ASSIGNMENT 1: Multiple Linear Regression

First order Model with Interaction Term (Quantitative and Qualitative Variable)

Problem 1. The amount of water used by the production facilities of a plant varies. Observations on water usage and other, possibility related, variables were collected for 250 months. The data are given in **water.csv file**. The explanatory variables are

TEMP= average monthly temperature (degree celsius)

PROD=amount of production (in hundreds of cubic)

DAYS=number of operationing day in the month (days)

HOUR=number of hours shut down for maintenance (hours)

The response variable is USAGE=monthly water usage (gallons/minute)

a. Fit the model containing all four independent variables. What is the estimated multiple regression equation?

```
#Question a,b
fullmodel=lm(USAGE~.,data=waterdata)
summary(fullmodel)
```

```
##
## Call:
## lm(formula = USAGE ~ ., data = waterdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4030 -1.1433  0.0473  1.1677  5.3999
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.891627   1.028794   5.727  3.0e-08 ***
## PROD         0.040207   0.001629  24.681  < 2e-16 ***
## TEMP         0.168673   0.008209  20.546  < 2e-16 ***
## HOUR        -0.070990   0.016992  -4.178  4.1e-05 ***
## DAYS        -0.021623   0.032183  -0.672    0.502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.768 on 244 degrees of freedom
## Multiple R-squared:  0.8885, Adjusted R-squared:  0.8867
## F-statistic:   486 on 4 and 244 DF,  p-value: < 2.2e-16
```

<!–*We fit the model using the command lm in R. See R code above. The estimated multiple regression equation is* $\hat{usage} = 5.891627 + 0.040207PROD + 0.168673TEMP - 0.070990HOUR - 0.021623DAYS$.

–>

b. Test the hypothesis for the full model i.e the test of overall significance. Use significance level 0.05.

<!– $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ *versus* H_a : *at least one* β_i *is not zero. Fcal= 486 with p-value* $< 2.2e\text{-}16 < 0.05$ *so we reject* H_o *at* $\alpha = 0.05$. *Therefore, at least one of the predictors must be related to the response water usage.*

–>

c. Would you suggest the model in part b for predictive purposes? Which model or set of models would you suggest for predictive purposes? Hint: Use Individual Coefficients Test (t-test) to find the best model.)

```
#Question c
reducedmodel=lm(USAGE~PROD+TEMP+HOUR,data=waterdata)
summary(reducedmodel)
```

```
##
## Call:
## lm(formula = USAGE ~ PROD + TEMP + HOUR, data = waterdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5066 -1.1356  0.0469  1.1519  5.3750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.307511   0.549483   9.659  < 2e-16 ***
## PROD         0.040115   0.001621  24.741  < 2e-16 ***
## TEMP         0.169188   0.008164  20.723  < 2e-16 ***
## HOUR        -0.070769   0.016970  -4.170 4.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.766 on 245 degrees of freedom
## Multiple R-squared:  0.8883, Adjusted R-squared:  0.8869
## F-statistic: 649.3 on 3 and 245 DF,  p-value: < 2.2e-16
```

<!– *The output from summary(fullmodel) (see R code above) provides a table with individual t-tests. It shows that the p-value for DAYS is 0.502>0.05, so we should clearly fail to reject the null hypothesis. Therefore, the predictor DAYS should be dropped out of the model. After dropping DAYS, the new estimated first order model becomes:* $us\hat{a}ge = 5.307511 + 0.040115PROD + 0.169188TEMP - 0.070769HOUR$.

–>

d. Use Partial F test to confirm that the independent variable (removed from part c) should be out of the model at significance level 0.05.

```
#Question a,b
fullmodel=lm(USAGE~.,data=waterdata)
reducedmodel=lm(USAGE~PROD+TEMP+HOUR,data=waterdata)
#Question d
anova(reducedmodel,fullmodel)
```

```
## Analysis of Variance Table
##
## Model 1: USAGE ~ PROD + TEMP + HOUR
## Model 2: USAGE ~ PROD + TEMP + HOUR + DAYS
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    245 764.47
## 2    244 763.06  1    1.4117 0.4514 0.5023
```

<!–

$$
\begin{aligned}
H_0 &: \beta_4 = 0 \text{ in the model} \quad Y = \beta_0 + \beta_1 PROD + \beta_2 TEMP + \beta_3 HOUR + \beta_4 DAYS + \epsilon \\
H_a &: \beta_4 \neq 0 \text{ in the model} \quad Y = \beta_0 + \beta_1 PROD + \beta_2 TEMP + \beta_3 HOUR + \beta_4 DAYS + \epsilon
\end{aligned}
$$

The Partial F test using the command R command anova(reducedmodel,fullmodel) gives $Fcal = 0.4514$ *with p-value = 0.5023> 0.05, confirming that the predictor DAYS clearly should be dropped out of the model.*

–>

e. Obtain a 95% confidence interval of regression coefficient for TEMP from the model in part c. Give an interpretation.

```
reducedmodel=lm(USAGE~PROD+TEMP+HOUR,data=waterdata)
#Question e
confint(reducedmodel)
```

```
##                  2.5 %      97.5 %
## (Intercept)  4.22519744  6.38982411
## PROD         0.03692098  0.04330837
## TEMP         0.15310634  0.18526907
## HOUR        -0.10419445 -0.03734272
```

<!– *We computed 95% confidence intervals using the command* confint(reducedmodel), *and we obtained this interval* $(0.15310634, 0.18526907)$ *from the R output, which means that the monthly water usage increases 0.153106434 (gallons/minute) to 0.18526907 (gallons/minute) for every 1 degree Celsius increase in average temperature.*

–>

f. Use the method of Model Fit to calculate R^2_{adj} and RMSE to compare the full model and the model in part c. Which model or set of models would you suggest for predictive purpose? For the final model, give an interpretation of R^2_{adj} and RMSE.

<!– *The model* $\hat{usage} = 5.307511 + 0.040115PROD + 0.169188TEMP - 0.070769HOUR$ *has* $R^2_{adj} = 0.8869$ *and* $RMSE = 1.766$. *Its* R^2_{adj} *is very high and RMSE is lower than the full model, so I would suggest this model for predicting* Y. *Interpretation: As* $R^2_{adj} = 0.8869$, *hence 88.69% of the variation of the water usage is explained by the model. An* $RMSE = 1.766$ *means that the standard deviation of the unexplained variance by the model is 1.766.*

–>

g. (From Exercise 2) Build an interaction model to fit the multiple regression model from the model in part f. From the output, which model would you recommend for predictive purposes?

```
#Question g
interacmodel1<-lm(USAGE~(PROD+TEMP+HOUR)^2,data=waterdata)
summary(interacmodel1)
```

```
##
## Call:
## lm(formula = USAGE ~ (PROD + TEMP + HOUR)^2, data = waterdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1941 -0.3165 -0.0502  0.2755  7.0985
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.294e+01  7.113e-01  18.193   <2e-16 ***
```

```
## PROD         -3.642e-03  2.565e-03  -1.420    0.157
## TEMP         -2.389e-02  2.129e-02  -1.122    0.263
## HOUR         -2.340e-01  2.512e-02  -9.316   <2e-16 ***
## PROD:TEMP     1.189e-03  6.932e-05  17.154   <2e-16 ***
## PROD:HOUR     7.767e-04  7.820e-05   9.933   <2e-16 ***
## TEMP:HOUR     7.600e-04  7.683e-04   0.989    0.324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9867 on 242 degrees of freedom
## Multiple R-squared:  0.9656, Adjusted R-squared:  0.9647
## F-statistic:  1131 on 6 and 242 DF,  p-value: < 2.2e-16
```

```r
interacmodel2<-lm(USAGE~PROD+TEMP+HOUR+PROD*TEMP+PROD*HOUR,data=waterdata)
summary(interacmodel2)
```

```
##
## Call:
## lm(formula = USAGE ~ PROD + TEMP + HOUR + PROD * TEMP + PROD *
##     HOUR, data = waterdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1423 -0.3148 -0.0358  0.3029  7.2555
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.243e+01  4.839e-01  25.679   <2e-16 ***
## PROD        -2.529e-03  2.305e-03  -1.097    0.274
## TEMP        -4.737e-03  8.859e-03  -0.535    0.593
## HOUR        -2.151e-01  1.624e-02 -13.242   <2e-16 ***
## PROD:TEMP    1.142e-03  5.009e-05  22.795   <2e-16 ***
## PROD:HOUR    7.873e-04  7.745e-05  10.165   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9866 on 243 degrees of freedom
## Multiple R-squared:  0.9654, Adjusted R-squared:  0.9647
## F-statistic:  1357 on 5 and 243 DF,  p-value: < 2.2e-16
```

```r
interacmodel3<-lm(USAGE~PROD+TEMP+HOUR+TEMP*HOUR+PROD*HOUR,data=waterdata)
summary(interacmodel3)
```

```
##
## Call:
## lm(formula = USAGE ~ PROD + TEMP + HOUR + TEMP * HOUR + PROD *
##     HOUR, data = waterdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9348 -0.6362  0.1191  0.7713  8.0030
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5980162  0.6796266   5.294 2.67e-07 ***
```

```
## PROD         0.0292074  0.0025348  11.523  < 2e-16 ***
## TEMP         0.2997804  0.0146449  20.470  < 2e-16 ***
## HOUR         0.0388793  0.0288840   1.346     0.18
## TEMP:HOUR   -0.0083505  0.0008247 -10.126  < 2e-16 ***
## PROD:HOUR    0.0006707  0.0001158   5.792 2.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 243 degrees of freedom
## Multiple R-squared:  0.9237, Adjusted R-squared:  0.9221
## F-statistic: 588.4 on 5 and 243 DF,  p-value: < 2.2e-16
```

```r
interacmodel4<-lm(USAGE~PROD+TEMP+HOUR+PROD*TEMP+TEMP*HOUR,data=waterdata)
summary(interacmodel4)
```

```
##
## Call:
## lm(formula = USAGE ~ PROD + TEMP + HOUR + PROD * TEMP + TEMP *
##     HOUR, data = waterdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1134 -0.4948  0.0303  0.4258  6.5290
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.043e+01  7.872e-01  13.249  < 2e-16 ***
## PROD         1.214e-02  2.384e-03   5.090 7.18e-07 ***
## TEMP        -3.042e-02  2.519e-02  -1.207    0.228
## HOUR        -1.305e-01  2.706e-02  -4.822 2.51e-06 ***
## PROD:TEMP    1.135e-03  8.182e-05  13.868  < 2e-16 ***
## TEMP:HOUR    1.808e-03  9.010e-04   2.006    0.046 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.168 on 243 degrees of freedom
## Multiple R-squared:  0.9515, Adjusted R-squared:  0.9505
## F-statistic: 954.2 on 5 and 243 DF,  p-value: < 2.2e-16
```

<!– *We fitted the following models in order to include potential interactions: i) we first include all possible interactions with varaibles involved in the model obtained in part f. Then ii) we remove individually interactions that are not significant. The R code to implement this procedure is shown above. After fitting the model with all interactions (interacmodel1), interaction TEMP x HOUR was not significant and we removed it from model interacmodel2. Model interacmodel2 has* $R^2_{adj} = 0.9647$ *(higher than* $R^2_{adj} = 0.8869$ *of reducedmodel) and* $RMSE = 0.9866$ *lower than the RMSE of reducedmodel. Hence we will recommend model interacmodel2 with interactions PROD x TEMP and PROD x HOUR. Moreover, we also fitted 2 models (interacmodel3 and interacmodel4) without interactions PROD x TEMP and PROD x HOUR respectively. We obtained respectively* $R^2_{adj} = 0.9221$ *and* $R^2_{adj} = 0.9505$ *which are lower than our final model.*

–>

Problem 2. A collector of antique grandfather clocks sold at auction believes that the price received for the clocks depends on both the age of the clocks and the number of bidders at the auction. Thus, (s)he hypothesizes the first-order model

$$
\begin{aligned}
y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\
\textit{where} & \\
y &= \text{Auction price (dollars)} \\
X_1 &= \text{Age of clock (years)} \\
X_2 &= \text{Number of bidders}
\end{aligned}
$$

A sample of 32 auction prices of grandfather clocks, along with their age and the number of bidders, is given in data file **GFCLOCKS.CSV**

a. Use the method of least squares to estimate the unknown parameters β_0, β_1, β_2 of the model.

```r
#Question a,c,f
fullmodel<-lm(PRICE~AGE+NUMBIDS,data=clock)
summary(fullmodel)
```

```
##
## Call:
## lm(formula = PRICE ~ AGE + NUMBIDS, data = clock)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -206.49 -117.34   16.66  102.55  213.50
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1338.9513   173.8095  -7.704 1.71e-08 ***
## AGE            12.7406     0.9047  14.082 1.69e-14 ***
## NUMBIDS        85.9530     8.7285   9.847 9.34e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 133.5 on 29 degrees of freedom
## Multiple R-squared:  0.8923, Adjusted R-squared:  0.8849
## F-statistic: 120.2 on 2 and 29 DF,  p-value: 9.216e-15
```

<!– *From the R command above summary(fullmodel), we have* $\hat{\beta}_0 = -1338.9513$, $\hat{\beta}_1 = 12.7406$ *and* $\hat{\beta}_2 = 85.9530$.

–>

b. Find the value of SSE that is minimized by the least squares method.

```r
fullmodel<-lm(PRICE~AGE+NUMBIDS,data=clock)
#Question b,e
anova(lm(PRICE~1,data=clock),fullmodel)
```

```
## Analysis of Variance Table
##
## Model 1: PRICE ~ 1
## Model 2: PRICE ~ AGE + NUMBIDS
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     31 4799790
## 2     29  516727  2   4283063 120.19 9.216e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<!– *From the command line anova(fullmodel,lm(PRICE~1,data=clock)), SSE = 516727.* –>

c. Estimate s, the standard deviation of the model, and interpret the result.

<!– *From the summary function summary(fullmodel), the output reports RMSE = 133.5. This value shows the standard deviation of the unexplained variance in Auction price. It shows how far off using the model is from actual value* Y.

–>

d. Find and interpret the adjusted coefficient of determination, R^2_{Adj}.

<!– *The adjusted* R^2 *is* $R^2_{adj} = 0.8849$ *i.e. the variation in auction price that can be explained by using this model is 88.49 %. The rest (11.51%) can be explained by other predictors.*

–>

e. Construct the Anova table for the model and test the global F-test of the model at the $\alpha = 0.05$ level of significance.

<!– *From the output obtained from the command line* anova(lm(PRICE~1,data=clock),fullmodel)*, we can obtain this ANOVA table.*

The ANOVA Table

Source of Variation	Df	Sum of squares	Mean squares	F-statistics
Regression	2	4283063	2141532	120.19
Residual	29	516727	17818.17	
Total	31	4799790		

We test $H_o : \beta_1 = \beta_2 = 0$ *versus H_a: at least one* $\beta_i \neq 0$. *Fcal= 120.2 on 2 and 29 DF, with p-value: 9.216e-15 <0.05 so we reject Ho at* $\alpha = 0.05$. *Therefore, at least one of the predictors must be related to the Auction price.*

–>

f. Test the hypothesis that the mean auction price of a clock increases as the number of bidders increases when age is held constant (i.e., when $\beta_2 \neq 0$). (Use $\alpha = 0.05$)

<!– *The test is* $H_0 : \beta_2 = 0$ *versus* $H_a : \beta_2 \neq 0$. $t_{cal} = 49.847$ *with p-value=9.34e-11<0.05, confirming that the predictor NUMBIDS should clearly be added into the model at* $\alpha = 0.05$.

–>

g. Find a 95% confidence interval for β_1 and interpret the result.

```
fullmodel<-lm(PRICE~AGE+NUMBIDS,data=clock)
#Question g
confint(fullmodel)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1694.43162 -983.47106
## AGE            10.89017   14.59098
## NUMBIDS        68.10115  103.80482
```

<!– *From the output obtained from the command confint(fullmodel), a 95% confidence Interval for AGE is (10.89017, 14.59098) which means that the auction price increases 10.89017 dollars to 14.59098 dollars for every 1 year.*

–>

h. Test the interaction term between the 2 variables at $\alpha = .05$. What model would you suggest to use for predicting y? Explain.

```
#Question h
interacmodel<-lm(PRICE~AGE+NUMBIDS+AGE*NUMBIDS,data=clock)
summary(interacmodel)
```

```
##
## Call:
## lm(formula = PRICE ~ AGE + NUMBIDS + AGE * NUMBIDS, data = clock)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -154.995  -70.431    2.069   47.880  202.259
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 320.4580   295.1413   1.086  0.28684
## AGE           0.8781     2.0322   0.432  0.66896
## NUMBIDS     -93.2648    29.8916  -3.120  0.00416 **
## AGE:NUMBIDS   1.2978     0.2123   6.112 1.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.91 on 28 degrees of freedom
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.9489
## F-statistic:   193 on 3 and 28 DF,  p-value: < 2.2e-16
```

<!– *From the output (from the command line summary(interacmodel)), comparing the first order model with the interaction models, it can be clearly seen that the interaction model* $\widehat{PRICE} = \hat{\beta}_0 + \hat{\beta}_1 AGE + \hat{\beta}_2 NUMBIDS + \hat{\beta}_3 AGE \times NUMBIDS$ *performs the best result* ($R^2_{Adj} = 0.9489$, $RMSE = 88.91$). *Therefore, for predicting the auction price, I would suggest to use this interaction model instead of the additive model.*

–>

Problem 3. Cooling method for gas turbines. Refer to the Journal of Engineering for Gas Turbines and Power (January 2005) study of a high pressure inlet fogging method for a gas turbine engine. The heat rate (kilojoules per kilowatt per hour) was measured for each in a sample of 67 gas turbines augmented with high pressure inlet fogging. In addition, several other variables were measured, including cycle speed (revolutions per minute), inlet temperature (degree celsius), exhaust gas temperature (degree Celsius), cycle pressure ratio, and air mass flow rate (kilograms persecond). The data are saved in the **TURBINE.CSV** file.

(a) Write a first-order model for heat rate (y) as a function of speed, inlet temperature, exhaust temperature, cycle pressure ratio, and air flow rate.

```
head(turbine)
```

```
##        ENGINE SHAFTS   RPM CPRATIO INLET.TEMP EXH.TEMP AIRFLOW POWER HEATRATE
## 1 Traditional      1 27245     9.2       1134      602       7  1630    14622
## 2 Traditional      1 14000    12.2        950      446      15  2726    13196
## 3 Traditional      1 17384    14.8       1149      537      20  5247    11948
## 4 Traditional      1 11085    11.8       1024      478      27  6726    11289
## 5 Traditional      1 14045    13.2       1149      553      29  7726    11964
## 6 Traditional      1  6211    15.7       1172      517     176 52600    10526
```

```
#Question a,b,c,d
fullmodel<-lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+CPRATIO+AIRFLOW,data=turbine)
```

RPM	CPRATIO	INLET-TEMP	EXH-TEMP	AIRFLOW	HEATRATE
27245	9.2	1134	602	7	14622
14000	12.2	950	446	15	13196
17384	14.8	1149	537	20	11948
11085	11.8	1024	478	27	11289
14045	13.2	1149	553	29	11964
.					
.					
18910	14.0	1066	532	8	12766
3600	35.0	1288	448	152	8714
3600	20.0	1160	456	84	9469
16000	10.6	1232	560	14	11948
14600	13.4	1077	536	20	12414

Source: Bhargava, R., and Meher-Homji, C. B. "Parametric analysis of existing gas turbines with inlet evaporative and overspray fogging," *Journal of Engineering for Gas Turbines and Power*, Vol. 127, No. 1, Jan. 2005.

Figure 1: The first and last five observations are listed in the table.

```r
summary(fullmodel)
```

```
##
## Call:
## lm(formula = HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP + CPRATIO +
##     AIRFLOW, data = turbine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1007.0  -290.9  -105.8   240.8  1414.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.361e+04  8.700e+02  15.649  < 2e-16 ***
## RPM          8.879e-02  1.391e-02   6.382 2.64e-08 ***
## INLET.TEMP  -9.201e+00  1.499e+00  -6.137 6.86e-08 ***
## EXH.TEMP     1.439e+01  3.461e+00   4.159 0.000102 ***
## CPRATIO      3.519e-01  2.956e+01   0.012 0.990539
## AIRFLOW     -8.480e-01  4.421e-01  -1.918 0.059800 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 458.8 on 61 degrees of freedom
## Multiple R-squared:  0.9235, Adjusted R-squared:  0.9172
## F-statistic: 147.3 on 5 and 61 DF,  p-value: < 2.2e-16
```

<!– *From the command line summary(fullmodel) (see R code above) after using the least square method, we can write the first-order model for heart rate as* $HEAT\hat{R}ATE = 13610 + 0.08879RMP - 9.201INLET.TEMP + 14.39EXH.TEMP + 0.3519CPRATIO - 0.848AIRFLOW$

–>

(b) Test the overall significance of the model using $\alpha = 0.01$

<!– *We would like to test* $H_0 : \beta_{RMP} = \beta_{INLET.TEMP} = \beta_{EXH.TEMP} = \beta_{CPRATIO} = \beta_{AIRFLOW} = 0$ *versus* H_a *: at least one* $\beta_i \neq 0$*. From the output of summary(fullmodel), Fcal= 147.3 on 5 and 61 DF, p-value: < 2.2e-16 <0.01 so we reject Ho at* $\alpha = 0.05$*. Therefore, at least one of the predictors must be related to heat rate (Y).*

–>

(c) Fit the model to the data using the method of least squares. (Suggestion! check both models with and without a predictor that has p-value close to 0.05, and propose the best model.)

```r
head(turbine)
```

```
##        ENGINE SHAFTS   RPM CPRATIO INLET.TEMP EXH.TEMP AIRFLOW POWER HEATRATE
## 1 Traditional      1 27245     9.2       1134      602       7  1630    14622
## 2 Traditional      1 14000    12.2        950      446      15  2726    13196
## 3 Traditional      1 17384    14.8       1149      537      20  5247    11948
## 4 Traditional      1 11085    11.8       1024      478      27  6726    11289
## 5 Traditional      1 14045    13.2       1149      553      29  7726    11964
## 6 Traditional      1  6211    15.7       1172      517     176 52600    10526
```

```r
#Question c
reducemodel<-lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+AIRFLOW,data=turbine) # with AIRFLOW without CPRATIO
summary(reducemodel)
```

```
##
## Call:
## lm(formula = HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP + AIRFLOW,
##     data = turbine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1007.7  -290.5  -106.0   240.1  1414.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.362e+04  8.133e+02  16.744  < 2e-16 ***
## RPM          8.882e-02  1.344e-02   6.608 1.02e-08 ***
## INLET.TEMP  -9.186e+00  7.704e-01 -11.923  < 2e-16 ***
## EXH.TEMP     1.436e+01  2.260e+00   6.356 2.76e-08 ***
## AIRFLOW     -8.475e-01  4.370e-01  -1.939    0.057 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 455.1 on 62 degrees of freedom
## Multiple R-squared:  0.9235, Adjusted R-squared:  0.9186
## F-statistic: 187.1 on 4 and 62 DF,  p-value: < 2.2e-16
```

```
reducemodel1<-lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP,data=turbine) # without AIRFLOW and CPRATIO
anova(reducemodel1,reducemodel)
```

```
## Analysis of Variance Table
##
## Model 1: HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP
## Model 2: HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP + AIRFLOW
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1     63 13620986
## 2     62 12841965  1    779021 3.7611 0.05701 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(reducemodel1) ## Model with AIRFLOW has a better adjusted R2 and a better RMSE.
```

```
##
## Call:
## lm(formula = HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP, data = turbine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1025.8  -297.9  -115.3   225.8  1425.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.436e+04  7.333e+02  19.582  < 2e-16 ***
## RPM          1.051e-01  1.071e-02   9.818 2.55e-14 ***
## INLET.TEMP  -9.223e+00  7.869e-01 -11.721  < 2e-16 ***
## EXH.TEMP     1.243e+01  2.071e+00   6.000 1.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 465 on 63 degrees of freedom
## Multiple R-squared:  0.9189, Adjusted R-squared:  0.915
## F-statistic: 237.9 on 3 and 63 DF,  p-value: < 2.2e-16
```

<!– *We would like to test* $H_0 : \beta_i = 0$ *versus* $H_a : \beta_i \neq 0$. *From the output, 1. We clearly see that the CPRATIO predictor must be dropped out of the model as the tcal=0.012 with the p-value* $= 0.990539 > 0.05$.

2. For AIRFLOW, since the p-value =0.0598, close to 0.05, so we have 2 possible additive models.

Model 1: $HEAT\hat{R}ATE = \hat{\beta}_0 + \hat{\beta}_1 RMP + \hat{\beta}_2 INLET.TEMP + \hat{\beta}_3 EXH.TEMP + \hat{\beta}_4 AIRFLOW$.

Model 2: $HEAT\hat{R}ATE = \hat{\beta}_0 + \hat{\beta}_1 RMP + \hat{\beta}_2 INLET.TEMP + \hat{\beta}_3 EXH.TEMP$.

From summary(reducemodel) and summary(reducemodel1), the adjusted R^2 *of Model 1 and 2 are respectively* R^2_{adj} *(model 1)= 0.9186 and* R^2_{adj} *(model 2)= 0.915. Hence, we would prefer Model 1. In addition, model 1 has the lowest RMSE=455.1.*

–>

(d) Test all possible interaction terms for the best model in part (c) at $\alpha = .05$. What is the final model would you suggest to use for predicting y? Explain.

```r
interacmodel1<-lm(HEATRATE~(RPM+INLET.TEMP+EXH.TEMP+AIRFLOW)^2,data=turbine)
summary(interacmodel1)
```

```
##
## Call:
## lm(formula = HEATRATE ~ (RPM + INLET.TEMP + EXH.TEMP + AIRFLOW)^2,
##     data = turbine)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -779.7 -211.0  -40.7  177.2 1370.3
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.650e+04  8.891e+03   2.981 0.004247 **
## RPM                  7.037e-02  1.485e-01   0.474 0.637512
## INLET.TEMP          -2.366e+01  7.364e+00  -3.213 0.002180 **
## EXH.TEMP            -4.555e+00  1.795e+01  -0.254 0.800610
## AIRFLOW              1.021e+01  6.279e+00   1.627 0.109455
## RPM:INLET.TEMP      -1.133e-04  8.720e-05  -1.299 0.199266
## RPM:EXH.TEMP         1.656e-04  3.116e-04   0.531 0.597314
## RPM:AIRFLOW         -8.257e-04  4.653e-04  -1.775 0.081414 .
## INLET.TEMP:EXH.TEMP  2.417e-02  1.457e-02   1.659 0.102791
## INLET.TEMP:AIRFLOW   1.418e-02  3.852e-03   3.681 0.000523 ***
## EXH.TEMP:AIRFLOW    -5.049e-02  1.357e-02  -3.720 0.000463 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.6 on 56 degrees of freedom
## Multiple R-squared:  0.9481, Adjusted R-squared:  0.9388
## F-statistic: 102.3 on 10 and 56 DF,  p-value: < 2.2e-16
```

```r
#dropping some interaction terms as they are nonsignificant.
interacmodel2<-lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+AIRFLOW+INLET.TEMP*AIRFLOW+EXH.TEMP*AIRFLOW,data=tur
summary(interacmodel2)
```

```
##
```

```
## Call:
## lm(formula = HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP + AIRFLOW +
##     INLET.TEMP * AIRFLOW + EXH.TEMP * AIRFLOW, data = turbine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -787.68 -189.26  -22.34  145.15 1307.53
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.360e+04  9.930e+02  13.699  < 2e-16 ***
## RPM                 4.578e-02  1.577e-02   2.902 0.005174 **
## INLET.TEMP         -1.280e+01  1.090e+00 -11.741  < 2e-16 ***
## EXH.TEMP            2.327e+01  2.901e+00   8.024 4.46e-11 ***
## AIRFLOW             1.347e+00  3.496e+00   0.385 0.701414
## INLET.TEMP:AIRFLOW  1.613e-02  3.640e-03   4.432 4.03e-05 ***
## EXH.TEMP:AIRFLOW   -4.150e-02  1.087e-02  -3.816 0.000323 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401.4 on 60 degrees of freedom
## Multiple R-squared:  0.9424, Adjusted R-squared:  0.9367
## F-statistic: 163.7 on 6 and 60 DF,  p-value: < 2.2e-16
```

```r
## Partial F test
anova(interacmodel2,interacmodel1)
```

```
## Analysis of Variance Table
##
## Model 1: HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP + AIRFLOW + INLET.TEMP *
##     AIRFLOW + EXH.TEMP * AIRFLOW
## Model 2: HEATRATE ~ (RPM + INLET.TEMP + EXH.TEMP + AIRFLOW)^2
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     60 9664946
## 2     56 8717640  4    947306 1.5213 0.2084
```

<!– *After fitting a model with all interactions (output from the command summary(interacmodel1)), we dropped nonsignificant interaction terms. Final model is then interacmodel2 and the estimation is obtained from the command line summary(interacmodel2). After testing all interaction terms from model 1, we found that the model* $\hat{Y} = 13600 + 0.04578RMP - 12.80INLET.TEMP + 23.27EXH.TEMP + 1.347AIRFLOW + 0.01613INLET.TEMP \times AIRFLOW + 0.00415EXH.TEMP \times AIRFLOW$ *is used for predicting Y with the* R^2_{Adj}*=0.9367 (higher that the adjusted* R^2 *for model 1) and RMSE= 401.4 (lower than the RMSE for model 1). To confirm that we should drop all those interactions together, we perform a partial F-test with the command anova(interacmodel2,interacmodel1). It gives a p-value of 0.2084, which confirms that we do not have enough evidence to keep those interactions in the model.*

–>

(e) Give practical interpretations of the β_i estimates.

<!– *Note that the final model has two significant interactions. Hence, we will not interpret directly the main effects for variables involved in the interaction term.*

1. $\hat{\beta}_{RMP} = 0.04578$ *means that for a given amount of other predictors (are held constant), an increase of 1 revolution per minute of the cycle speed leads to an increase in the heat rate by 0.04578 revolutions per minute.*

2. The effect of INLET.TEMP is 12.80+0.016AIRFLOW, means that for a given amount of EXH.TEMP *and RPM (are held constant), an increase of 1 degree Celsius of the inlet temperature leads to an increase in the heat rate by -12.80+0.016AIRFLOW revolutions per minute.*

3. The effect of EXH.TEMP is 23.27-0.041AIRFLOW, means that for a given amount of INLET.TEMP and RPM (are held constant), an increase of 1 degree Celsius of the exhaust gas temperature leads to an increase in the heat rate by 23.27-0.041AIRFLOW revolutions per minute.

4. The effect of AIRFLOW is 1.34+0.016INLET.TEMP-0.041EXH.TEMP, means that for a given amount of RPM (is held constant), an increase of 1 kilogram persecond of air mass flow rate leads to an increase in the heat rate by 1.34+0.016INLET.TEMP-0.041EXH.TEMP revolutions per minute.

–>

(f) Find RMSE, s from the model in part (d)

<!– *RMSE for the model is 401.4* –>

(g) Find the adjusted-R2 value from the model in part (d) and interpret it.

<!– *The adjusted* R^2 *is* R^2_{Adj}*=0.9367 implies that 93.67% of the variation in the heart rate is explained by this model.*

–>

(h) Predict a heat rate (y) when a cycle of speed = 273,145 revolutions per minute, inlet temperature= 1240 degree celsius, exhaust temperature=920 degree celsius, cycle pressure ratio=10 kilograms persecond, and air flow rate=25 kilograms persecond.

```
#dropping some interaction terms as they are nonsignificant.
interacmodel2<-lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+AIRFLOW+
                  +INLET.TEMP*AIRFLOW+EXH.TEMP*AIRFLOW,data=turbine)
#Question h
newdata = data.frame(RPM=273145, INLET.TEMP=1240,EXH.TEMP=920,AIRFLOW=25)
predict(interacmodel2,newdata, interval="predict")
```

```
##        fit      lwr     upr
## 1 31227.97 24067.74 38388.2
```

<!–

From the R command predict (see R code above), with 95% confidence interval, the heat rate (Y) is between 24067.74 revolutions per minute to 38388.2 revolutions per minute when a cycle of speed = 273,145 revolutions per minute, inlet temperature= 1240 degree Celsius, exhaust temperature=920 degree Celsius, cycle pressure ratio=10 kilograms per second, and air flow rate=25 kilograms per second.