Name _____ Spring, 2016

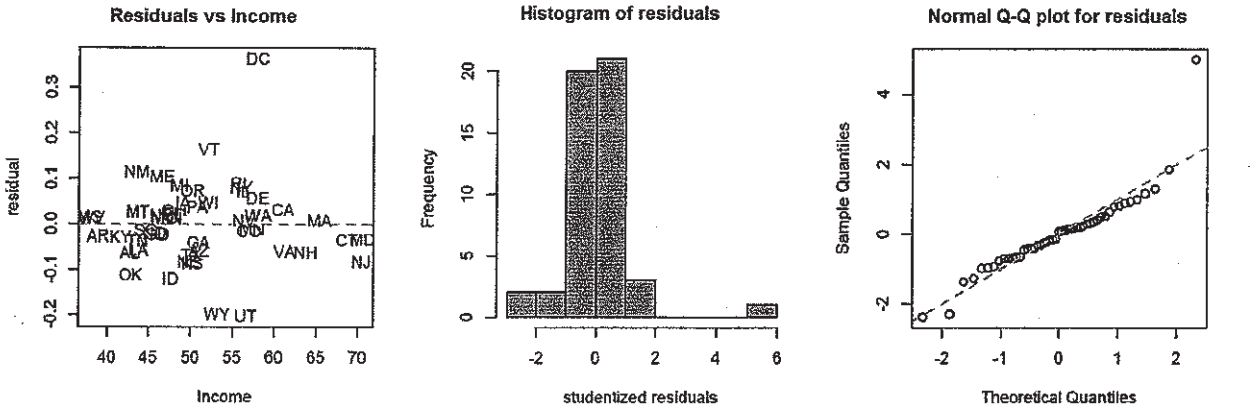
Applied Statistics Comprehensive Examination

- Calculators are permitted on this examination.
- When you are asked to construct a confidence interval, always interpret the interval in terms of the problem.
- When you are asked to perform a hypothesis test, always write down the null and alternative hypotheses, and write the conclusions in terms of this problem.
- There are 200 points for the entire examination.

1. (20 points) A recent study found that the average movie length for the top 25 most popular films in 2015 (as rated by IMDb) is 11 minutes longer than in 1985. However, a local movie buff hypothesizes that since the number of films made by lower-budget production studios has increased significantly, the average movie length for all films released into theaters (not just the top 25) would actually be shorter in 2015 than in 1985. The movie buff obtains two independent random samples of 10 movies from 1985 and 2015 with the following summary measures in minutes: $\overline{y}_{1985} = 105.0$, $s_{1985} = 14.5$, $\overline{y}_{2015} = 100.9$, $s_{2015} = 11.3$. Test the local movie buff's claim at the 0.05 level. You may make any assumptions you feel are necessary, but list these assumptions explicitly.

2. (15 points) In this study, we collected the results for the 2008 presidential election as well as income information for each state/district in the lower continental US. The goal of the study is to relate Obama's share of the presidential vote in each state, y, to the median income (in \$1000) for each state, x.

Consider the following plots for the model $y = \beta_0 + \beta_1 x + \epsilon$.



(a) (10 points) What problem(s) do you see? How would you fix it(them)? Justify your changes to the regression.

(b) (5 points) Do you think that your fix will lead to big changes in the least-squares regression coefficients? Why or why not?

3. (20 points) The editorial staff of the student newspaper at a large state university is interested in predicting how students will vote in an upcoming election. To gather information, they draw a stratified random sample of 200 students by taking 50 from each class year. They then interview the students, obtaining the preference counts given in the table below. Using level 0.05, test for evidence that the level of support for Candidate A differs from one class year to another.

Candidate	Freshman	Sophomore	Junior	Senior
A	28	29	31	24
B	22	21	19	26

4. (30 points) Suppose a (3×2) factorial design is run in a completely randomized manner and results in the data which appear below. The data will be analyzed assuming a model of the form: $y_{ijk} = \mu+\alpha_i+\beta_j+\gamma_{ij}+\epsilon_{ijk}$ where α_i is a fixed effect corresponding to rows, β_j is a fixed effect corresponding to columns, γ_{ij} is a fixed effect corresponding to the interaction between rows and columns and ϵ_{ijk} is a random error term which is assumed to be $IIDN(0,\sigma^2)$. Note that $i = 1,2,3$, $j = 1,2$ and $k = 1,2,\ldots n_{ij}$.

3,3	5,7
6	7,8,9
8,9,10	9,11

(a) (10 points) Write the population marginal means for row 2 and column 1 and calculate estimates of these two population marginal means.

(b) (10 points) Write any two linearly independent contrasts in interaction effects. State whether or not these two contrasts are orthogonal and justify your answer.

(c) (10 points) Determine if $\alpha_1 - \alpha_3 + \gamma_{11} + \gamma_{31} - 2\gamma_{32}$ is estimable and justify your answer.

5. (20 points) You are tasked with conducting a test of the hypotheses $H_0 : \mu \leq 3$ vs. $H_a : \mu > 3$ based on a sample of 15 observations in a case where the standard deviation σ is known to be 5 and the population distribution is normal. Suppose you decide to reject H_0 whenever $\overline{Y} > 5$.

   (a) (10 points) What is the probability of Type I error for this test?

   (b) (10 points) What is the power for this test if the true mean is 6?

6. (35 points) A study examines the relationship between the size of the acorn and the geographic range of the oak tree species. Twenty seven (27) species of oak from the Atlantic region and 11 from the California region were studied. The size of each species' acorns was measured to see whether acorn size is related to geographic range. It is suggested that a plant's seed size may have an effect on the geographic range of the plant because larger acorns can only be carried away by larger animals who in turn have a wider territorial range. Since both variables are skewed and the transformed variables are considered as follows,

y = (natural) log of the species range area (100km^2)
x_1 = (natural) log of the acorn (seed) size (cm^3)

The following table provides residual sums of squares from 5 different models using y, x_1 and the indicator x_2 that is 1 if the species is from the Atlantic region and 0 if it is from the California region.

Model	Terms in Model	Residual Sum of Squares	Residual degrees of freedom
A	Intercept	82.9	37
B	Intercept, x_1	78.9	36
C	Intercept, x_2	36.1	36
D	Intercept, x_1, x_2	23.2	35
E	Intercept, x_1, x_2, x_1x_2	21.9	34

(a) (5 points) What is the estimate of σ_ϵ^2 from Model D?

(b) (5 points) Calculate and interpret the value of R^2 for Model D.

(c) (10 points) Test for the overall significance of regression for Model D at $\alpha = 0.05$.

(d) (10 points) Given the presence of x_1 and x_2 in Model E, is the interaction term x_1x_2 significant? Conduct the appropriate hypothesis test at $\alpha = 0.05$.

(e) (5 points) Suppose Model D is adequate, the fitted model is obtained below,

$$\hat{y} = 5.7 + 2.8x_1 + 0.63x_2.$$

Interpret the coefficient of x_2 in terms of the species range area.

7. (25 points) A local manufacturer is interested in finding a faster way to fill trucks with manufactured product. Three possible strategies have been suggested by workers and management. On one particular morning when 12 trucks are scheduled to be loaded, the loading dock manager randomly assigns four trucks to each strategy. She then records the time in minutes required to load each truck, obtaining the numbers in the table below.

Strategy	Values	Sample Mean	Sample SD
1	10, 9, 12, 7	9.5	2.1
2	13, 8, 12, 11	11.0	2.2
3	15, 12, 21, 10	14.5	4.8

(a) (8 points) State an appropriate mathematical model, including all assumptions.

(b) (10 points) Using level 0.05, test for a difference in mean loading time for the three strategies.

(c) (7 points) How would you check the assumptions? You need not do the checking.

8. (20 points) An experiment was conducted to investigate the warping of copper plates. The two factors studied were temperature (50° and 100°) and copper content of the plates (40% and 80%) and the response variable was the amount of warping. Three observations were obtained at each of the four combinations of the two factors and the resulting averages were recorded:

	Content (%)	
Temp (°)	40	80
50	2	4
100	10	8

(a) (10 points) Write the normal equations assuming the following model: $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$ where α_i is a fixed effect corresponding to temperature, β_j is a fixed effect corresponding to content, γ_{ij} is a fixed effect corresponding to the interaction between temperature and content and ϵ_{ijk} is a random error term which is assumed to be $IIDN(0, \sigma^2)$. Note that $i = 1, 2$, $j = 1, 2$ and $k = 1, 2, 3$.

(b) (10 points) Briefly describe why restrictions are often used to obtain a solution to the normal equations. Name a method other than restrictions which may be used to solve the normal equations.

9. (15 points) As part of a feasibility study for increasing the capacity of a large hotel, a statistician wishes to estimate the standard deviation in the number of rooms rented each night. For 31 randomly chosen nights, he finds a sample standard deviation of 15.7 rooms.

   (a) (10 points) Construct a 99% confidence interval for the standard deviation in the number of rooms rented nightly at this large hotel.

   (b) (5 points) Given the sample size of 31, is a normality assumption necessary to construct the confidence interval in part a? Briefly explain.