Name _____

Spring 2019 **STAT 587C** **Final exam**
(100 points)

Instructions:

1. Full credit will be given only if you show your work.
2. The questions are not necessarily ordered from easiest to hardest.
3. You are allowed to use any resource except aid from another individual.
4. Aid from another individual will automatically earn you a 0.
5. Feel free to tear off the last page. There is no need to turn it in.

Regression calculation

1. Suppose we have the following summary statistics for 100 pairs of response-explanatory variables with a sample correlation of -0.7.

	Explanatory (x)	Response (y)
mean	55	-221
standard deviation	9	49

Provide estimates of the following quantities. (2 points each)

Answer:

```
n = 100
Xbar = 55
Ybar = -221
s_X  = 9
s_Y = 49
r_XY = -0.7

SXX <- (n-1)*s_X^2
SYY <- (n-1)*s_Y^2
SXY <- (n-1)*s_X*s_Y*r_XY
```

(a) Intercept, $\hat{\beta}_0$

Answer:

```
beta1 <- SXY/SXX
beta0 <- Ybar - beta1 * Xbar
beta0
## [1] -11.38889
```

(b) Slope, $\hat{\beta}_1$

Answer:

```
beta1
## [1] -3.811111
```

(c) Coefficient of determination, R^2

Answer:

```
R2 <- r_XY^2; R2
## [1] 0.49
```

(d) Residual sum of squares, SSE

Answer:

```
SSE <- SYY*(1-R2); SSE
## [1] 121226.5
```

(e) Error variance, $\hat{\sigma}^2$

Answer:

```
sigma2 <- SSE/(n-2); sigma2
## [1] 1237.005
```

(f) Standard error for the intercept, $SE(\hat{\beta}_0)$

Answer:

```
sigma = sqrt(sigma2)
SE_beta0 <- sigma*sqrt(1/n + Xbar^2/SXX); SE_beta0
## [1] 21.88617
```

(g) Standard error for the slope, $SE(\hat{\beta}_1)$

Answer:

```
SE_beta1 <- sigma*sqrt( 1/SXX ); SE_beta1
## [1] 0.3927585
```

(h) Mean response when the explanatory variable is 70, $E[Y|X = 70]$

Answer:

```
x = 70
beta0 + beta1 * x
## [1] -278.1667
```

(i) Standard error of the mean response when X is 70, $SE(E[Y|X = 70])$

Answer:

```
sigma*sqrt( 1/n + (Xbar-x)^2/SXX )
## [1] 6.861369
```

(j) Standard error of prediction when X is 70, $SE(Pred\{Y|X = 70\})$

Answer:

```
sigma*sqrt( 1 + 1/n + (Xbar-x)^2/SXX )
## [1] 35.83411
```

Model comparisons

2. An unreplicated completely randomized block design (CRBD) experiment is being designed to study the effect of laser etching on hardness of Boron Nitride. The experiment has 2 blocks, 3 levels of laser intensity, and 4 levels of laser speed.

   (a) Complete the degrees of freedom portion of the following ANOVA table. (12 points)

Factor	df
Block	
Intensity	
Speed	
Intensity:Speed	
Error	
Total	

Answer:

Factor	df
Block	1
Intensity	2
Speed	3
Intensity:Speed	6
Error	11
Total	23

   (b) The following table provides residual sums-of-squares for a sequence of nested models.

Term included				
Block	Intensity	Speed	Intensity:Speed	SSE
X				55
X	X			43
X	X	X		30
X	X	X	X	8

Conduct an F-test to compare the model with Block to the model that includes Block, Intensity, Speed, and the interaction between Intensity and Speed. Report the numerator and denominator degrees of freedom, estimate for $\hat{\sigma}^2$, F-statistic, p-value, and an interpretation for the test. (8 points)

Answer:

```
n_df = 11 # 2+3+6
d_df = 11 # error degrees of freedom
SSE_full = 8
SSE_reduced = 55
sigma2 = 8/11 # SSE for full model/error degrees of freedom
Fstat = (SSE_reduced-SSE_full)/n_df/sigma2; Fstat
## [1] 5.875
p = 1-pf(Fstat,n_df,d_df); p
## [1] 0.003324157
```

$$F_{11,11} = \frac{(55-8)/11}{8/11} = 5.875$$

$$p = P(F_{11,11} > 5.875) = 1 - P(F_{11,11} < 5.875) = 0.0033242$$

The data are incompatible with the null model which is the regression model whose mean only includes block. If we believe the other regression model assumptions, i.e. independent errors, normal errors, and errors with constant variance, then we should improve the mean structure by including intensity, speed, their interaction, or some combination of these.

Escaping hydrocarbons

3. The file `hydrocarbons.csv` is a `csv` version of the data set found at `http://people.sc.fsu.edu/~jburkardt/datasets/regression/x14.txt`. Use the following code to fit a main effects model:

```
d = read.csv("hydrocarbons.csv")
m = lm(Hydrocarbons.escaping..grams. ~ ., data = d)
```

Answer the following questions based on this model fit.

(a) For each of the following regression assumptions, state whether the assumption is reasonably met and state what your evidence for this conclusion. As an example, for the linearity (mean structure) assumption, you might say

> The linearity assumption is reasonably met based on no curvature observed in scatterplots of residuals vs each individual explanatory variable.

(5 points each)

i. Normality
Answer: The linearity assumption appears reasonably met according to the normal qq-plot. There is some slight evidence of heavy-tailed errors, but nothing too concerning.

ii. Constant variance
Answer: The constant variance assumption appears reasonably met according to the plot of residuals vs fitted values (no shotgun pattern) and the absolute residuals vs fitted values or scale-location plot (smoothing line is reasonably straight).

iii. Independence
Answer: Independence is hard to evaluate. Plotting the residuals vs index which may give an indication of a temporal pattern in the errors may show reduced variability for residuals with a low index compared to residuals with a higher index. This would actually be a violation of the constant variance assumption. No clear pattern is observed in the residuals and therefore there is no clear evidence that the independence assumption is violated.

(b) Explain why the observation in line 18 has high leverage. (5 points)
Answer: An observation has high leverage when the explanatory variables are far from the explanatory variables for the other observations. Observation 18 has tank temperature of 90F which is almost the maximum in the data set, initial tank pressure of 7.32psi which is close to the maximum, petrol pressure of 7.2psi which is also close to the maximum, and petrol temperature of 60F which is close to the 3rd quartile. These values are all on the high end compared to the other observations and thus give observation 18 high leverage.

Escaping hydrocarbons (continued)

(c) Provide estimates of the following quantities based on the model fit. (2 points each)

i. Intercept
Answer:

```
coef(m)[1]
## (Intercept)
##    1.015018
```

ii. Coefficient for 'Tank temperature (F)'
Answer:

```
coef(m)[2]
## Tank.temperature..F.
##          -0.02860886
```

iii. Coefficient for 'Petrol temperature (F)'
Answer:

```
coef(m)[3]
## Petrol.temperature..F.
##              0.2158169
```

iv. Coefficient for 'Initial tank pressure (pounds/square inch)'
Answer:

```
coef(m)[4]
## Initial.tank.pressure..pounds.square.inch.
##                                  -4.320052
```

v. Coefficient for 'Petrol pressure (pounds/square inch)'
Answer:

```
coef(m)[5]
## Petrol.pressure..pounds.square.inch.
##                               8.974889
```

vi. Error variance
Answer:

```
summary(m)$sigma^2
## [1] 7.452874
```

vii. Coefficient of determination
Answer:

```
summary(m)$r.squared
## [1] 0.92606
```

(d) Conduct a t-test to determine whether an interaction between 'Tank temperature (F)' and 'Initial tank pressure (pounds/square inch)' is supported by the data. Report the following quantities for this test. (1 point each)

Answer:

```
mI = lm(Hydrocarbons.escaping..grams. ~ . +
          Tank.temperature..F.:Initial.tank.pressure..pounds.square.inch.,
        data = d)
```

i. Estimated coefficient
Answer:

```
coef(mI)[6]
## Tank.temperature..F.:Initial.tank.pressure..pounds.square.inch.
##                                                      -0.03523913
```

ii. t-statistic
Answer:

```
summary(mI)$coefficients[6,3]
## [1] -1.506864
```

iii. degrees of freedom
Answer:

```
mI$df.residual
## [1] 26
```

iv. p-value
Answer:

```
summary(mI)$coefficients[6,4]
## [1] 0.1439004
```

(e) Provide an interpretation for the result of this test. (2 points)
Answer: Since the p-value is not small, there is no evidence the data are incomppatible with the regression model that does not include the interaction term.

Story County Commercial Property Sales

4. The page entitled `R Code - Story County Commercial Property Sales` provides an analysis of properties sold from 2005 through February 2013 based on `price` ($), `year` built, building `area` (ft^2), and `land` area (ft^2). (5 points each)

   (a) Provide an interpretation (including units) for the following quantities:

      i. $e^{13.278083} = 584249$
      Answer: The estimated median price for a property built in 1960, with 10,000 ft^2 of building area, and 100,000 ft^2 of land area, is $584,249.

      ii. $e^{0.007551} = 1.008$
      Answer: The multiplicative effect on the median price for each additional year is 1.008 when building area is 10,000 ft ^{2} and land area is held constant.

      iii. $10^{0.522834} = 3.33$
      Answer: The multiplicative effect on the median price for each 10-fold increase in building area is 3.33 when year is 1960 and land area is held constant.

   (b) Construct a two-sided equal-tail 90% confidence interval for the interaction coefficient. (5 points)
   Answer:

```
0.00468 + c(-1,1)*qt(.95, df=224)*0.001612
## [1] 0.002017484 0.007342516
```

R Code - Story County Commercial Property Sales

Feel free to remove this page

```
summary(m)
##
## Call:
## lm(formula = log(price) ~ I(year - 1960) + log(area/10000) +
##     log(land/1e+05) + I(year - 1960):log(area/10000), data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4587 -0.5373 -0.0024  0.5002  3.5129
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    13.278083   0.107976 122.973  < 2e-16 ***
## I(year - 1960)                  0.007551   0.002327   3.245  0.00135 **
## log(area/10000)                 0.522834   0.076494   6.835 7.68e-11 ***
## log(land/100000)                0.117985   0.059950   1.968  0.05029 .
## I(year - 1960):log(area/10000)  0.004680   0.001612   2.903  0.00407 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7726 on 224 degrees of freedom
## Multiple R-squared:  0.5314,Adjusted R-squared:  0.5231
## F-statistic: 63.51 on 4 and 224 DF,  p-value: < 2.2e-16
```

(intentionally left blank)