603 Assignment One

Maria Delgado

2024-03-07

```
#Problem 1
library("mosaic")
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order
to add
## additional features.  The original behavior of these functions should not
be affected by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
water <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment
One/water.csv")
head(water)
```

```
##    PROD TEMP HOUR USAGE DAYS
## 1 171.3 39.7  9.5  19.0   20
## 2  19.4 16.0 20.0   6.6   21
```

```
## 3  18.7 12.1 26.0   6.7   21
## 4  25.6 39.0 24.0   9.5   21
## 5  25.6 39.0 23.0   9.5   21
## 6 139.2 14.3 16.0  12.2   21

##A
water_full_m = lm(USAGE~PROD+TEMP+HOUR+DAYS, data=water)
water_full_m$coefficients

## (Intercept)        PROD        TEMP        HOUR        DAYS
##  5.89162697  0.04020739  0.16867306 -0.07099009 -0.02162304

#Estimated multiple regression equation: Water_hat = 5.89162697 + 0.04020739
* PROD + 0.16867306 * TEMP - 0.07099009 * HOUR - 0.02162304 * DAYS

##B (H0): All slope coefficients are zero, implying that the model itself
contributes nothing useful. The alternative hypothesis (Ha): At least one
slope coefficient is not zero, indicating that the model is useful in
predicting the response variable.

anova(water_full_m)

## Analysis of Variance Table
##
## Response: USAGE
##            Df Sum Sq Mean Sq   F value    Pr(>F)
## PROD        1 4210.3  4210.3 1346.3213 < 2.2e-16 ***
## TEMP        1 1813.7  1813.7  579.9440 < 2.2e-16 ***
## HOUR        1   54.3    54.3   17.3516 4.313e-05 ***
## DAYS        1    1.4     1.4    0.4514    0.5023
## Residuals 244  763.1     3.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(water_full_m)

##
## Call:
## lm(formula = USAGE ~ PROD + TEMP + HOUR + DAYS, data = water)
##
## 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
```

#The output of "summary(water_full_m)" shows that F=486.02 with 4, 244 degrees of freedom (p-value < 2.2e-16 < alpha = 0.05), which indicates that we should reject the null. The large F statistic suggests that at least one coefficient should be significant

##C

```
summary(water_full_m)
```

```
##
## Call:
## lm(formula = USAGE ~ PROD + TEMP + HOUR + DAYS, data = water)
##
## 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
```

#From the individual coefficients test (t-test), the output shows: The intercept and coefficients for the variables PROD, TEMP, and HOUR are statistically significant (p-values < 0.05), indicating that they have a significant influence on water USAGE. The coefficient for the variable DAYS has a p-value (0.502) > 0.05, suggesting that it is not statistically significant and does not have a significant influence on water USAGE.

```
water_current_model = lm(USAGE~PROD+TEMP+HOUR, data=water)
water_current_model$coefficients
```

```
## (Intercept)        PROD        TEMP        HOUR
##  5.30751078  0.04011468  0.16918771 -0.07076858
```

```r
#Current valid model is: USAGE_hat = 5.30751078 + 0.04011468 * PROD +
0.16918771 * TEMP - 0.07076858 * HOUR

##D

water_full_m = lm(USAGE~PROD+TEMP+HOUR+DAYS, data=water)
water_current_model = lm(USAGE~PROD+TEMP+HOUR, data=water)
anova(water_current_model, water_full_m)
```

```
## 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
```

```r
#Null hypothesis (H0) is that the coefficient of DAYS (beta(DAYS)) in the
full model is zero, implying that DAYS does not contribute significantly to
the model. Alternative hypothesis (Ha) is that beta(DAYS) is not equal to
zero, suggesting that DAYS does contribute significantly to the model. In
running the Partial F-test to confirm if the DAYS variable should be dropped:
P-value of 0.5023 (> 0.05), we fail to reject the null hypothesis. Therefore,
based on the partial F-test, we should drop the variable DAYS from the model,
as it does not contribute significantly to explaining the variation in the
response variable (USAGE) at the 5% significance level.

##E
confint(water_current_model)
```

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

```r
#The 95% confidence interval for the coefficient of TEMP is given as
(0.15310634, 0.18526907). This interval indicates that we are 95% confident
that the true effect of TEMP on water USAGE falls between 0.15310634 and
0.18526907 gallons/minute for every increase in TEMP by 1 degree Celsius,
holding HOUR and PROD constant. Since the confidence interval does not
include zero, it implies that the effect of TEMP on water USAGE is
statistically significant at the 5% level. We can conclude that there is a
significant positive relationship between TEMP and water USAGE.

##F

water_full_m = lm(USAGE~PROD+TEMP+HOUR+DAYS, data=water)
water_current_model = lm(USAGE~PROD+TEMP+HOUR, data=water)
summary(water_full_m)$adj.r.squared
```

```
## [1] 0.886658
```

```
summary(water_current_model)$adj.r.squared
```

```
## [1] 0.8869118
```

```
#Model with all predictors has adjusted r-squared = 0.886658
#Reduced Model has adjusted r-squared = 0.8869118
```

```
sigma(water_full_m)
```

```
## [1] 1.768414
```

```
sigma(water_current_model)
```

```
## [1] 1.766433
```

```
#Model with all predictors has RMSE = 1.768414
#Reduced Model has RMSE = 1.766433
```

```
#I would suggest the reduced model: Based on the higher adjusted R-squared
and lower RMSE values, the reduced model (using only PROD, TEMP, and HOUR
predictors) appears to be better for predictive purposes compared to the full
model. The reduced model explains approximately 88.69% of the variation in
water USAGE, and its RMSE can be interpreted as the standard deviation of the
unexplained variance.
```

```
##G
```

```
interacmodel = lm(USAGE~PROD+TEMP+HOUR+PROD*TEMP+PROD*HOUR+TEMP*HOUR,
data=water)
summary(interacmodel)
```

```
##
## Call:
## lm(formula = USAGE ~ PROD + TEMP + HOUR + PROD * TEMP + PROD *
##     HOUR + TEMP * HOUR, data = water)
##
## 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

#As seen from the summary of summary(interacmodel), the interaction terms
PROD:TEMP and PROD:HOUR have t-values of 17.154 and 9.933, respectively, and
very low p-values (< alpha=0.05), indicating they are statistically
significant. This means that we failed to reject the null hypothesis that
these beta hat coefficients are zero and should be included in the model.

Recommended_model = lm(USAGE~PROD+TEMP+HOUR+PROD*TEMP+PROD*HOUR, data=water)
summary(Recommended_model)

##
## Call:
## lm(formula = USAGE ~ PROD + TEMP + HOUR + PROD * TEMP + PROD *
##     HOUR, data = water)
##
## 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

Recommended_model$coefficients

##   (Intercept)          PROD          TEMP          HOUR     PROD:TEMP
## 12.4257346600 -0.0025288273 -0.0047367734 -0.2150726648  0.0011417022
##     PROD:HOUR
##  0.0007873227

#I recommend: USAGE_hat = 12.4257346600 - 0.0025288273 * PROD - 0.0047367734
* TEMP - 0.2150726648 * HOUR + 0.0011417022 * PROD * TEMP + 0.0007873227 *
PROD * HOUR
```

```r
#It has significantly lower RMSE (0.9866) and substantially higher adjusted
r-squared (0.9647) than previous models.

#Problem 2
##A

gfclocks <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment
One/GFCLOCKS.csv")
head(gfclocks)
```

```
##   AGE NUMBIDS PRICE
## 1 127      13  1235
## 2 115      12  1080
## 3 127       7   845
## 4 150       9  1522
## 5 156       6  1047
## 6 182      11  1979
```

```r
gfclocks_Model = lm(PRICE~AGE+NUMBIDS, data=gfclocks)
gfclocks$coefficients
```

```
## NULL
```

```r
#Full model: Price_hat = -1338.95134 + 12.74057 * AGE + 85.95298 * NUMBIDS

##B
sse_method1 = sigma(gfclocks_Model)^2 * (32-2-1)
print(sse_method1)
```

```
## [1] 516726.5
```

```r
##C
price_hat = predict(gfclocks_Model, gfclocks[c('AGE', 'NUMBIDS')])
errors = gfclocks$PRICE - price_hat
squared_errors = errors ^ 2
sse_method2 = sum(squared_errors)
print(sse_method2)
```

```
## [1] 516726.5
```

```r
rmse = sqrt(sse_method2 / (32-2-1))
print(rmse)
```

```
## [1] 133.4847
```

```r
#From my calculation, the root mean square error (RMSE) is $133.4847, which
represents the standard deviation of the unexplained variance. Essentially,
it indicates the average discrepancy of $133.48 between the predicted price
from the model and the actual price in the dataset.

##D
summary(gfclocks_Model)$adj.r.squared
```

```
## [1] 0.8849194

#The adjusted R-squared, standing at 0.8849, signifies that approximately
88.49% of the total variation in the response variable PRICE is accounted for
by the regression model.

##E

#Null Hypothesis (H0): The model does not provide any meaningful
contribution, and all slope coefficients are zero: β(AGE) = β(NUMBIDS) = 0.
Alternative Hypothesis (Ha): At least one slope coefficient is not zero.

#The ANOVA test output reveals an F-value of 120.19 (with degrees of freedom
= 2, 29) and a p-value < 9.216e-15, which is significantly less than the
significance level of 0.05. Therefore, we reject the null hypothesis. The
substantial F-test indicates that at least one coefficient should be
statistically significant.

anova(lm(PRICE~1, data = gfclocks), gfclocks_Model)

## 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

#The ANOVA table results indicate an F-statistic of 120.19 with 2 and 29
degrees of freedom and a p-value less than 9.216e-15. Since this p-value is
much smaller than the significance level of 0.05, we reject the null
hypothesis. The large F-test suggests that at least one coefficient should be
significant, implying a linear relationship with PRICE.

##F
#Null Hypothesis (H0): β(NUMBIDS) = 0, implying that on average, price
remains unchanged when the number of bidders changes (age held
constant).Alternative Hypothesis (Ha): β(NUMBIDS) ≠ 0, indicating that on
average, price does change when the number of bidders changes (age held
constant).

gfclocks_Model = lm(PRICE~AGE+NUMBIDS, data=gfclocks)
summary(gfclocks_Model)

##
## Call:
## lm(formula = PRICE ~ AGE + NUMBIDS, data = gfclocks)
##
## 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
```

#Based on the summary output, the t-value for NUMBIDS stands at 9.847, with a p-value close to zero (much lower than the significance level of 0.05). This strongly suggests that it's improbable for β(NUMBIDS) to be zero. Consequently, with 95% confidence, we reject the null hypothesis, indicating a linear relationship between our tested variable and price. The estimated value, 85.953, is positive, signifying that price increases by approximately $85.95 on average when the number of bidders increases by 1 person, holding other variables constant.

##G

```
confint(gfclocks_Model)
```

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

 # We are 95% confident that the true value of beta1 (AGE) lies within the interval of 10.89 to 14.59. This indicates that, on average, the price increases by these amounts given a one-year increase in the clock's age, while holding other variables constant.

##H

```
clocks_interaction_model = lm(PRICE~AGE+NUMBIDS+AGE*NUMBIDS, data = gfclocks)
summary(clocks_interaction_model)
```

```
##
## Call:
## lm(formula = PRICE ~ AGE + NUMBIDS + AGE * NUMBIDS, data = gfclocks)
##
## 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
```

#The interaction term AGE:NUMBIDS is indeed statistically significant, as evidenced by its t-value of 6.112 and a very low p-value of 1.35e-06, which is much less than the significance level of 0.05. Thus, it should be included in the model.

```
cat("The model containing all predictors has adjusted r-squared =",
    summary(gfclocks_Model)$adj.r.squared,
    "\n",
    "The model that includes interaction term has adjusted r-squared =",
    summary(clocks_interaction_model)$adj.r.squared,
    "\n\n")
```

```
## The model containing all predictors has adjusted r-squared = 0.8849194
##  The model that includes interaction term has adjusted r-squared =
0.9489395
```

```
cat("The model containing all predictors has RMSE =",
    sigma(gfclocks_Model),
    "\n",
    "The model that includes interaction term has RMSE =",
    sigma(clocks_interaction_model))
```

```
## The model containing all predictors has RMSE = 133.4847
##  The model that includes interaction term has RMSE = 88.91451
```

#The adjusted R-squared value increases from 0.8849 in the full model to 0.9489 in the model that includes the interaction term. This indicates that the model with the interaction term explains a higher proportion of the variation in the response variable, suggesting a better fit.Additionally, the RMSE decreases from 133.4847 in the full model to 88.91451 in the model with the interaction term. A lower RMSE indicates better predictive accuracy.Therefore, the model with the interaction term is preferred over the full model without it, as it provides a better fit to the data and yields more accurate predictions.

*#The model that is recommended for predicting: PRICE_hat = 320.4579934 + 0.8781425 * AGE - 93.2648244 * NUMBIDS + 1.2978458 * AGE * NUMBIDS*

```r
#Question 3
##A
TURBINE <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment One/TURBINE.csv")
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
turbine_full_model = lm(HEATRATE~CPRATIO+RPM+INLET.TEMP+EXH.TEMP+AIRFLOW, data = TURBINE)
turbine_full_model$coefficients
```

```
##   (Intercept)       CPRATIO           RPM    INLET.TEMP      EXH.TEMP
##  1.361446e+04  3.519043e-01  8.878591e-02 -9.200873e+00  1.439385e+01
##       AIRFLOW
## -8.479583e-01
```

```r
#First-order mode: HEATRATE_hat = 1.361446e+04 + 3.519043e-01 * CPRATIO + 8.878591e-02 * RPM - 9.200873 * INLET.TEMP + 1.439385e+01 * EXH.TEMP - 8.479583e-01 * AIRFLOW

##B
#Null Hypothesis (H0): The model itself does not add any value, and all slope coefficients are zero: β(RPM) = β(INLET.TEMP) = β(EXH.TEMP) = β(POWER) = β(AIRFLOW) = 0.Alternative Hypothesis (Ha): At least one slope coefficient differs from zero.

summary(turbine_full_model)
```

```
##
## Call:
## lm(formula = HEATRATE ~ CPRATIO + RPM + INLET.TEMP + EXH.TEMP +
##     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 ***
## CPRATIO      3.519e-01  2.956e+01   0.012 0.990539
## 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 ***
## 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

anova(lm(HEATRATE~1, data = TURBINE), turbine_full_model)

## Analysis of Variance Table
##
## Model 1: HEATRATE ~ 1
## Model 2: HEATRATE ~ CPRATIO + RPM + INLET.TEMP + EXH.TEMP + AIRFLOW
##   Res.Df       RSS Df Sum of Sq     F    Pr(>F)
## 1     66 167897208
## 2     61  12841935  5 155055273 147.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#The results of the summaries reveal an F-statistic of 147.3 with 5 and 61
degrees of freedom (p-value < 2.2e-16 < alpha = 0.01). This indicates a
significant rejection of the null hypothesis. The substantial F-test value
suggests that at least one coefficient is likely to be statistically
significant, indicating a linear relationship with HEATRATE.

##C
summary(turbine_full_model)

##
## Call:
## lm(formula = HEATRATE ~ CPRATIO + RPM + INLET.TEMP + EXH.TEMP +
##     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 ***
## CPRATIO      3.519e-01  2.956e+01   0.012 0.990539
## 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 ***
## 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
```

#The summary of the turbine full model reveals significant coefficients for most predictors. However, individual t-tests indicate that CPRATIO is not significant (p-value = 0.990539), therefore, it can be confidently removed from the model. For AIRFLOW, its significance is slightly below the adjusted significance level with a p-value of 0.0598. With an alpha level set at 0.06, AIRFLOW would still be considered non-significant. To make an informed decision about AIRFLOW's inclusion, I will proceed with comparing the adjusted R-squared values of two models: one with AIRFLOW and one without. It's worth noting that CPRATIO should be removed from both models as it's not significant.

```
cat("The model with AIRFLOW predictor has adjusted r-squared =",
    summary(lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+AIRFLOW, data =
TURBINE))$adj.r.squared,
    "\n",
    "Model without AIRFLOW predictor has adjusted r-squared =",
    summary(lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP, data =
TURBINE))$adj.r.squared,
    "\n\n")
```

```
## The model with AIRFLOW predictor has adjusted r-squared = 0.9185783
##  Model without AIRFLOW predictor has adjusted r-squared = 0.9150099
```

#Given that the model containing the AIRFLOW predictor exhibits a slightly higher adjusted R-squared value, I intend to incorporate this variable into my model. The output provided above offers evidence supporting the superiority of the model including this variable.

#This is the advised model:

```
turbine_current_model = lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+AIRFLOW, data =
TURBINE)
turbine_current_model$coefficients
```

```
##   (Intercept)           RPM    INLET.TEMP      EXH.TEMP       AIRFLOW
##  1.361792e+04  8.882334e-02 -9.185605e+00  1.436283e+01 -8.475203e-01
```

##D

```
turbine_interaction_model = lm(HEATRATE~(RPM+INLET.TEMP+EXH.TEMP+AIRFLOW)^2,
data = TURBINE)
summary(turbine_interaction_model)
```

```
##
## 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
```

#The summary of the turbine interaction model reveals coefficients for various interaction terms between RPM, INLET.TEMP, EXH.TEMP, and AIRFLOW. Among these, two interactions stand out as statistically significant: INLET.TEMP x AIRFLOW and EXH.TEMP x AIRFLOW. These significant interactions indicate that the combined effects of INLET.TEMP and AIRFLOW, as well as EXH.TEMP and AIRFLOW, have a notable impact on HEATRATE. Therefore, I will incorporate these interactions into the model.

```
turbine_model_adj =
lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+AIRFLOW+INLET.TEMP*AIRFLOW+EXH.TEMP*AIRFL
OW, data = TURBINE)
summary(turbine_model_adj)
```

```
##
## 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
anova(lm(HEATRATE~1, data = TURBINE), turbine_model_adj)
```

```
## Analysis of Variance Table
##
## Model 1: HEATRATE ~ 1
## Model 2: HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP + AIRFLOW + INLET.TEMP *
##     AIRFLOW + EXH.TEMP * AIRFLOW
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1     66 167897208
## 2     60   9664946  6 158232262 163.72 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#Based on the provided output, with an F-statistic of 163.7 and 6 and 60 degrees of freedom & p-value: < 2.2e-16, we have strong evidence to reject the null hypothesis, which suggests that the model incorporating interaction terms does not provide any valuable contributions, and that all slope coefficients are zero.

```r
turbine_model_adj$coefficients
```

```
##        (Intercept)                RPM         INLET.TEMP
EXH.TEMP
##       1.360331e+04       4.577613e-02      -1.279883e+01
2.327429e+01
##            AIRFLOW INLET.TEMP:AIRFLOW   EXH.TEMP:AIRFLOW
##       1.346949e+00       1.613280e-02      -4.149806e-02
```

*#My Recommended model is: HEATRATE_hat = 1.360331e+04 + 4.577613e-02 * RPM - 1.279883e+01 * INLET.TEMP + 2.327429e+01 * EXH.TEMP + 1.346949 * AIRFLOW + 1.613280e-02 * INLET.TEMP:AIRFLOW - 4.149806e-02 * EXH.TEMP:AIRFLOW*

```r
##E
#To interpret the coefficients, let's break down the main effects and interaction effects:
#Main Effects: RPM: For each increase of 1 revolution per minute, the heat rate, on average, increases by 0.0458 kilojoules per kilowatt per hour, holding other variables constant. INLET.TEMP: With every 1-degree Celsius increase in inlet temperature, the heat rate, on average, decreases by 12.80 kilojoules per kilowatt per hour, holding other variables constant. EXH.TEMP: A rise of 1 degree Celsius in exhaust gas temperature corresponds to an increase in the heat rate, on average, by 23.27 kilojoules per kilowatt per hour, holding other variables constant.
#Interaction Effects: The heat rate is influenced by the AIRFLOW variable, which exhibits a linear dependence on both INLET.TEMP and EXH.TEMP.

##F
sigma(turbine_model_adj)

## [1] 401.3508

##G
summary(turbine_model_adj)$adj.r.squared

## [1] 0.9366789

#The adjusted R-squared value of 0.9366789 indicates that approximately 93.67% of the total variation in the response variable, HEATRATE, can be explained by the regression model. This adjustment clarifies that the proportion of explained variation is represented by the adjusted R-squared value.

##H
favstats(~RPM, data = TURBINE)

##   min   Q1 median    Q3   max     mean       sd  n missing
##  3000 3600   5100 12610 33000 8326.642 7023.311 67       0

favstats(~INLET.TEMP, data = TURBINE)

##  min   Q1 median   Q3  max     mean       sd  n missing
##  888 1078   1149 1288 1427 1174.313 137.4331 67       0

favstats(~EXH.TEMP, data = TURBINE)

##  min    Q1 median    Q3 max     mean       sd  n missing
##  444 512.5    532 568.5 626 536.0896 44.13984 67       0

favstats(~AIRFLOW, data = TURBINE)

##  min Q1 median    Q3 max    mean      sd  n missing
##    3 27    172 442.5 737 240.791 226.714 67       0

new_data = data.frame(RPM=273145, INLET.TEMP=1240, EXH.TEMP=920, AIRFLOW=25)
predict(turbine_model_adj, new_data, interval = "predict")
```

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

##The 95% confidence interval for the HEATRATE, based on the provided parameters, ranges from 24067.74 to 38388.2 kilojoules per kilowatt per hour. However, it's important to interpret this result cautiously because some of the predictor values in our new data exceed the range of our sample data. Specifically, the cycle speed of 273,145 revolutions per minute exceeds the maximum speed observed in our sample, which is 33,000. Similarly, the exhaust temperature of 920 degrees Celsius exceeds our sample maximum of 626. These values fall outside the bounds of our sample data, potentially affecting the accuracy of the prediction.