you like!

This exercise gives you practice working with analysis of variance tables. In an an experiment to describe the toxic action of a certain chemical on silkworm larvae, the relationship of $\log_{10}$(dose) and $\log_{10}$(larva weight) to $\log_{10}$(survival) was sought. The data, obtained by feeding each larva a precisely measured dose of the chemical in an aqueous solution and then recording the survival time (ie time until death) are given in the table. Also given are relevant computer results and the analysis of variance table.

Larva	1	2	3	4	5	6	7	8
Y = $\log_{10}$(survival time)	2.836	2.966	2.687	2.679	2.827	2.442	2.421	2.602
X_1=$\log_{10}$(dose)	0.150	0.214	0.487	0.509	0.570	0.593	0.640	0.781
X_2=$\log_{10}$(weight)	0.425	0.439	0.301	0.325	0.371	0.093	0.140	0.406

Larva	9	10	11	12	13	14	15
Y = $\log_{10}$(survival time)	2.556	2.441	2.420	2.439	2.385	2.452	2.351
X_1=$\log_{10}$(dose)	0.739	0.832	0.865	0.904	0.942	1.090	1.194
X_2=$\log_{10}$(weight)	0.364	0.156	0.247	0.278	0.141	0.289	0.193

#3 - continued.

$$Y = 2.952 - 0.550\ (X_1)$$
$$Y = 2.187 + 1.370\ (X_2)$$
$$Y = 2.593 - 0.381\ (X_1) + 0.871\ (X_2)$$

Source	DF	Sum of Squares
Regression on X_1	1	0.3633
Residual	13	0.1480

Source	DF	Sum of Squares
Regression on X_2	1	0.3367
Residual	13	0.1746

Source	DF	Sum of Squares
Regression on X_1 , X_2	2	0.4642
Residual	12	0.0471

3a. Test for the significance of the overall regression involving both independent variables X_1 and X_2.

3b. Test to see whether using X_1 alone significantly helps in predicting survival time.

3c. Test to see whether using X_2 alone significantly helps in predicting survival time.

3d. Compute R^2 for each of the three models.

3e. Which independent predictor do you consider to be the best single predictor of survival time?

3f. Which model involving one or both of the independent predictors do you prefer and why?