STAT GR5205 Homework 2 [100 pts]

Due Wednesday, October 9th at 8:40am

Problem 1 (**2.7 KNN**)

Sixteen batches of plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume the first-order regression model (1.1) is appropriate (**(2.1) in the notes**).

Data not displayed

Use `R` to perform the following tasks:

i. Estimate the change in the mean hardness when the elapsed time increases by one hour. Use a 99 percent confidence interval. Interpret your interval estimate.

ii. The plastic manufacturer has stated that the mean hardness should increase by 2 Brinell units per hour. Conduct a two-sided test to decide whether this standard is being satisfied; use $\alpha = .01$.

iii. Set up the ANOVA table.

iv. Test by means of an F-test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use $\alpha = .01$.

v. Does t^2_{calc} from part [ii] equal f_{calc} from part [iv]? Explain why this identity holds or does not hold.

vi. Construct 95% Bonferroni joint confidence intervals for estimating both the true intercept β_0 and the true slope β_1.

vii. Construct 95% Bonferroni joint confidence intervals for predicting the true average hardness corresponding to elapsed times 20, 28 and 36 hours.

Problem 2

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

i. Assuming $H_0 : \beta_1 = 0$ is true, use `R` to simulate the sampling distribution of the F-statistic

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}.$$

Assume $\beta_0 = 10$, $\sigma = 3$, $n = 30$ and run the loop 10,000 times to generate the sampling distribution. Run the following code preceding the loop so that everyone has the same seed and X data vector. Fill in the missing code to receive full credit.

```
# Set seed
set.seed(0)
# Assign sample size and create x vector
n <- 30
# Empty list for f-statistics
f.list <- NULL
x <- runif(n,min=0,max=100)
# Run loop
for (i in 1:10000) {

# Fill in the body of the loop here...

}
```

ii. From the simulated sampling distribution, plot a histogram and overlay the *correct F-density* on the histogram. Adjust the bin-size to *breaks=50* in the histogram. Overlay the F-density in red.

iii. Compute the $95th$ percentile of both the simulated sampling distribution and the *correct* F-distribution. Compare these values.

Problem 3

Consider splitting the response values $y_1, \ldots, y_n$ into two groups with respective sample sizes n_1 and n_2. Define the **dummy** variable

$$x_i = \begin{cases} 1 & \text{if group one} \\ 0 & \text{if group two} \end{cases} \tag{1}$$

Show that the least squares estimators of β_1 and β_0 are respectively

$$\hat{\beta}_1 = \bar{y}_1 - \bar{y}_2 \quad \text{and} \quad \hat{\beta}_0 = \bar{y}_2,$$

where $\bar{y}_1$ and $\bar{y}_2$ are the respective sample means of each group.

Problem 4

Fusible interlinings are being used with increasing frequency to support outer fabrics and improve the shape and drape of various pieces of clothing. The article *Compatibility of Outer and Feasible Interlining Fabrics in Tailored Garments* gave the accompanying data on extensibility (%) at 100 gm/cm for both high-quality (H) fabric and poor-quality (P) fabric specimens.

H	1.2	.9	.7	1.0	1.7	1.7	1.1	.9	1.7
	1.9	1.3	2.1	1.6	1.8	1.4	1.3	1.9	1.6
	.8	2.0	1.7	1.6	2.3	2.0			
P	1.6	1.5	1.1	2.1	1.5	1.3	1.0	2.6	

Use `R` to perform the following tasks.

i. Create an appropriate graphic to visualize the relationship between extensibility and quality. Do you think there is a relationship between extensibility and quality? Make sure to label the plot.

ii. Using the indicator variable

$$x = \begin{cases} 1 & \text{if high quality} \\ 0 & \text{if low quality} \end{cases}$$

run a regression analysis to test if the average fabric extensibility differs per group.

Problem 5

i. Consider the *regression through the origin model* given by

$$Y_i = \beta x_i + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{2}$$

Derive the maximum likelihood estimators of β and σ^2.

ii. Consider the residuals e_i related to the regression through the origin model (2). Prove that

$$\sum_{i=1}^{n} e_i x_i = 0.$$

Also, in the regression through the origin model (2), is the sum of residuals equal to zero? I.e., is the following relation true?

$$\sum_{i=1}^{n} e_i = 0.$$

Explain your answer in a few sentences or less.

iii. Consider testing the null/alternative pair

$$H_0 : \beta = \beta' \quad \text{v.s.} \quad H_A : \beta \neq \beta'.$$

Note that β' is the hypothesized value. Show that the likelihood-ratio test can be based on the rejection region $|T| > k$ with test statistic

$$T = \frac{\hat{\beta} - \beta'}{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}x_i)^2/(n-1)}{\sum_{i=1}^{n} x_i^2}}}.$$

Note that k is some positive real number and $\hat{\beta}$ is the maximum likelihood estimator of β.

iv. Under H_0, what is the probability distribution of the above test statistic T?

Hints: To solve 5 Part iii:

(a) Compute the likelihood-ratio test statistic (λ) from Definition 2.4 on Page 45 of the class notes.

(b) When simplifying the expression, the following trick might be useful:

$$\sum_{i=1}^{n}(Y_i - \beta' x_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}x_i + \hat{\beta}x_i - \beta' x_i)^2.$$

(c) After simplifying $\lambda < c$, find a suitable transformation of λ that yields the desired test statistic and rejection rule.