DATA 603 HW 3

Kane Smith

2022-10-30

Contents

Problem 1

```
water <- read.csv("water.csv")
```

a)

```
water_fit <- lm(USAGE~TEMP+PROD+DAYS+HOUR, data = water)
summary(water_fit)$coef
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  5.89162697 1.028794291  5.7267298 2.998277e-08
## TEMP         0.16867306 0.008209366 20.5464176 3.901112e-55
## PROD         0.04020739 0.001629089 24.6809092 2.863363e-68
## DAYS        -0.02162304 0.032182776 -0.6718825 5.022941e-01
## HOUR        -0.07099009 0.016992311 -4.1777776 4.102173e-05
```

From our fitted model, we get a estimated multiple regression equation of:

$\widehat{USAGE} = 5.8916 + 0.1687TEMP + 0.0402PROD - 0.02162DAYS - 0.0710HOUR.$

b)

To test the hypothesis of the whole model, we will conduct a global F test. We will begin by stating the null and alternative hypothesis.

Null hypothesis: $H_0 : \beta_{TEMP} = \beta_{PROD} = \beta_{DAYS} = \beta_{HOUR} = 0$

Alternative hypothesis: At least one $\beta_i (i = TEMP, PROD, DAYS, HOUR)$ does not equal 0.

We will set the alpha value to 0.05.

```
summary(water_fit)
```

```
##
## Call:
## lm(formula = USAGE ~ TEMP + PROD + DAYS + HOUR, 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 ***
## TEMP         0.168673   0.008209  20.546  < 2e-16 ***
## PROD         0.040207   0.001629  24.681  < 2e-16 ***
## DAYS        -0.021623   0.032183  -0.672    0.502
## HOUR        -0.070990   0.016992  -4.178  4.1e-05 ***
## ---
## 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 summary output of our fitted model shows an Fcalc of 486 and a p-value of 2.2e-16. Since 2.2e-16 is less than our set alpha value of 0.05, we have compelling evidence against the null hypothesis and can therefore reject it. We can then conclude that at least one of the regression coefficients does not equal zero and the full model is overall significant.

c)

Although the model in b) is significant overall, I would not recommend it for predictive purposes. Below are the t-statistics and p-values for the individual coefficients in the model.

```
q1_t_vals <- c(20.546, 24.681, -0.672, -4.178)
q1_p_vals <-c(2e-16, 2e-16, 0.502, 4.1e-05)
q1_col_names <- c('variable', 't-value', 'p-value')
variable <- c('TEMP', 'PROD', 'DAYS', 'HOUR')
q1_df <- data.frame(variable, q1_t_vals, q1_p_vals)
colnames(q1_df) <- q1_col_names
knitr::kable(q1_df, caption = "T-values and P-values for Predictors")
```

Table 1: T-values and P-values for Predictors

variable	t-value	p-value
TEMP	20.546	0.000000
PROD	24.681	0.000000
DAYS	-0.672	0.502000
HOUR	-4.178	0.000041

All variables have very low p-values and are significant except for DAYS which has a p-value of 0.502. Therefore I would use a model with the variables TEMP, PROD, and HOUR to predict USAGE.

$\widehat{USAGE} = 5.8916 + \hat{\beta}_{TEMP}TEMP + \hat{\beta}_{PROD}PROD - \hat{\beta}_{HOUR}HOUR.$

d)

To conduct a partial F-test, we will begin by stating our null and alternative hypothesis.

Null hypothesis: $H_0 : \beta_{DAYS} = 0$ in the model $USAGE = \beta_0 + \beta_{TEMP}TEMP + \beta_{PROD}PROD - \beta_{DAYS}DAYS - \beta_{HOUR}HOUR + \epsilon$.

Alternative hypothesis: At least one $\beta_{DAYS} \neq 0$ in the model $USAGE = \beta_0 + \beta_{TEMP}TEMP + \beta_{PROD}PROD - \beta_{DAYS}DAYS - \beta_{HOUR}HOUR + \epsilon$.

We will set the alpha value to 0.05.

```
water_reduced <- lm(USAGE~TEMP+PROD+HOUR, data=water) # Fit model without DAYS variable
anova(water_reduced, water_fit)
```

```
## Analysis of Variance Table
##
## Model 1: USAGE ~ TEMP + PROD + HOUR
## Model 2: USAGE ~ TEMP + PROD + DAYS + HOUR
##   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
```

After dropping the DAYS variable, our ANOVA output gives a Fcalc of 0.4514 and a p-value of 0.5023, meaning that we fail to reject the null hypothesis that the coefficient for DAYS is equal to zero. Therefore, we conclude that we should drop the DAYS variable from our model.

e)

```
confint(water_reduced)
```

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

Using confint on the model without DAYS, we get a 95% confidence interval for TEMP of **(0.1531, 0.1853)**. This means that we can say with 95% confidence that the coefficient for TEMP is between 0.1531 and 0.1853.

In other words, if the average monthly temperature increases by one degree Celsius, we can say with 95% confidence that the mean monthly water usage will increase by between 0.1531 and 0.1853 gallons per minute.

f)

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

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

```
sigma(water_fit)
```

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

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

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

```
sigma(water_reduced)
```

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

For the full model from a), we get an adjusted R-squared of 0.886658 and a RMSE of 1.768414.

For the model without DAYS from c), we get an adjusted R-squared of 0.8869118 and a RMSE of 1.766433.

R-squared indicates how much of the variance in the dependent variable is explained by the model. Adjusted R-squared compensates for the fact that adding variables to the model will always increase R-squared. This means that both models explain about ~88% of the variance in USAGE, with the model from c) being slightly higher. Having higher adjusted R-squared is better.

RMSE shows the standard deviation of the variation that is not explained by the model. The lower RMSE, the better. The model built in part c) has a slightly lower RMSE than the model built in part a).

Because the model without the DAYS variable has a higher adjusted r-squared and a lower RMSE compared to the full model, I would choose the model without the DAYS variable for prediction purposes.

g)

```
water_reduced2 <- lm(USAGE~TEMP*PROD*HOUR, data=water)
summary(water_reduced2)
```

```
##
## Call:
## lm(formula = USAGE ~ TEMP * PROD * HOUR, data = water)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1433 -0.3150 -0.0509  0.2716  7.1922
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.316e+01  8.051e-01  16.343  < 2e-16 ***
## TEMP           -3.254e-02  2.604e-02  -1.249    0.213
## PROD           -5.399e-03  3.978e-03  -1.357    0.176
## HOUR           -2.453e-01  3.181e-02  -7.710 3.29e-13 ***
```

```
## TEMP:PROD       1.252e-03  1.297e-04   9.659  < 2e-16 ***
## TEMP:HOUR       1.224e-03  1.111e-03   1.101    0.272
## PROD:HOUR       8.891e-04  2.096e-04   4.243 3.15e-05 ***
## TEMP:PROD:HOUR -4.218e-06  7.294e-06  -0.578    0.564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.988 on 241 degrees of freedom
## Multiple R-squared:  0.9656, Adjusted R-squared:  0.9646
## F-statistic: 966.9 on 7 and 241 DF,  p-value: < 2.2e-16
```

From our output, the statistically significant interaction terms are (TEMPxPROD) and (PRODxHOUR) with p-values of 3.15e-05 and < 2e-16 respectively. Therefore, we should keep these interaction terms and drop the rest.

```
water_reduced3 <- lm(USAGE~TEMP+PROD+HOUR+PROD:HOUR+ TEMP:PROD, data=water)
summary(water_reduced3)
```

```
##
## Call:
## lm(formula = USAGE ~ TEMP + PROD + HOUR + PROD:HOUR + TEMP:PROD,
##     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 ***
## TEMP        -4.737e-03  8.859e-03  -0.535    0.593
## PROD        -2.529e-03  2.305e-03  -1.097    0.274
## HOUR        -2.151e-01  1.624e-02 -13.242   <2e-16 ***
## PROD:HOUR    7.873e-04  7.745e-05  10.165   <2e-16 ***
## TEMP:PROD    1.142e-03  5.009e-05  22.795   <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
```

The model I would recommend for prediction purposes is: $\widehat{USAGE} = 1.243 - 0.0047373TEMP - 0.002529PROD - 0.2151HOUR + 0.0007873(PROD * HOUR) + 0.001142(TEMP * PROD)$.

Problem 2

```
gfclocks <- read.csv("GFCLOCKS.csv")
```

a)

```
gfc_fit <- lm(PRICE~AGE+NUMBIDS, data=gfclocks)
summary(gfc_fit)$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -1338.95134 173.8094707 -7.703558 1.705814e-08
## AGE            12.74057   0.9047403 14.082023 1.692756e-14
## NUMBIDS        85.95298   8.7285233  9.847368 9.344953e-11
```

Using the least squares method, we get the following estimates:

β_0 (Intercept): -1338.95134

β_1 (AGE): 12.74057

β_2 (NUMBIDS): 85.95298

b)

To calculate SSE, we can use the following equation: $SSE = \sum (\hat{y}_i - y_i)^2$.

```
sse <- sum((gfc_fit$fitted.values-gfclocks$PRICE)^2)
sse
```

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

We get an SSE of 516,726.5.

c)

```
sigma(gfc_fit)
```

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

RMSE is in the units of our predicted variable. So from our output, we get a RMSE of $133.48. This is the standard deviation of the variation that is not explained by the fitted model.

d)

```
summary(gfc_fit)$adj.r.square
```

```
## [1] 0.8849194
```

We get an adjusted R-square of 0.8849194. This means about 88.49194% of the variance in the auction price of clocks is explained by the fitted model using age of the clock and number of bidders in the auction.

e)

Below is the calculation of the ANOVA table:

```
reg_df = 2
res_df = nrow(gfclocks) - reg_df - 1
total_df = reg_df + res_df
ssr = sum((gfc_fit$fitted.values - mean(gfclocks$PRICE))^2)
sst = sse + ssr
msr = ssr/reg_df
mse = sse/res_df
f_calc = msr/mse
col_1 <- c("Regression", "Residual", "Total")
col_2 <- c(reg_df, res_df, total_df)
col_3 <- c(ssr, sse, sst)
col_4 <- c(msr, mse, "")
col_5 <- c(f_calc, "", "")
anova_df <- data.frame(col_1, col_2,col_3,col_4, col_5)
colnames(anova_df) <- c("Source of Variation", "Df", "Sum of Squares", "Mean Square", "F-Statistic")
knitr::kable(anova_df, caption = "ANOVA Table")
```

Table 2: ANOVA Table

Source of Variation	Df	Sum of Squares	Mean Square	F-Statistic
Regression	2	4283063.0	2141531.48005036	120.188161679417
Residual	29	516726.5	17818.1565482511	
Total	31	4799789.5		

To do conduct a global F-test for our model, we must first state our null and alternative hypothesis:

Null hypothesis: $H_0 : \beta_1 = \beta_2 = 0$

Alternative hypothesis: At least one $\beta_i \neq 0$ $(i = 1, 2)$

We will set an alpha value of 0.05.

```
summary(gfc_fit)
```

```
##
## 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
```

The summary output of our fitted model shows an Fcalc of 120.2 and a p-value of 9.216e-15 which is less than our set alpha value of 0.05. Therefore we can reject our null hypothesis that all of the regression coefficients are equal to zero. We can then conclude that at least one of the regression coefficients does not equal zero and the full model is overall significant.

f)

To test whether mean auction price of a clock increases as the number of bidders increases when age is held constant, we need to test whether we can say that the coefficient of NUMBIDS is different from 0. To do this, we will do a partial F-test. We will start by stating the null and alternative hypothesis.

Null hypothesis: $H_0 : \beta_2 = 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

Alternative hypothesis: $\beta_2 \neq 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

We will set an alpha value of 0.05.

```
gfc_reduced <- lm(PRICE~AGE, data = gfclocks)
anova(gfc_reduced, gfc_fit)
```

```
## Analysis of Variance Table
##
## Model 1: PRICE ~ AGE
## Model 2: PRICE ~ AGE + NUMBIDS
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     30 2244565
## 2     29  516727  1   1727838 96.971 9.345e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From our test we get a Fcalc of 96.971 and a p-value of 9.345e-11 which is less than our set alpha value of 0.05. This means we can reject the null hypothesis and conclude with 95% significance level than $\beta_1 \neq 0$ meaning that the mean auction price of clocks increase as the number of bidders increase when age is held constant and therefore we should keep NUMBIDS in our model.

We can also refer to the summary of the full fitted model to see the t-stat 9.847 and p-value of 9.34e-11 for NUMBIDS which confirms our conclusion from the partial F-test.

g)

```
confint(gfc_fit)
```

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

We get a 95% confidence interval for β_1 (AGE) of **(10.8902, 14.5910)**. This means that when the age of the clock increases by 1 year, we can say with 95% confidence that the mean auction price of the clock will increase between 10.89 and 14.59 dollars.

h)

```
gfc_interact <- lm(PRICE~AGE*NUMBIDS, data = gfclocks)
summary(gfc_interact)
```

```
##
## Call:
## lm(formula = PRICE ~ 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
```

From our output, we get a p-value for the interaction variable between AGE and NUMBIDS of 1.35e-06 which is less than our set alpha value of 0.05. Therefore, we can keep the interaction variable in our model. The model I would suggest using to predict the auction price of a clock (y) would be: $\hat{y} = 320.458 + 0.8781X_1 - 93.2648X_2 + 1.2978(X_1 * X_2)$

Problem 3

```
turbine <- read.csv("TURBINE.csv")
```

a)

```
#Fit the full model
turbine_fit <- lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+CPRATIO+AIRFLOW, data = turbine)
summary(turbine_fit)
```

```
##
## 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
```

First order model:

$\widehat{HEATRATE} = 13610 + 0.08879RPM - 9.201INLET + 14.39EXH + 0.3519CPRATIO - 0.8480AIRFLOW$

b)

To test the overall significance of the model, we will do a global F-test. First we will state the null and alternative hypothesis.

Null hypothesis: $H_0 : \beta_{RPM} = \beta_{INLET} = \beta_{EXH} = \beta_{CPRATIO} = \beta_{AIRFLOW} = 0$

Alternative hypothesis: At least one $\beta_i \neq 0$ $(i = RPM, INLET, EXH, CPRATIO, AIRFLOW)$

We will set an alpha value of 0.01.

```
turbine_fit <- lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+CPRATIO+AIRFLOW, data = turbine) # Fit the full mode
turbine_null <- lm(HEATRATE~1, data = turbine) # Fit the model with only intercept
anova(turbine_null, turbine_fit)
```

```
## Analysis of Variance Table
##
## Model 1: HEATRATE ~ 1
## Model 2: HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP + CPRATIO + 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
```

From our global F-test, we get an Fcalc of 147.3 and a p-value of <2.2e-16. The p-value is less than our set alpha value of 0.01, meaning we can reject our null hypothesis that all the coefficients in our model are 0. We can then conclude with a significance level of 0.01 that at least one of the coefficients is not equal to 0 and our overall model is significant.

c)

CPRATIO has a t-stat of 0.012 and p-value of 0.990539, meaning we should take this out of our model as it is not a significant predictor of HEATRATE.

AIRFLOW is the variable that has a p-value close to 0.05. We will evaluate the model with and without this variable. If the addition of AIRFLOW shows an increase in model performance, we will keep it in.

Null hypothesis: $H_0 : \beta_{AIRFLOW} = 0$ in the model $Y = \beta_0 + \beta_{RPM} RPM + \beta_{INLET} INLET + \beta_{EXH} EXH + \beta_{AIRFLOW} AIRFLOW + \epsilon$.

Alternative hypothesis: $\beta_{AIRFLOW} \neq 0$ in the model $Y = \beta_0 + \beta_{RPM} RPM + \beta_{INLET} INLET + \beta_{EXH} EXH + \beta_{AIRFLOW} AIRFLOW + \epsilon$.

```
turbine_reduced <- lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+AIRFLOW, data = turbine) # With AIRFLOW
turbine_reduced2<- lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP, data = turbine) # Without AIRFLOW
summary(turbine_reduced)
```

```
##
## 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
```

```
summary(turbine_reduced2)
```

```
##
## 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
```

```
anova(turbine_reduced2, turbine_reduced)
```

```
## 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
```

```
sigma(turbine_reduced)
```

```
## [1] 455.1137
```

```
sigma(turbine_reduced2)
```

```
## [1] 464.9797
```

Both model provide a p-value of 2.2e-16. The model with AIRFLOW provides a higher adjusted R-squared value (0.9186 vs. 0.915) and a lower RMSE (455.1137 vs. 464.9797). Therefore, we will keep AIRFLOW in the model for predictive purposes as it improves the adjusted R-squared and RMSE albeit only slightly.

The model I would recommend for predictive purposes is:

$\widehat{HEATRATE} = 13620 + 0.08882RPM - 9.186INLET + 14.36EXH - 0.8475AIRFLOW$

d)

Creating a two-way interaction model of all our variables:

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

```
##
## 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
```

Evaluating individual t-test with alpha value of 0.05.

From our output,the interaction variable that are statically significant according to their t-tests are (INLET x AIRFLOW) and (EXH x AIRFLOW). Therefore, we will include these in our interaction model.

```
turbine_interaction_final <- lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+AIRFLOW+INLET.TEMP:AIRFLOW+EXH.TEMP:AI
summary(turbine_interaction_final)
```

```
##
## 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
```

Although from the individual t-test, AIRFLOW is statistically insignificant, it's interaction with INLET and EXH is significant so we should keep it in the model based on the hierarchical principle.

Therefore, the final model I would suggest for predicting HEATRATE (y) would be: $\widehat{HEATRATE} = 13600 + 0.04578RPM - 12.80INLET + 23.27EXH + 1.347AIRFLOW + 0.01613(INLET * AIRFLOW) - 0.04150(EXH * AIRFLOW)$

e)

Practical interpretations:

β_0 **(Intercept)**: When RPM, inlet temperature, exhaust temperature are all 0, HEATRATE will be 13600 kJ/KW/h.

RPM: For an increase in the RPM of 1 rev/min, HEATRATE will increase by an average 0.04578 kJ/KW/h.

INLET: For an increase in the inlet temperature of 1 degree Celsius, HEATRATE will decrease by an average 12.80 kJ/KW/h.

EXH: For an increase in the exhaust temperature of 1 degree Celsius, HEATRATE will increase by an average 23.27 kJ/KW/h.

AIRFLOW: For an increase in the airflow of 1 kg/s, HEATRATE will increase by an average 1.347 kJ/KW/h.

(INLET x AIRFLOW): For an increase in the inlet temperature of 1 degree Celsius, HEATRATE will increase by an average 0.01613*AIRFLOW kJ/KW/h.

(EXH x AIRFLOW): For an increase in the exhaust temperature of 1 degree Celsius, HEATRATE will decrease by an average 0.04150*AIRFLOW kJ/KW/h.

f)

```
sigma(turbine_interaction_final)
```

```
## [1] 401.3508
```

From our output, we get a RMSE of 401.3508 kJ/KW/h.

g)

```
summary(turbine_interaction_final)
```

```
##
## 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
```

The adjusted R-squared value of the model from d) is 0.9367. This means that about 93.67% of the variance in HEATRATE can be explained using our model.

h)

Predicting HEATRATE when:

RPM = 273145

INLET.TEMP = 1240

EXH.TEMP = 920

CPRATIO = 10 (NOT IN OUR MODEL)

AIRFLOW = 25

```
13600 + 0.04578*273145 - 12.80*1240 + 23.27*920 + 1.347*25 + 0.01613*(1240*25)-0.04150*(920*25)
```

```
## [1] 31220.18
```

We get a predicted HEATRATE of **31,220.18 kJ/KW/h.**

However, we must check that the values we are plugging into our model are within the range of our original data used to fit the model. If it is not, we extrapolating. This is bad because we do not know how the data behaves outside of our range, and therefore should not be making predictions.

```
fav_stats <- favstats(~RPM, data = turbine)[c("min", "max")]
fav_stats <- rbind(fav_stats, favstats(~INLET.TEMP, data = turbine)[c("min", "max")], favstats(~EXH.TEM
fav_stats["variable"] <- c("RPM", "INLET.TEMP", "EXH.TEMP", "CPRATIO", "AIRFLOW")
knitr::kable(fav_stats, caption = "Min and Max")
```

Table 3: Min and Max

	min	max	variable
	3000.0	33000	RPM
1	888.0	1427	INLET.TEMP
2	444.0	626	EXH.TEMP
3	6.6	35	CPRATIO
4	3.0	737	AIRFLOW

In our case, our values for EXH.TEMP fall out of the range of our original data. Therefore we should not trust our predicted value of HEATRATE from our model due to our input values.

Session Info:

```
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-conda-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.1 LTS
##
## Matrix products: default
## BLAS/LAPACK: /opt/conda/lib/libopenblasp-r0.3.21.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] mosaic_1.8.3      ggridges_0.5.3    mosaicData_0.20.2 ggformula_0.10.1
## [5] ggstance_0.3.5    Matrix_1.4-1      lattice_0.20-45   dplyr_1.0.9
## [9] ggplot2_3.3.6
##
## loaded via a namespace (and not attached):
##  [1] ggrepel_0.9.1     Rcpp_1.0.9        tidyr_1.2.0       assertthat_0.2.1
##  [5] digest_0.6.29     utf8_1.2.2        ggforce_0.3.4     R6_2.5.1
##  [9] plyr_1.8.7        backports_1.4.1   labelled_2.9.1    evaluate_0.16
## [13] highr_0.9         pillar_1.8.1      rlang_1.0.4       rstudioapi_0.14
## [17] rmarkdown_2.15    splines_4.1.3     readr_2.1.2       stringr_1.4.1
## [21] htmlwidgets_1.5.4 polyclip_1.10-0   munsell_0.5.0     broom_1.0.0
## [25] compiler_4.1.3    xfun_0.32         pkgconfig_2.0.3   htmltools_0.5.3
## [29] tidyselect_1.1.2  tibble_3.1.8      gridExtra_2.3     mosaicCore_0.9.0
## [33] fansi_1.0.3       tzdb_0.3.0        withr_2.5.0       MASS_7.3-58.1
## [37] grid_4.1.3        gtable_0.3.0      lifecycle_1.0.1   DBI_1.1.3
## [41] magrittr_2.0.3    scales_1.2.1      cli_3.3.0         stringi_1.7.8
## [45] farver_2.1.1      leaflet_2.1.1     ellipsis_0.3.2    ggdendro_0.1.23
## [49] generics_0.1.3    vctrs_0.4.1       tools_4.1.3       forcats_0.5.2
## [53] glue_1.6.2        tweenr_2.0.1      purrr_0.3.4       hms_1.1.2
## [57] crosstalk_1.2.0   fastmap_1.1.0     yaml_2.3.5        colorspace_2.0-3
## [61] knitr_1.39        haven_2.5.0
```