Fall 2019 Exam I

STAT 8020

September 27, 2019

Name:__________________________

Directions

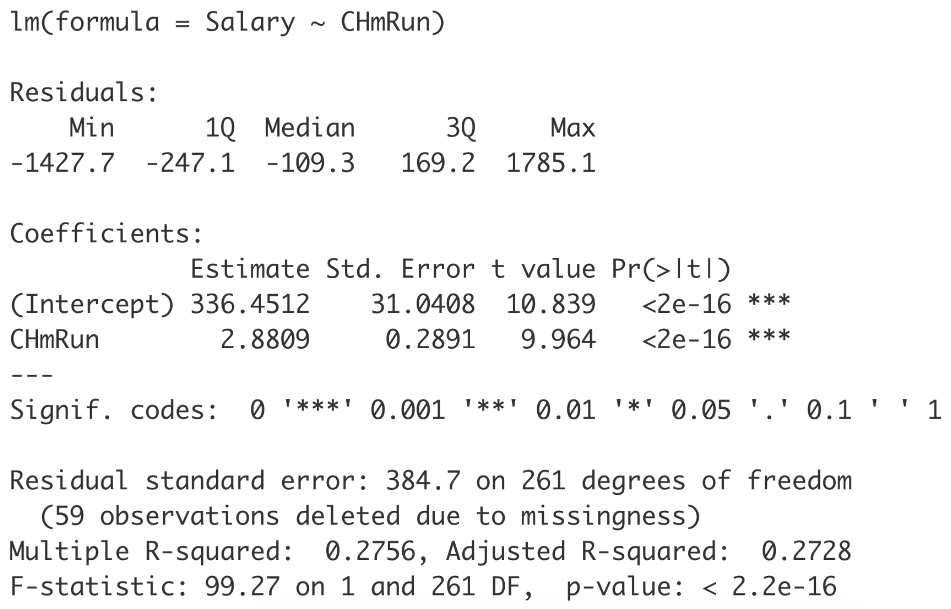
1. Show your work on ALL questions (except those multiple choice questions). Unsupported work will NOT receive full credit.
2. Decimal answers should be exact, or to exactly 2 significant digits.
3. Please write legibly. If I cannot read your writing, NO credit will be given.
4. You are allowed the following aids:
   (a) a one-page A4 handwritten cheat sheet
   (b) A scientific Calculator
5. Turn off your cell phone before the exam begins.

Use your time wisely. Good Luck!!!

Problem	Points Possible	Points Earned
1	60	
2	20	
3	20	
Total	100	

Problem 1

A baseball fan would like to study the relationship between the annual salary **Salary** (in thousands of dollars) of major league players and the number of home runs during his career **CHmRun**. A simple linear regression is performed where **Salary** is the response. Use the R output below to answer the following questions: **(12 points for each question.)**



```
lm(formula = Salary ~ CHmRun)

Residuals:
    Min      1Q  Median      3Q     Max 
-1427.7  -247.1  -109.3   169.2  1785.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 336.4512    31.0408  10.839   <2e-16 ***
CHmRun        2.8809     0.2891   9.964   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 384.7 on 261 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.2756, Adjusted R-squared:  0.2728 
F-statistic: 99.27 on 1 and 261 DF,  p-value: < 2.2e-16
```

1. Write down the least squares regression line and compute the fitted value with **CHmRun** $= 100$.

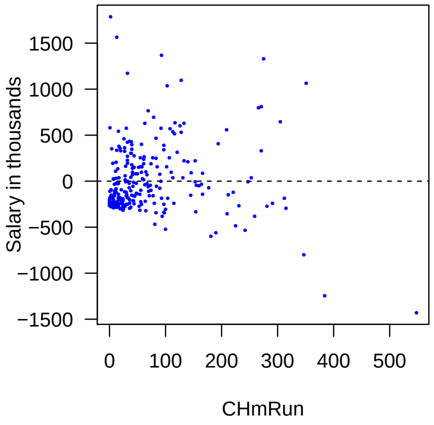
2. Construct the 95% confidence interval (using $t(0.975, df = 261) = 1.97$ and $\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 1330.484$) for β_1.

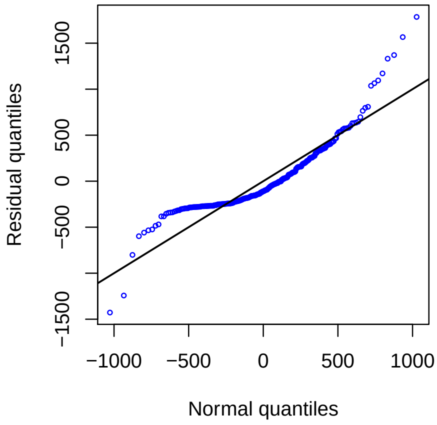
3. Test the following hypothesis: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ with $\alpha = 0.05$. State your conclusion in plain language in the present context.

4. Fill in the missing values in the ANOVA table below and compute the R^2, the coefficient of determination.

Source	df	SS	MS	F
Model	?	SSR = 14692193	MSR = ?	F* =?
Error	261	SSE = 38626920	MSE = ?	
Total	?	SST =?		

5. Do the residual plot and the Normal Q-Q plot below suggest any regression assumptions may be violated? Explain your answer.

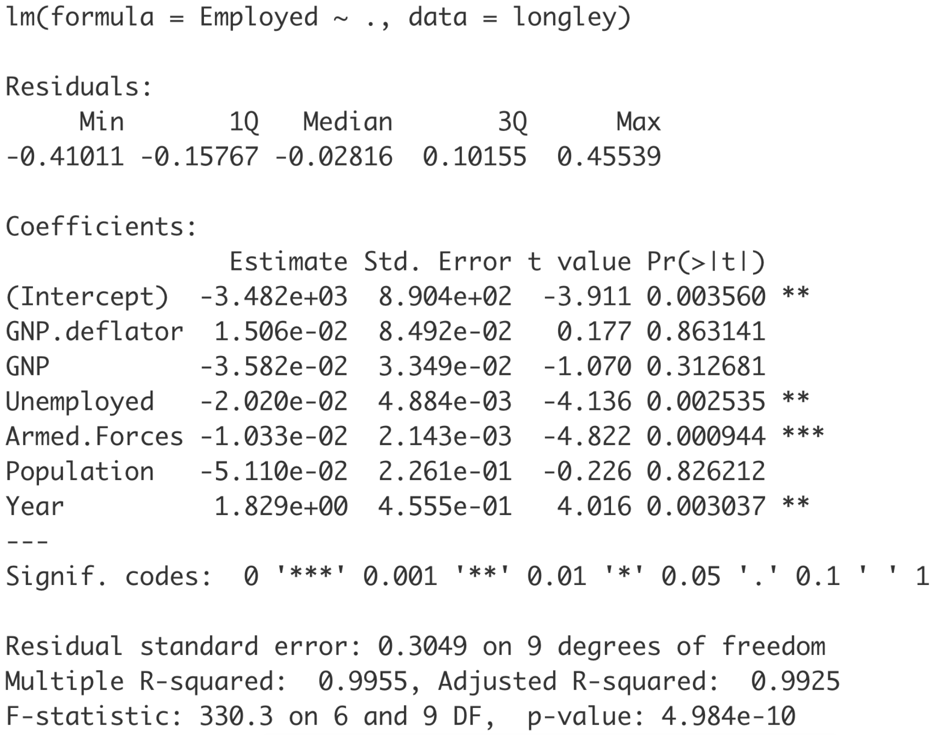




Problem 2

A researcher performs a multiple linear regression, using the Longley's macroeconomic data set, to study the relationship between `Employed` (number of people employed) and `GNP.deflator`, `GNP`, `Unemployed`, `Armed.Forces`, `Population`, and `Year`. Use the R outputs below to answer the following questions:

Full model Fit:



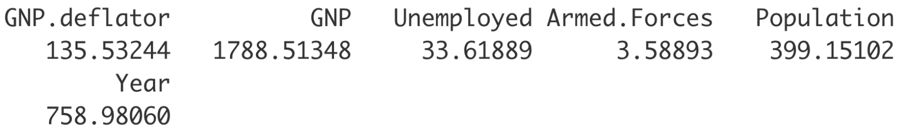
```
lm(formula = Employed ~ ., data = longley)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.41011 -0.15767 -0.02816  0.10155  0.45539 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 ** 
GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141    
GNP          -3.582e-02  3.349e-02  -1.070 0.312681    
Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 ** 
Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
Population   -5.110e-02  2.261e-01  -0.226 0.826212    
Year          1.829e+00  4.555e-01   4.016 0.003037 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-squared:  0.9955, Adjusted R-squared:  0.9925 
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

VIF:



```
GNP.deflator          GNP   Unemployed Armed.Forces   Population 
   135.53244   1788.51348     33.61889      3.58893    399.15102 
        Year 
   758.98060 
```

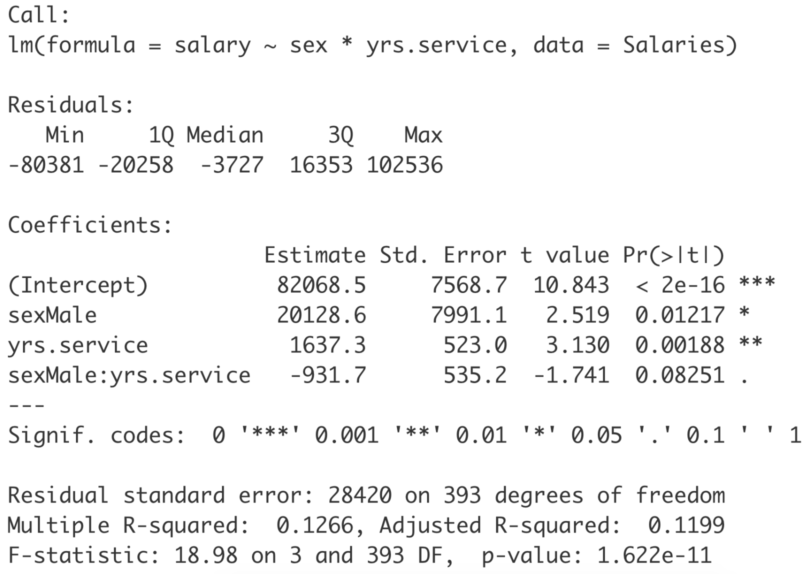
1. **(10 points)** Explain why the full model is highly significant (overall F-test P-value $< 5 \times 10^{-10}$ and with a very high R^2) but still have very high p-values on some of the regressor's t tests? (Hint: Check the VIF values.)

2. **(10 points)** Perform a general linear test using the R output below:

```
## Analysis of Variance Table
##
## Model 1: Employed ~ GNP + Unemployed + Armed.Forces + Year
## Model 2: Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces + Populat
ion +
##     Year
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     11 0.85868
## 2      9 0.83642  2  0.022256 0.1197 0.8885
```

Problem 3

The dean of a college in a University would like to monitor salary differences between male and female faculty members and she performed a multiple linear regression where the response variable `salary` is regressed on `sex` (male, female) and `yrs.service` (years of service). Use the R output below to answer the following question:



```
Call:
lm(formula = salary ~ sex * yrs.service, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max 
-80381 -20258  -3727  16353 102536 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          82068.5     7568.7  10.843  < 2e-16 ***
sexMale              20128.6     7991.1   2.519  0.01217 *  
yrs.service           1637.3      523.0   3.130  0.00188 ** 
sexMale:yrs.service   -931.7      535.2  -1.741  0.08251 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28420 on 393 degrees of freedom
Multiple R-squared:  0.1266, Adjusted R-squared:  0.1199 
F-statistic: 18.98 on 3 and 393 DF,  p-value: 1.622e-11
```

1. **(20 points)** Write down the regression equation for male and female faculty, respectively.