Exam I

MTH 4230 Regression and Computational Statistics

1 For this problem refer to the **R commands** and **output** below. The data analyzed are the **cylinder volumes** (cubic inches) and **gas mileages** (mpg) for a sample of cars.

```
my.reg <- lm(Mpg ~ CylinderVol)
```

```
summary(my.reg)

##
## Call:
## lm(formula = Mpg ~ CylinderVol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8883 -0.7861  0.1954  0.6325  2.1070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.00918    2.23689  15.204 1.28e-06
## CylinderVol -0.08372    0.01838   ?????  0.00262
##
## Residual standard error: 1.259 on 7 degrees of freedom
## Multiple R-squared:  0.7478,        Adjusted R-squared:  0.7117
## F-statistic: 20.75 on 1 and 7 DF,  p-value: 0.002619
```

```
summary(my.reg)
```

a) Write the (theoretical) **regression model** including the **assumptions** about the random component of the model.

b) How many cars are included in the sample (i.e. what's the **sample size**)?

c) Notice that the value of the F statistic is 20.75 and the corresponding p-value is 0.002619. What **null** and **alternative hypotheses** does this test? Write $\boldsymbol{H_0}$ and $\boldsymbol{H_a}$ below.

d) Notice that the t statistic value for the test of β_1 is missing. Find the value of the missing $\boldsymbol{t}$ **statistic**.

e) The 97.5th percentile (t critical value) $t_{0.025,7}$ of the $t(7)$ distribution is

```
qt(0.975, df = 7)

## [1] 2.364624
```

Use this and the R output from `summary()` above to a compute a **95% confidence interval** for the true slope β_1.

2 For this question refer to the **data**, **R commands**, **output**, and **plot** below. The data analyzed are **peak power loads** for a power plant and **daily high temperatures** (Fahrenheit) for each of 10 days.

```
power.plant.data

##    Day HighTemp PeakLoad
## 1    1       95      214
## 2    2       82      152
## 3    3       90      156
## 4    4       81      129
## 5    5       99      254
## 6    6      100      266
```

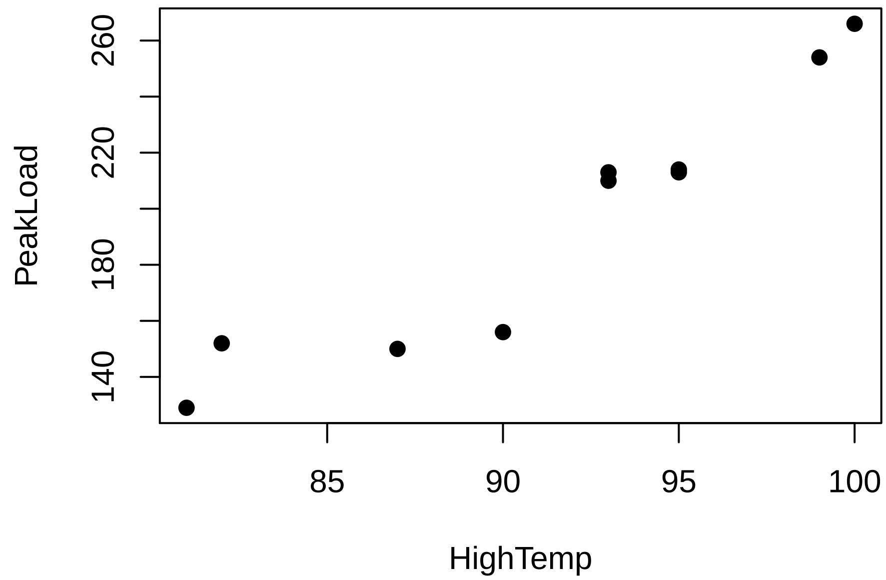
```
## 7    7       93      210
## 8    8       95      213
## 9    9       93      213
## 10  10       87      150
```

```
my.reg <- lm(PeakLoad ~ HighTemp)
```

```
summary(my.reg)

##
## Call:
## lm(formula = PeakLoad ~ HighTemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.500  -6.251   4.401   7.248  20.902
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -426.5240    72.5759  -5.877 0.000371
## HighTemp       6.8003     0.7914   8.593  2.6e-05
##
## Residual standard error: 15.44 on 8 degrees of freedom
## Multiple R-squared:  0.9022,Adjusted R-squared:   0.89
## F-statistic: 73.84 on 1 and 8 DF,  p-value: 2.601e-05
```

```
plot(HighTemp, PeakLoad, pch=19)
```



a) Give the equation of the **estimated regression line**.

b) By how much does peak load tend to increase for each one-degree (Fahrenheit) increase in temperature?

c) Is the observed increase in peak load **statistically significant**? Give the value of a **test statistic** with **degrees of freedom** and a **p-value** to support your answer.

d) One of the intervals below is a **confidence interval for the mean** peak load when the temperature is 85 degrees, and the other is a **prediction interval** for a new individual peak load when temperature is 85. Which is the **confidence interval** and which is the **prediction interval**? How do you know?

$$(135.1,\ \ 167.9) \qquad (112.3,\ \ 190.7)$$

e) Suppose that the temperatures are converted from Fahrenheit to Celsius, and the linear regression model is refit to the data. What would be the **equation** of the **new fitted regression line**? ***Hint***: If F is a temperature in Fahrenheit, then that temperature in Celsius, C, is $C = \frac{5}{9} \cdot (F - 32)$.

3 A high school physics class performed an experiment see how pressure applied along a steel wire affects its length. Six **force** levels (kg) were applied, and the **change in length** of the wire (mm) was recorded for each force level. The data are below.

```
physics.data

##   Force ChangeInLength
## 1  29.4            4.3
## 2  39.2            5.3
## 3  49.0            6.5
## 4  58.8            7.9
## 5  68.6            8.8
## 6  74.4           10.0
```

A linear regression model was fit to the data. Refer to the **R commands** and **output** below.

```
my.reg <- lm(ChangeInLength ~ Force)
```

```
summary(my.reg)

##
## Call:
## lm(formula = ChangeInLength ~ Force)
```

```
##
## Residuals:
##        1        2        3        4        5        6
##  0.13373 -0.08629 -0.10632  0.07366 -0.24636  0.23158
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.506203   0.283678   1.784    0.149
## Force       0.124492   0.005109  24.369 1.68e-05
##
## Residual standard error: 0.1977 on 4 degrees of freedom
## Multiple R-squared:  0.9933,Adjusted R-squared:  0.9916
## F-statistic: 593.8 on 1 and 4 DF,  p-value: 1.682e-05
```

a) We want to summarize how well the linear regression model fits the data. Give the numerical value of the **most appropriate statistic** for such a summary.

b) **Predict** the wire's **change in length** when the **force** applied is **6** kg.

c) Find the value of the **mean square for regression**, **MSR**.

d) Write out the **design matrix** $\boldsymbol{X}$ below.

e) Notice that the intercept β_0 was included in the model. Give an argument that it would have been more appropriate to fit a model through the origin.

4 A simple linear regression model was fit to a data set with $n = 3$ observations. The resulting fitted (predicted) values $\hat{Y}_1, \hat{Y}_2, \hat{Y}_3$ and observed responses Y_1, Y_2, Y_3 are shown below.

Observed response Y_i	Fitted value $\hat{Y}_i$
4	2.9
2	3.8
6	5.3

a) Calculate MSE and MSR.

b) Calculate the F statistic for the test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.