

Instructions: Show all work. Answers without necessary formulas and steps will not receive any credit.

1. (75 points) The Summer Olympic Games have been held from 1896 until 1996. Over this 100 year period, there have been a total of 40 times that the 100 meter dash has been run: 16 times for women and 24 times for men. For each event, the winning (gold medal) times have been recorded. For example, the winning men's time in 1996 was 9.84 seconds, while the winning women's time was 10.94 seconds.

An example of the data is given below:

Time (seconds)	Gender	Year
12.20	Women	1928
11.90	Women	1932
.....		
11.00	Men	1904
11.20	Men	1906

A simple linear regression model with the winning times (y) as the response and the year (x) as the predictor was fit to the 40 observations:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, 40$$

The ANOVA results are shown below:

Analysis of Variance

Source	DF	Sum of Squares	Mean Squares	F Value	Prob>F
Model	1	3.01308	3.01308	9.210	0.0043
Error	38	12.43200	0.32716		
Total	39	15.44508			

The least squares estimates for the intercept and slope are $\hat{\beta}_0 = 29.119$ and $\hat{\beta}_1 = -0.009$, respectively.

(a) (15 points) What is the value of r^2 , the coefficient of determination, for this simple linear regression model? How is this measure interpreted here?

(b) (10 points) What is the interpretation of the estimated slope coefficient, $\hat{\beta}_1$, for this model?

Next a linear regression model with separate slopes for men and women was fit to the data. A dummy variable was defined as follows:

$$G_i = \begin{cases} 1 & \text{if Gender is men;} \\ 0 & \text{if Gender is women,} \end{cases} \quad i = 1, \dots, 40$$

The new model is $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i \times G_i + \varepsilon_i$, $i = 1, \dots, 40$. Note that this model imposes a common intercept β_0 and separate slope parameters for men and women.

(c) (15 points) What are the slopes for men and women? Write down your answers in symbols.

The ANOVA table for the multiple regression model is shown below:

Analysis of Variance

Source	DF	Sum of Squares	Mean Squares	F Value	Prob>F
Model	2	13.40617	6.70309	121.641	0.0001
Error	37	2.03890	0.05511		
Total	39	15.44508			

(d) (20 points) Use the ANOVA results from the simple and multiple linear regressions to test whether the slopes for men and women are different. Include all steps of a statistical test and interpret your results in the context of the setting. Use $\alpha = 0.05$.

(e) (15 points) The following diagnostic values (Cook's distance, studentized deleted residuals, and leverages) were computed for the earliest men's 100 meter events (from 1896 to 1906), based on the multiple regression model:

Year	Cook's Dist.	Stud. Del. Residual	Leverage
1896	0.579	4.2807	0.1222
1900	0.022	-0.7273	0.1097
1904	0.008	-0.4494	0.0984
1906	0.011	0.5725	0.0931

Based on these three diagnostic measures, comment on whether any of the above 4 observations are influential.

2. (70 points) An experiment was set up to compare the effect of different soil pH and calcium additives on the increase in trunk diameters for orange trees. Annual applications of elemental sulfur, gypsum, soda ash, and other ingredients were applied to provide pH value levels of 4, 5, 6 and 7. Three levels of a calcium supplement (100, 200 and 300 pounds per acre) were also applied. All factor-level combinations of these two variables were used in the experiment. At the end of a 2-year period, three diameters were examined at each factor-level combination. The data are given below.

(a) (15 points) Construct an interaction plot. What do you learn from the interaction plot?

(b) (15 points) Identify the experimental design. Write an appropriate statistical model and define every term (on both sides of the "=" sign) in the model.

pH	Calcium		
	100	200	300
4.0	5.2	7.4	6.3
	5.9	7.0	6.7
	6.3	7.6	6.1
5.0	7.1	7.4	7.3
	7.4	7.3	7.5
	7.5	7.1	7.2
6.0	7.6	7.6	7.2
	7.2	7.5	7.3
	7.4	7.8	7.0
7.0	7.2	7.4	6.8
	7.5	7.0	6.6
	7.2	6.9	6.4

(c) (20 points) Use the computer output given below to perform an overall F test. Include all steps of a statistical test and provide conclusions in the context of the setting. Use $\alpha = 0.05$.

Source	DF	Sum of Squares	Mean Squares
PH	3	4.4608	1.4869
CA	2	1.4672	0.7336
PH*CA	6	3.2550	0.5425
Error	24	1.6267	0.0678
Total	35	10.8097	

(d) (20 points) Test for the interaction effects. Include all steps of a statistical test and provide conclusions in the context of the setting. Use $\alpha = 0.05$.

3. (55 points) Consider a completely randomized design for 5 treatments, with a single covariate x_1 and six observations per treatment. Assume that the response y is linearly related to the covariate x_1 for each treatment.

(a) (15 points) Write an appropriate statistical model allowing different linear relationships (lines) between y and x_1 for each treatment. Define the terms in your model.

(b) (15 points) Write a reduced model of the complete model in (a) assuming that the lines are parallel but do not coincide. Show a graph with the intercept and slope of each line marked by symbols in your model.

(c) (10 points) Write a reduced model of the complete model in (a) assuming that the lines are parallel and coincident.

(d) (15 points) Give the complete model and reduced model for testing the parallelism among the lines for the five treatment groups. Provide a test statistic and specify which model the sum of squares and degrees of freedom are from.