ST518

Fall Semester, 2021

Midterm Exam

Name:_____

Directions: Answer questions as directed. Please show work. Partial credit may be awarded for correct expressions given in incomplete answers. To save time, it is perfectly ok to give answers as numeric expressions without carrying out every last operation on the calculator. If the question asked for the standard error of the estimated mean of a population at $x = 14$ in a simple linear regression, the following response would receive full credit:

$$\widehat{SE}(\hat{\beta}_0 + 14\hat{\beta}_1) = \sqrt{16.4\left(\frac{1}{38} + \frac{(14 - 10.8)^2}{190}\right)}$$

or in typed format,

```
estimated SE(betahat0+14betahat1)=sqrt(16.4(1/38+(14-10.8)^2/190))
```

There are 6 problems worth a total of 100 points. You may skip any single numbered problem and all its components and still receive full credit, but only if you type `SKIP` for that problem.

1. (20 pts) Consider a simple linear regression of y on x, with model $Y_i = \beta_0 + \beta_1 x_i + E_i$ where $E_i \overset{iid}{\sim} N(0, \sigma^2)$. SAS PROC REG was used to fit the model with output below:

```
The REG Procedure
                          Analysis of Variance

                                    Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       1     187.26667      187.26667      26.90    0.0013
Error                       7      48.73333        6.96190
Corrected Total             8     236.00000

Root MSE              2.63854    R-Square      xxxxxx
Dependent Mean       50.00000    Adj R-Sq      xxxxxx

                     Parameter        Standard
Variable     DF       Estimate           Error    t Value    Pr > |t|
Intercept     1       23.50000         5.18466       4.53      0.0027
x             1        1.76667         0.34063       5.19      0.0013
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
 x     y     fitted    residual       qnorm

17     50    53.5333    -3.53333    -1.28155
14     46    48.2333    -2.23333    -0.84162
12     43    44.7000    xxxxxxxx    xxxxxxxx
15     49    50.0000    -1.00000    xxxxxxxx
13     46    46.4667    -0.46667    xxxxxxxx
19     57    57.0667    -0.06667     xxxxxxx
18     58    55.3000     2.70000     0.52440
11     46    42.9333     3.06667     0.84162
16     55    51.7667     3.23333     1.28155
```

(a) Report an unbiased estimate of the slope.

(b) Report an unbiased estimate of the change in $E(Y|X = x)$ as x is increased by 2 units.

(c) Report an estimate of the standard deviation of the estimate in (b).

(d) Estimate the mean of the response when $x = \overline{x} = 15$.

(e) Report an estimate of the standard error of the estimate in part (d).
(Note that $\text{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x) = \sigma^2(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}})$.)

(f) Report an estimate of the standard deviation of a future observation Y_{10}, given that $X_{10} = \overline{x}$, $SD(Y_{10}|X_{10} = \overline{x})$.

(g) For the $i = 3$ observation, $x_3 = 12$. Report the fitted value $\widehat{y}_3$ and the residual for the $i = 3$ observation.

(h) Given that $\overline{x} > 0$, are the estimated intercept and slope (independent/positively correlated/negatively correlated). (Choose one answer.)

(i) How many observations were there?

(j) Quantiles from the normal distribution to be used in a normal q-q plot appear as `qnorm`. What value should be used for the $i = 5^{th}$ ordered residual, $e_{(5)} = -0.47$?

2. (10 pts)

Consider a linear regression of y on x. Suppose the observed variance among $n = 11$ observations of y is $s_y^2 = 16$ and the sample correlation is $r_{xy} = 0.5$. The sample variance of the x variable is denoted s_x^2, but you do not need to know its value.

(a) Report the coefficient of determination from the regression.

(b) Estimate the difference $E(Y|X = \overline{x} + 2s_x) - E(Y|X = \overline{x})$.

3. (10 pts)

Recall the fuel efficiency dataset from the R package entitled `mtcars`. Two explanatory variables were added for the square of weight and horsepower (`wt2` and `hp2` respectively). A multiple linear regression model was fit using SAS with output below.

```
The SAS System
The REG Procedure
Dependent Variable: mpg
                            Analysis of Variance

                                     Sum of            Mean
Source                    DF        Squares          Square    F Value    Pr > F
Model                      5     1003.19216       200.63843      42.46    <.0001
Error                     26      122.85503         4.72519
Corrected Total           31     1126.04719

                  Parameter      Standard
Variable    DF     Estimate         Error  t Value  Pr > |t|    Type I SS   Type II SS
Intercept    1     50.27529       5.44339     9.24    <.0001         12916    403.07828
am           1     -0.27556       1.37255    -0.20    0.8424     405.15059      0.19045
wt           1     -9.56268       2.87219    -3.33    0.0026     442.57690     52.37831
wt2          1      0.88147       0.34375     2.56    0.0165      86.61636     31.07082
hp           1     -0.09460       0.03256    -2.91    0.0074      47.61265     39.89652
hp2          1   0.00017753    0.00008374     2.12    0.0437      21.23567     21.23567
```

(a) How many cars were there in this study?

(b) What is the multiple coefficient of determination of the fitted model?

(c) `am` is an indicator variable for whether a car has an automatic transmission. Use the regression model to estimate the fuel efficiency (`mpg`) among cars with automatic transmissions (`am=1`), that weigh 2000 pounds (`wt=2`), and that have 70 horsepower (`hp=70`). A expression involving only numbers suffices.

(d) Report an F-ratio (or t-statistic if you prefer) and degrees of freedom for a test of equality of fuel efficiency for cars with and without an automatic transmission ...

   i. after controlling for effects of weight and horsepower

   ii. without controlling for effects of weight and horsepower.

(e) Report the F-ratio and degrees of freedom for a test comparing the full model fit here with a reduced model in which there is no dependence on horsepower or its square.

4. (20 pts)

A statistics professor (not for ST518) is an enthusiastic golfer. He designs an experiment to compare the distances travelled by golfballs from 4 manufacturers. He believes (H_0) that the manufacturers are all the same and all balls will have mean distance $\mu = 250$ but will vary from 210 to 290 yards when he hits them. Let Y_{ij} denote the distance of the j^{th} ball from manufacturer $i = 1, ..., 4$. His colleague suggests that he choose the sample size necessary to detect an alternative with means as or more different than those below:

$$H_a : E(Y_{1j}) = 220, E(Y_{2j}) = 240, E(Y_{3j}) = 260, E(Y_{4j}) = 280$$

but with the same variances as under his null belief. He will conduct an analysis of variance under the usual assumptions.

(a) What are those assumptions? Write out a factorial effects model for Y_{ij}.

(b) If he uses a total of $N = 4(3) = 12$ golf balls (3 from each manufacturer), what will the distribution of the F-ratio, $F = MS(\text{manufacturer})/MS(E)$ be

i. under H_0?

ii. under H_a?

(Give degrees of freedom for both parts.)

(c) The appropriate $\alpha = 0.05$ critical value for the experiment which uses a total of $N = 12$ golf balls is 4.07. The areas under the F-distributions to the left of this critical value are 0.95 under H_0 and 0.28 under H_a.

i. What is the power of his experiment to detect the putative alternative?

ii. What is the chance that he fails to detect a real difference as large as that specified by his colleague?

iii. Suppose the variability of distances (σ) is not as large as speculated for the calculations above. Will the power go up or down or stay the same?

iv. Consider the following alternative instead:

$$H_a : E(Y_{1j}) = 230, E(Y_{2j}) = 250, E(Y_{3j}) = 250, E(Y_{4j}) = 270$$

Will the power be equal to, greater than or less than (i) above?

v. Suppose he uses only two balls per manufacturer instead of three. Will the power go up or down or stay the same?

vi. Suppose he relaxes the amount of evidence he requires to be convinced that manufacturer matters and instead adopts a level of significance of $\alpha = 0.10$. Will the power go up, down or stay the same?

5. (20 pts) A statistics professor (not for ST518) who is an enthusiastic golfer. He runs a randomized experiment to compare the distances travelled by golfballs from 4 manufacturers. Let Y_{ij} denote the distance travelled by the j^{th} ball from manufacturer $i = 1, ..., 4$. A factorial effects model with fit with SAS with output below

   (a) Estimate the mean difference in distance between golfballs from manufacturer 4 and manufacturer 3. Report a standard error and associated degrees of freedom.

   (b) Consider all six pairwise comparisons among the manufacturers. What is the definition of the familywise/experimentwise error (FWE) rate?

$$FWE = ?$$

   (c) Use the output from Tukey's procedure to identify which of the six differences among treatment means may be declared significant at familywise error rate 0.05.

   (d) What does HSD abbreviate?

   (e) Use Fisher's LSD procedure to conduct all 6 pairwise comparisons among the four averages. Be sure to describe the protocol for employing the procedure. Use the 95^{th} percentile of the appropriate t distribution, `qt(.975,8)=2.31` to show that Fisher's LSD is 29.0.

```
The GLM Procedure

Class             Levels    Values
manufacturer           4    1 2 3 4

Dependent Variable: dist
                                   Sum of
Source                      DF       Squares   Mean Square  F Value  Pr > F
Model                        3   3648.250000   1216.083333     5.15  0.0284
Error                        8   1888.000000    236.000000
Corrected Total             11   5536.250000

Source                      DF     Type I SS   Mean Square  F Value  Pr > F
manufacturer                 3   3648.250000   1216.083333     5.15  0.0284


Level of               -------------dist------------
manufacturer     N             Mean          Std Dev

1                3       222.333333        7.0237692
2                3       242.000000       22.0680765
3                3       245.333333       13.5030861
4                3       271.333333       15.0111070

Tukey's Studentized Range (HSD) Test for dist

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                                     0.05
Error Degrees of Freedom                     8
Error Mean Square                          236
Critical Value of Studentized Range    4.52877
Minimum Significant Difference          40.168
```

6. (20 pts) Consider a crossed, completely randomized design with two two-level factors A and B in which the following averages were observed for the four treatment combinations:

Level of Factor A	Level of Factor B 1	2	Average
1	14	10	12
2	20	12	16
Average	17	11	

Estimate "effects" by subtracting level 1 means from level 2 means:

(a) Estimate the simple effect of factor B at the first level of A.

(b) Estimate the simple effect of factor B at the second level of A.

(c) Estimate the main effect of factor B.

(d) Estimate the main effect of factor A.

(e) Estimate the interaction effect.