

STAT230A

Due at 3:00 p.m. on Feb 24, 2017 (@ Yuting Ye's mailbox at 367 Evans Hall)

Problem Set 3

Homework Policy:

You're encouraged to discuss with your fellows but need to finish your own write-up. Delay is not allowed. The write-up is encouraged to be a pdf version, either by Latex, Markdown or other formats. A tidy and clean write-up deserves a higher grade when there are mistakes. When there are real-data problems, R is required and only allowed. R Markdown and R Sweave (used in Rstudio) are recommended to report the results of the real-data problems.

Problem 1 (10 pts, ★☆☆☆☆)

Consider a set of seemingly unrelated regression equations

$$Y_i = X_i\beta_i + e_i, \quad e_i \sim N(0, \sigma^2 I), \quad i = 1, \dots, r$$

where $X_i \in \mathbb{R}^{n_i \times p}$ and the e_i 's are independent. Find the test for $H_0 : \beta_1 = \dots = \beta_r$.

Problem 2 (15 pts, ★☆☆☆☆)

Imagine we observe $(x_1, y_1), \dots, (x_n, y_n)$ where (x_i, y_i) are multivariate normal with mean μ_x, μ_y , $\text{Var}(x_i) = \text{Var}(y_i) = \sigma^2$ and correlation ρ . We are interested in testing the null hypothesis that $\mu_x = \mu_y$. Under the null hypothesis, we know

$$t = \frac{(\bar{x} - \bar{y})}{s_{pooled} \sqrt{2/n}}$$

is distributed as a Student's t with $2n - 2$ degrees of freedom, where s_{pooled} is the pooled sample standard deviation. See any undergraduate text (or Wikipedia page "Student's t-test") if you are unfamiliar with the t distribution.

1. Write s_{pooled} in terms of x_i , y_i , \bar{x} and \bar{y} (this is a standard definition).

2. What is the expectation of s_{pooled}^2 ?

3. The statement above (on the t-statistics) isn't quite right. Are any additional assumptions needed?

Consider doing a paired t-test with the same data. The test statistics here is

$$t_{paired} = \frac{\bar{x} - \bar{y}}{s_{diff} \sqrt{1/n}}$$

4. Write s_{diff} in terms of x_i , y_i , \bar{x} and \bar{y} . (another standard definition)



5. What distribution does t_{paired} have?



6. What is the expectation of s_{diff}^2 ?

7. Compare $\frac{s_{diff}^2}{n}$ to $\frac{s_{pooled}^2}{n}$. When is $\frac{s_{diff}^2}{n} < \frac{s_{pooled}^2}{n}$? When is $\mathbb{E}(\frac{s_{diff}^2}{n}) < \mathbb{E}(\frac{s_{pooled}^2}{n})$?

8. From these computations, what do you learn?

Problem 3 (20 pts, ★★☆☆☆)

After the final exam of spring quarter, 30 of the subjects of the previous experiment decided to test the sturdiness of 3 brands of sport coats and 2 brands of shirts. In this study, sturdiness was measured as the length of time before tearing when the instructor was hung by his collar out of his second-story office window. Each brand was randomly assigned to 6 students, but the instructor was occasionally dropped before his collar tore, resulting in some missing data. The data are listed below.

Coat 1 :	2.34	2.46	2.83	2.04	2.69	
Coat 2 :	2.64	3.00	3.19	3.83		
Coat 3 :	2.61	2.07	2.80	2.58	2.98	2.30
Shirt 1 :	1.32	1.62	1.92	0.88	1.50	1.30
Shirt 2 :	0.41	0.83	0.53	0.32	1.62	


- (a) Give an ANOVA table for these data, and perform and interpret the F test for the differences between brands
- (b) Test whether, on average, these brands of coats are sturdier than these brands of shirts.
- (c) Give three contrasts that are mutually orthogonal and orthogonal to the contrast used in (b). Compute the sums of squares for all four contrasts.
- (d) Give a 95% confidence interval for the difference in sturdiness between shirt Brands 1 and 2. Is one brand significantly sturdier than the other?

(Hint: Read Chapter 4 of Christensen's book.)

Problem 4 (20 pts, ★★☆☆☆)

An experiment was conducted to see which of four brands of blue jeans were most resistant to wearing out as a result of students kneeling before their linear models instructor begging for additional test points. In a class of 32 students, 8 students were randomly assigned to each brand of jeans. Before being informed of their test score, each student was required to fall to his/her kneew and crawl 3 meters to the instructor's desk. This was done after each of 5 mid-quarter and 3 final exam. (The jeans were distributed along with each of the 8 test forms and were collected again 36 hours after grades were posted.) A fabric wear score was determined for each pair of jeans. The scores are listed below.

Brand 1:	3.41	1.83	2.69	2.04	2.83	2.46	1.84	2.34
Brand 2:	3.58	3.83	2.64	3.00	3.19	3.57	3.04	3.09
Brand 3:	3.32	2.62	3.92	3.88	2.50	3.30	2.28	3.57
Brand 4:	3.22	2.61	2.07	2.58	2.80	2.98	2.30	1.66

- (a) Give an ANOVA table for these data, and perform and interpret the F test for the differences between brands.
- (b) Brands 2 and 3 were relatively inexpensive, while Brands 1 and 4 were very costly. Based on these facts, determine an appropriate set of orthogonal contrasts to consider in this problem.  Find the sums of squares for the contrasts.
- (c) Compare all pairs of means for the blue jeans by using the following methods:
 - (i) Scheffe's method, $\alpha = 0.01$,
 - (ii) LSD method, $\alpha = 0.01$,
 - (iii) Bonferroni method, $\alpha = 0.012$,
 - (iv) Tukey's HSD method, $\alpha = 0.01$,
 - (v) Newman-Keuls method, $\alpha = 0.01$.

(Hint: Read Chapter 5 of Christensen's book.)

Problem 5 (20 pts, ★☆☆☆☆)

This problem focuses on the collinearity problem.

- (a) Perform the following commands in R:

```

1 > set.seed(240)
2 > x1=runif(100)
3 > x2=0.5*x1+rnorm(100)/10
4 > y=2+2*x1+0.3*x2+rnorm(100)

```

The last line corresponds to creating a linear model in which y is a function of x_1 and x_2 . Write out the form of the linear model. What are the regression coefficients?

- (b) What is the correlation between x_1 and x_2 ? Create a scatterplot displaying the relationship between the variables.
- (c) Using this data, fit a least squares regression to predict y using x_1 and x_2 . Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0 , β_1 and β_2 ? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?
- (d) Now fit a least squares regression to predict y using only x_1 . Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?
- (e) Now fit a least squares regression to predict y using only x_2 . Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

- (f) Do the results obtained in (c)-(e) contradict each other? Explain your answer.
- (g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
1 > x1=c(x1, 0.1)
2 > x2=c(x2, 0.8)
3 > y=c(y, 6)
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

Problem 6 (15 pts, ★☆☆☆☆)

This question should be answered using the *Carseats.RData* data set.

- Fit a multiple regression model to predict *Sales* using *Prices*, *Urban*, and *US*.
- Provide an interpretation of each coefficient in the model. Be careful – some of the variables in the model are qualitative!
- Write out the model in equation form, being careful to handle the qualitative variables properly.
- For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?
- On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. Conduct a F test on the covariates you include.
- How well do the models in (a) and (e) fit the data?
- Using the model from (e), obtain 95% confidence intervals for the coefficient(s).
- Is there evidence of outliers or high leverage observations in the model from (e)?