Name: _____ Spring, 2018

Applied Statistics Comprehensive Examination

- Calculators are permitted on this examination.
- When you compute a confidence interval, always give an interpretation of the interval in the context of the problem.
- When you perform a hypothesis test, always write down the null and alternative hypotheses, and write the conclusion in the context of the problem.
- There are 200 points on this examination.
- You must give complete explanations to receive full credit.
- Please put your answers and explanations on the separate sheets provided.

1. (20 points) A local businesswoman owns two car washes, one in a downtown area and one in a suburban shopping district. Each car wash allows customers to choose from one of three service options: standard car wash, car wax and polish, and car detailing. In preparation for an upcoming advertising campaign, the businesswoman wants to know whether the customers' choices are the same or different at the two locations. To answer this question, she obtains a systematic sample of 100 cars at each location, with the following results:

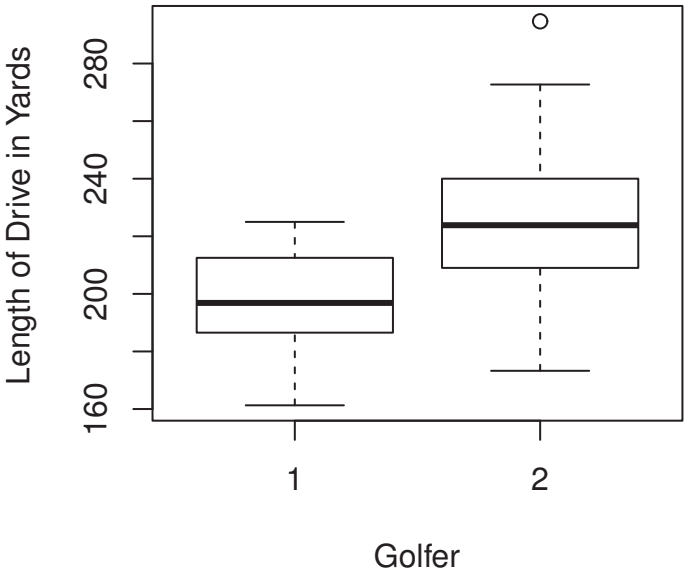
	Location	
Service	Downtown	Suburban
Car Wash	52	68
Wax and Polish	22	18
Car Detailing	26	14

   (a) (15 points) Test whether the distribution of service choices is different at the two locations. Use significance level 0.05.

   (b) (5 points) State the assumptions and conditions needed for the test in (a), and discuss whether they are met.

2. (35 points) Two local golfers recently visited a driving range. Each golfer hit 40 drives, and the length in yards of each drive was recorded. Summary statistics and boxplots for the two golfers are given below.

Golfer	Sample Mean	Sample Standard Deviation
1	198.4	16.1
2	224.7	27.6



(a) (10 points) Using significance level 0.05, test for a difference in the mean drive length for the two golfers.

(b) (5 points) State the assumptions needed for the test in (a), and discuss whether they are met.

(c) (10 points) Using significance level 0.05, test for a difference in the variance of drive length for the two golfers.

(d) (5 points) State the assumptions needed for the test in (c), and discuss whether they are met.

(e) (5 points) Your boss glances at the two boxplots, commenting that both distributions appear to be unimodal. Please discuss the validity of this comment.

3) (20 Points) Parents are frequently concerned when their child seems slow to begin walking. A recent paper reported on an experiment which compared the effects of several different treatments on the age (in months) at which a child first walks. Children in the first group were given special walking exercises for 12 minutes per day beginning at age 1 week and lasting 7 weeks. The second group of children received daily exercises but not the walking exercises administered to the first group. The third group was a control group – they received no special treatment. Summary statistics in months are given for each of the three groups in the following table:

Group	n	Mean	Standard Deviation
1 – Special Walking Exercises	6	9.8	0.7
2 – Daily Walking Exercises	6	10.5	0.8
3 – Control Group	5	12.4	1.0

a. (10 Points) Complete the following ANOVA table by filling in the x's and test at the 0.05 level to determine whether any of the population means differ.

Source	DF	Sum of Squares	Mean Square	F
Treatment	x	x	x	x
Error	x	9.804	x	
Total	x	29.404		

b. (10 Points) Construct and interpret a 95% confidence interval for the difference between the mean of the active treatment groups versus the mean of the control group.

4) (35 Points) The table below displays the means of three observations which were obtained from a study of the effect of oven temperature and baking time on the life (in hours) of an electrical component.

Oven Temperature (F)	Baking Time (min) 5	10
600	199	215
620	181	202
640	144	140

The total number of observations is 18 (3 per cell) and the mean squared error from an ANOVA table containing factors for temperature, baking time, and the interaction is 26. Use the effects model given below to answer the questions that follow.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where
α_i= the effect of the i^{th} oven temperature. i = 1, 2, 3
β_j= the effect of the j^{th} baking time. j = 1, 2
γ_{ij}= the interaction effect of the i^{th} oven temperature and j^{th} baking time.

a. (5 Points) Create an interaction plot and discuss the type of interaction in the data.
b. (5 Points) Write a complete set of orthogonal contrasts for oven temperature.
c. (10 Points) At the 0.05 level, test to determine whether the interaction between oven temperatures 600 and 620 (F) and baking times 5 and 10 minutes is significant.
d. (10 Points) Using only one observation per oven temperature and baking time combination, write the design matrix for the effects model above using set-to-zero restrictions.
e. (5 Points) Is $(\alpha_1 - \alpha_2)$ estimable? Explain.

5. (20 points) We are given 25 independent observations from a normal distribution with unknown mean μ and known standard deviation 10. Suppose that we test $H_0 : \mu = 100$ against $H_a : \mu > 100$ by rejecting H_0 when the sample mean exceeds 103.

   (a) (10 points) Find the α level for the test.

   (b) (10 points) Sketch the power curve for the test over the interval $[100, 110]$, showing the exact power for at least three different points in the interval.

6. (15 points) Villanova basketball player Mikal Bridges made 44 of his 112 three-point attempts during the 2016-2017 season. During the 2017-2018 season, he made 104 of his 239 three-point attempts.

   (a) (10 points) Using significance level 0.10, test for evidence that Bridges improved his three-point success rate between 2016-2017 and 2017-2018.

   (b) (5 points) State the assumptions needed for the test in (a) and discuss whether they are met.

7. (55 points) One key method of assessing fertility in women is the antral follicle count (AFC), which can be measured with non-invasive ultrasound. Researchers gathered various data on women who were having trouble getting pregnant. They were interested in how well AFC can be predicted by the other variables included in the study. The study variables included:

AFC	Antral follicle count
Age	Age in years
FSH	Maximum follicle stimulation hormone level
FL	Fertility level
TGL	Total gonadotropin level
Ooctyes	Number of egg cells
Embryos	Number of embryos

a. (10 points) The (partial) output from the **full model** can be found on the next page. It was obtained using *AFC* as the dependent variable and all other variables as independent variables. Recall that "residual standard error" is the square root of the mean squared error. Conduct the (global) hypothesis test on the regression model at the 5% level of significance.

b. (10 points) For the **full model**, interpret the value of the coefficient for *Oocytes* and perform a hypothesis test for this coefficient at the 5% level of significance.

c. (15 points) The researchers are interested in determining whether *Oocytes* and *Embryos* are predictors of *AFC* when controlling for the other variables in the model. Output from two reduced models is given below. Conduct the appropriate partial F-test at the 5% level of significance.

d. (5 points) An analyst wanted to conduct a stepwise regression based on the **full model**. She suggests two possible approaches as follows:
   i. Her first approach is to use a backwards elimination process by removing *Age* since it has the highest p-value at 0.577. She will then rerun the model without this predictor to determine the next step.
   ii. Her second approach is to use a forward selection process by selecting *TGL* for the model since it has the lowest p-value at 9.5×10^{-6}. She will then rerun models with *TGL* and each other predictor to determine the next step.

   Which of these two would you suggest she pursue? Briefly explain your answer.

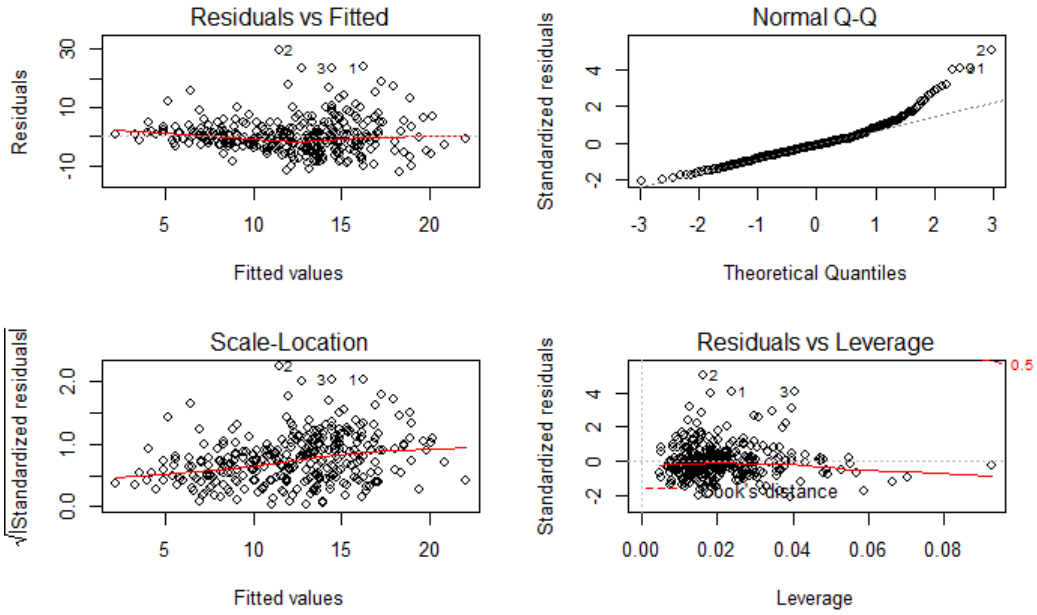
e. (15 points) State the assumptions / conditions for multiple linear regression. For each one, verify it using the plots below, which are from the **full model**. Please identify specific plots using the plot title included. If you cannot verify it using the plots below, state how you would determine whether the assumption / condition is met.

This page was intentionally left blank.

Full Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.654828   3.050252    6.12  2.7e-09
Age         -0.045596   0.081570   -0.56    0.577
FSH         -0.425999   0.195025   -2.18    0.030
FL          -0.048149   0.021686   -2.22    0.027
TGL         -0.001385   0.000308   -4.50  9.5e-06
Oocytes      0.208886   0.087005    2.40    0.017
Embryos      0.179317   0.122728    1.46    0.145

Residual standard error: 5.94 on 326 degrees of freedom
Multiple R-squared:  0.277,     Adjusted R-squared:  0.264
F-statistic: 20.8 on 6 and 326 DF
```



Reduced Model 1 (for part c)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.585477   2.953907    8.32  2.3e-15
Age         -0.038900   0.084562   -0.46   0.6458
FSH         -0.628208   0.198671   -3.16   0.0017
FL          -0.063778   0.022326   -2.86   0.0046
TGL         -0.001612   0.000316   -5.10  5.8e-07

Residual standard error: 6.17 on 328 degrees of freedom
Multiple R-squared:  0.214,     Adjusted R-squared:  0.204
F-statistic: 22.3 on 4 and 328 DF
```

Reduced Model 2 (for part c)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.901      0.788    8.75  < 2e-16
Oocytes        0.344      0.091    3.78  0.00019
Embryos        0.195      0.132    1.48  0.13933

Residual standard error: 6.4 on 330 degrees of freedom
Multiple R-squared:  0.151,     Adjusted R-squared:  0.146
F-statistic: 29.4 on 2 and 330 DF,  p-value: 1.85e-12
```