Midterm 3 Review **Math 222**

These are problems similar to the ones that might be on midterm 3. Be sure to also review class workshops and problems from the textbook.

1. In each case, state the specific statistical procedure that is appropriate for the given situation. Be specific: identify the response variable and the explanatory variable(s). If there are any categorical variables present, state how many levels each categorical variable has.

   (a) You want to study whether men and women get different average amounts of sleep at night.

   Solution: Two-sample t-test. The explanatory variable is gender, and the response variable is hours of sleep at night.

   (b) You want to predict life satisfaction based on several factors, including income, regional cost of living, commuting time, and number of children.

   Solution: This is multiple regression. The response variable is life satisfaction, and the explanatory variables are income, cost of living, commuting time, and number of children.

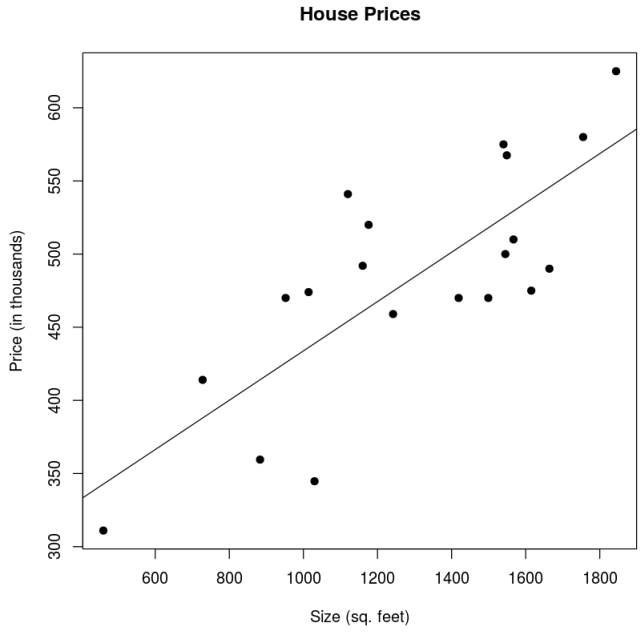
   (c) You want to determine if there are significant differences between the cost of housing in three different cities.

   Solution: This is 1-way ANOVA. The response variable is cost of housing. The explanatory variable is the city (3 levels).

2. Suppose you are performing one-way ANOVA to test for a difference in means for 4 groups. Each group contains 10 individuals that are randomly selected from a large population. Before conducting the test, you conduct a quick power computation for a specific alternative hypothesis where $\mu_1 = 10$, $\mu_2 = 11$, $\mu_3 = 12$ and $\mu_4 = 13$. You need to estimate σ for the computation, and so you choose $\sigma = 3$, which seems reasonable. Would the power be larger, smaller, or about the same if the true σ was actually larger than 3?

   Solution: The power would be larger if the variance were smaller.

3. The scatterplot below shows the relationship between size (in square feet) and price (in thousands of dollars) of a random sample of 20 houses sold recently in Arroyo Grande, CA.



Below is a summary of the least squares regression model for this scatterplot.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 265.22212   42.64202   6.220 7.21e-06 ***
myData$Size   0.16859    0.03188   5.288 5.00e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.31 on 18 degrees of freedom
Multiple R-squared:  0.6084, Adjusted R-squared:  0.5866
```

(a) Is the trend statistically significant? How can you tell?

Solution: The trend is significant because the p-value for the slope is 5.00 times 10^{-5}.

(b) If $SE_{\hat{\mu}} = 55.18$, find a 95% confidence interval for the mean home price of a 1200 square foot house.

Solution: The confidence interval for $\hat{\mu}$ is $\hat{y} \pm t^* SE_{\hat{\mu}}$ where t^* has $N-2 = 18$ degrees of freedom. Use the t-distribution chart, $t^* = 2.101$. Also, $\hat{y} = 265.22 + 0.16859(1200) = 467.5$. So the confidence interval is: 467.5 ± 115.9 or equivalently: 351.6 to 583.4 thousand dollars

(c) Find a 95% prediction interval for the price of a 1200 square foot house (recall that $SE_{\hat{y}}^2 = SE_{\hat{\mu}}^2 + s^2$ where s is the residual standard error).

Solution: The standard error in $\hat{y}$ is $\sqrt{55.18^2 + s^2}$ where s^2 is the variance of the residuals and is equal to the square of the residual standard error which is 51.31 in the chart above. So $SE_{\hat{y}} = \sqrt{55.18^2 + 51.31^2} = 75.35$. Then $\hat{y} \pm t^* SE_{\hat{y}}$ is 467.5 ± 158.3 which is from 309.2 to 625.8 thousand dollars.

(d) Use the fact that $R^2 = 0.6084$ and $s_y = \$79{,}801.5$ and $n = 20$ houses to fill in the following ANOVA table for this example.

	DF	SS	MS	F
Size				
Residuals				
Total				

Solution:

	DF	SS	MS	F
Size	1	7.36×10^{10}	7.36×10^{10}	27.97
Residuals	18	4.74×10^{10}	2.63×10^{9}	
Total	19	1.21×10^{11}	6.368×10^{9}	

4. This example is based on data from 78 seventh-grade students in a rural midwestern school. The researcher was interested in the relationship between the students' "self-concept" and their academic performance. The data included each student's grade point average (GPA) on a ten-point scale, score on a standard IQ test, and gender, taken from school records. Gender is coded as 1 for female and 2 for male. The final variable is each student's score on the Piers-Harris Children's Self-Concept Scale, a psychological test administered by the researcher. Below is a summary of the multiple linear regression model for this data in R.

```
Call:
lm(formula = gpa ~ iq + gender + concept, data = myData)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5769 -0.7493  0.1984  0.9577  2.4089

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.83463    1.28584  -2.205 0.030641 *
iq           0.08079    0.01336   6.045 5.78e-08 ***
gender      -0.82214    0.31354  -2.622 0.010630 *
concept      0.05048    0.01396   3.616 0.000548 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.323 on 73 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.561, Adjusted R-squared:  0.543
F-statistic:  31.1 on 3 and 73 DF,  p-value: 4.643e-13
```

(a) What is the formula for predicting GPA from IQ, Gender, and Self-Concept using this regression model?

Solution:

$$\text{GPA} = -2.83463 + 0.08079\,\text{IQ} - 0.82214\,\text{Gender} + 0.05048\,\text{Self-Concept}$$

(b) What percent of the variability in GPA is explained by this model?

Solution: The $R^2 = 56.1\%$.

(c) Estimate the GPA of a male student who has an IQ of 110, and a self-concept score of 60.

Solution:

$$\widehat{\text{GPA}} = 7.437$$

5. Do people from different cultures experience emotions differently? One study designed to examine this question collected data from 410 college students from five different cultures. 9 The participants were asked to record, on a 1 (never) to 7 (always) scale, how much of the time they typically felt eight specific emotions. These were averaged to produce the global emotion score for each participant. Here is a summary of this measure:

Culture	n	$\bar{x}$	SD
European American	46	4.39	1.06
Asian American	33	4.35	1.18
Japanese	91	4.72	1.13
Indian	160	4.34	1.26
Hispanic American	80	5.04	1.16

(a) Complete the ANOVA table below for these results by filling in the five missing entries:

	Df	SS	MS	F
Culture		31.268		
Residuals			1.4044	n/a
Total	409	600.04	1.4671	n/a

Solution: The filled in ANOVA table is:

	Df	SS	MS	F
Culture	4	31.268	7.817	5.566
Residuals	405	568.772	1.4044	n/a
Total	409	600.04	1.4671	n/a

(b) What are the null and alternative hypotheses for this ANOVA test?

Solution:

H_0 : The mean is the same for all cultures.

H_A : There are differences in the means.

(c) It turns out that the p-value for the F-statistic above is 2.27×10^{-4}. What does that mean in this situation?

Solution: We should reject the null hypothesis and conclude that there are statistically significant differences in the means.

(d) Why is it reasonable to assume that each group has the same population standard deviation in this example?

Solution: Because the sample standard deviations for each group are all very similar (all within a factor of 2).

(e) What number is the best estimate for the common standard deviation for each group?

Solution: The pooled standard deviation in ANOVA is $s_p = \sqrt{MSE}$. Here $s_p = 1.185$.

(f) Why don't we need to worry very much about whether the assumption of normality is met for this data?

Solution: The sample sizes are so large (the smallest is group has 30 people), so normality will not be a big concern.

(g) Recall that the confidence interval for the difference between the means of two groups is $\bar{x}_A - \bar{x}_B \pm t^{**} s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$, where t^{**} is the adjusted critical value with the Bonferroni correction. According to the Bonferroni method, what adjusted confidence level should we use to be 95% certain that we capture the true difference in population mean given that there are 10 possible pairs of groups to compare? *You don't need to compute the confidence interval.*

Solution: There are 10 different pairwise comparisons to consider. Therefore we need to use a 1 - 0.05/10 = 0.995 = 99.5% confidence level for each confidence interval.

6. Determine whether each statement below is True or False.

(a) In one way ANOVA the response variable is categorical and the explanatory variable is quantitative.

Solution: False. The response variable is quantitative and the explanatory variable is categorical.

(b) Linear regression assumes that the residuals are normally distributed.

Solution: True.

(c) One of the assumptions made in the application of the one-way ANOVA F test is homogeneity of variance (i.e., the variances for all populations are assumed to be the same).

Solution: True.

(d) If the data in each group is strongly right skewed, it is okay to do an ANOVA F-test as long as the sample sizes are large.

Solution: True.

(e) When testing differences between population means using the One-Way Analysis of Variance (ANOVA) statistical method, the region of rejection is always in the left tail of the F distribution.

Solution: False. The rejection region is the right tail of the F-distribution.

(f) In multilinear regression models, removing variables always decreases the adjusted R^2.

Solution: False.

(g) If the null hypothesis is rejected when conducting a one-way ANOVA F-test, then there are statistically significant differences between all pairs of means.

Solution: False. Not all pairs have to have significant differences.

7. Esophageal cancer can spread to the lymph nodes, and the larger the tumor is, the more likely it is to spread. Below is the R output for a logistic regression model based on a sample of 31 cancer patients. The explanatory variable is the size of the tumor in centimeters and the response variable is whether or not the cancer has spread to the lymph nodes.

```
## Call:  glm(formula = spread ~ size, family = "binomial", data = cancer)
##
## Coefficients:
## (Intercept)         size
##      -2.086       0.5117
```

(a) What is the formula for the log-odds in the model described above?

Solution:

$$\log(\text{odds}) = -2.086 + 0.5117(\text{size})$$

(b) What are the odds that the cancer has spread to the lymph nodes if a patient has a 6 cm tumor?

Solution: The odds are $\exp(-2.086 + 0.5117(6)) = 2.676$.

(c) What is the probability that the cancer has spread to the lymph nodes if a patient has a 6 cm tumor?

Solution: The probability is $\dfrac{2.676}{2.676 + 1} = 72.8\%$.

(d) How many times higher are the odds of the cancer spreading for every extra centimeter of tumor?

Solution: The odds-ratio for each extra centimeter is $\exp(0.5117) = 1.668$, so the odds of the cancer spreading is 1.668 times higher for every extra centimeter.