DATA 603 Assignment 1

Michael Ellsworth

November 1st, 2019

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?

```
water_model_full = lm(data = water, USAGE ~ PROD + TEMP + HOUR + DAYS)
coefficients(water_model_full)
```

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

From the coefficients above, the estimated regression equation is as follows:

$$\widehat{Usage} = 5.89 + 0.04 * Production + 0.17 * Temperature - 0.07 * Hours - 0.02 * Days$$

b

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

The hypothesis is:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_A : \text{at least one } \beta_i \text{ is not zero}$$

```
# Create Model with only the intercept
water_model_intercept <- lm(data = water, USAGE ~ 1)

# Create ANOVA table inputs using anova function with our two models
anova(water_model_intercept, water_model_full)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	248	6842.7189	NA	NA	NA	NA
2	244	763.0582	4	6079.661	486.0171	6.503243e-115

2 rows

Since the P-value calculated above is ≈ 0, which is less than 0.05, we can say the model is significant or we can reject the null hypothesis.

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.

By testing the individual coefficients for significance (individual t-test), we can see that the DAYS variable is not significant (P-value is greater than 0.05). This would suggest the model in part b should not be used for predictive purposes.

```
summary(water_model_full)
```

```
##
## 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
```

```
# Drop the DAYS variable from the model
water_model_reduced <- lm(data = water, USAGE ~ PROD + TEMP + HOUR)
summary(water_model_reduced)
```

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

```
coefficients(water_model_reduced)
```

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

By dropping the DAYS variable, each variable is now significant and the model I would suggest for predictive purposes becomes:

$$\widehat{Usage} = 5.31 + 0.04 * Production + 0.17 * Temperature - 0.07 * Hours$$

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.

```
# Test if H0: DAYS = 0
anova(water_model_reduced, water_model_full)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	245	764.4699	NA	NA	NA	NA
2	244	763.0582	1	1.411739	0.4514261	0.5022941

2 rows

By removing the DAYS variable from the model, we can test the effect that DAYS has in the full model. Since the P-value in the analysis of variance is greater than 0.05, we cannot reject H_0 and can drop DAYS from the model as it does not have a significant effect.

e

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

```
# Complete an interval estimation of the regression coefficients in the reduced model
confint(water_model_reduced)
```

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

The 95% confidence interval of the regression coefficient for the variable TEMP is:

$$0.153 \leq \beta \leq 0.185$$

This would suggest that as the average monthly temperature increases by 1C, the water usage will increase between 0.153 and 0.185 gallons per minute with 95% confidence.

f

Use the method of Model Fit to calculate R_a^2dj 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_a^2dj and RMSE.

```
# Calculate the Adjusted R-Squared for the full and reduced models
summary(water_model_full)$adj.r.squared
```

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

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

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

```
#Calculate the RMSE for the full and reduced models
sigma(water_model_full)
```

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

```
sigma(water_model_reduced)
```

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

As the Adjusted R-Squared is higher and the RMSE is lower for the reduced model compared with the full model, I would suggest the reduced model (without the DAYS variable) for predictive purposes. The R_a^2dj suggests that approximately 89% of the variance in water usage can be explained by the model described in part c. The RMSE for the reduced model is 1.77 gallons/minute which is the standard deviation of the unexplained variance of

g

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?

```
water_model_interaction <- lm(data = water, USAGE ~ (PROD + TEMP + HOUR)**2)
summary(water_model_interaction)
```

```
##
## Call:
## lm(formula = USAGE ~ (PROD + TEMP + HOUR)^2, 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
```

From the summary above, it appears as though the interaction between TEMP and HOUR does not have a significant effect on the model. This interaction will be removed from the interaction model and a reduced interaction model will be proposed.

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

```
##
## 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
```

The model I would recommend for predictive purposes is as follows:

$$\widehat{Usage} = 12.43 - 0.003 * PROD - 0.005 * TEMP - 0.2 * HOUR + 0.001 * PROD * TEMP + 0.0008 * PROD * HOUR$$

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.

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 $\beta_0, \beta_1, \beta_2$ *of the model.*

```
clocks_model_full <- lm(data = clocks, PRICE ~ AGE + NUMBIDS)
clocks_model_full$coefficients
```

```
## (Intercept)         AGE     NUMBIDS
## -1338.95134    12.74057    85.95298
```

Based on the least square estimate, the model parameters are estimated as follows:

$$\widehat{PRICE} = -1338.95 + 12.74 * AGE + 85.95 * NUMBIDS$$

Where:

$$\beta_0 = -1338.95$$
$$\beta_1 = 12.74$$
$$\beta_2 = 85.95$$

b

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

```
clocks_model_intercept <- lm(data = clocks, PRICE ~ 1)
anova(clocks_model_intercept, clocks_model_full)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	31	4799789.5	NA	NA	NA	NA
2	29	516726.5	2	4283063	120.1882	9.216359e-15

2 rows

Based on the anova function above, the SSE = 516727.

c

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

```
# Calculate RMSE using SSE, n and p
RMSE <- (SSE / (n - p - 1))**0.5
RMSE
```

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

```
# Alternative - calculated RMSE using the sigma function
sigma(clocks_model_full)
```

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

Based on the calculations above, the RMSE = 133.48. This is the standard deviation of the unexplained variance meaning the unexplained variance in the auction price has a standard deviation of $133.48.

d

Find and interpret the adjusted coefficient of determination

```
# Calculate the adjusted r squared using SSE, SST, n and p
adj_r_squared <- 1 - (SSE / (n - p - 1)) / (SST / (n - 1))
adj_r_squared
```

```
## [1] 0.8849193
```

```
# Alternative - calculate the adjusted r squared using the summary function
summary(clocks_model_full)$adj.r.squared
```

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

Based on the calculated adjusted R-squared value above, we can say that 88% of the variation in the clock price can be explained by the model described in part a.

e

Construct the Anova table for the model and test the global F-test of the model at the α = 0.05 level of significance.

```
# Create anova table inputs
df_resid <- n - p - 1
MSR <- SSR / p
MSE <- SSE / df_resid
F_stat <- MSR / MSE

# Create anova table
header <- c("Source of Variation", "Df", "Sum of Squares", "Mean Squares", "F-Statistic")
anova_table <- data.frame(rbind(c("Regression", p, SSR, MSR, F_stat),
                                c("Residual", df_resid, SSE, MSE, ""),
                                c("Total", n - 1, SSR + SSE, "", "")))
names(anova_table) <- header
anova_table
```

Source of Variation <fctr>	Df <fctr>	Sum of Squares <fctr>	Mean Squares <fctr>	F-Statistic <fctr>
Regression	2	4283063	2141531.5	120.188055781873
Residual	29	516727	17818.1724137931	
Total	31	4799790		

3 rows

```
anova(clocks_model_intercept, clocks_model_full)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	31	4799789.5	*NA*	*NA*	*NA*	*NA*
2	29	516726.5	2	4283063	120.1882	9.216359e-15

2 rows

From the anova function above, we can see the P-value (9.216e-15) is less than 0.05.

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 $\alpha = 0.05$

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0$$

```
clocks_model_reduced <- lm(data = clocks, PRICE ~ AGE)
anova(clocks_model_reduced, clocks_model_full)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	30	2244565.0	NA	NA	NA	NA
2	29	516726.5	1	1727838	96.97066	9.344953e-11

2 rows

Since the P-value from the anova calculation above is less than 0.05, we can conclude that β_1 is not 0 (reject the null hypothesis). The mean auction price of a clock increases as the number of bidders increases when age is held constant.

g

Find a 95% confidence interval for $_1 and interpret the result.

```
confint(clocks_model_full)
```

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

From the confidence interval function above, we can see that with 95% confidence, β_1, will be a positive number between 10.9 and 14.6. Because of this, we can say that the average auction price of a clock will increase as the age of clocks increases. Additionally, we can say with 95% confidence that the auction price of a clock will increase between $10.9 and $14.6 when the age of the clock increases by 1 year.

h

Test the interaction term between the 2 variables at $\text{\alpa} = 0.05$. *What model would you suggest to use for predicting y? Explain.*

```
clocks_model_interact <- lm(data = clocks, PRICE ~ NUMBIDS + AGE + NUMBIDS * AGE)
summary(clocks_model_interact)
```

```
##
## Call:
## lm(formula = PRICE ~ NUMBIDS + AGE + NUMBIDS * AGE, data = clocks)
##
## 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
## NUMBIDS     -93.2648    29.8916  -3.120  0.00416 **
## AGE           0.8781     2.0322   0.432  0.66896
## NUMBIDS:AGE   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 summary of the interaction model, we can see that the P-value for the interaction term is less than 0.05 meaning we can use this interaction term to help us model the price of clocks. The model for predicting y would become:

$$\widehat{PRICE} = 320.46 + 0.88 * AGE - 93.26 * NUMBIDS + 1.30 * NUMBIDS * AGE$$

Based on the Adjusted R-squared value in the summary of the interaction model, it is 0.9489 which is higher than the Adjusted R-squared value calculated in part d for the model without the interaction term. Additionally, the RMSE is also lower in the model that includes the interaction term. This further supports us using the interaction model to predict the auction price of clocks.

Problem 3

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.

```
turbines_model <- lm(data = turbines, HEATRATE ~ RPM + INLET + EXHAUST + CPRATIO + AIRFLOW)
coefficients(turbines_model)
```

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

The model parameters are estimated as follows:

$$\widehat{Heatrate} = 13614.46 + 0.09 * RPM - 9.2 * Inlet + 14.4 * Exhaust + 0.35 * CPR - 0.85 * Airflow$$

b

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

```
# Anova function
turbines_model_intercept <- lm(data = turbines, HEATRATE ~ 1)
anova(turbines_model_intercept, turbines_model)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	66	167897208	NA	NA	NA	NA
2	61	12841935	5	155055273	147.3045	1.06715e-32

2 rows

```
# Alternative
summary(turbines_model)
```

```
##
## Call:
## lm(formula = HEATRATE ~ RPM + INLET + EXHAUST + CPRATIO + AIRFLOW,
##     data = turbines)
##
## 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       -9.201e+00  1.499e+00  -6.137 6.86e-08 ***
## EXHAUST      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
```

The P-value from the anova function and the summary of the First Order model is < 2.2e-16 which is less than 0.01. The model is significant as a result.

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.)

From the summary in part b, we can see that the CPRATIO and AIRFLOW independent variables have P-values greater than 0.05. In order to determine the best model, adjusted R-squared values will be calculated for models without CPRATIO, without AIRFLOW and without both CPRATIO and AIRFLOW.

```
# Test model without CPRATIO
turbines_model_reduced1 <- lm(data = turbines, HEATRATE ~ RPM + INLET + EXHAUST + AIRFLOW)
summary(turbines_model_reduced1)$adj.r.squared
```

```
## [1] 0.9185783
```

```
# Test model without AIRFLOW
turbines_model_reduced2 <- lm(data = turbines, HEATRATE ~ RPM + INLET + EXHAUST + CPRATIO)
summary(turbines_model_reduced2)$adj.r.squared
```

```
## [1] 0.9136684
```

```
# Compare models
turbines_model_reduced3 <- lm(data = turbines, HEATRATE ~ RPM + INLET + EXHAUST)
summary(turbines_model_reduced3)$adj.r.squared
```

```
## [1] 0.9150099
```

```
summary(turbines_model_reduced1)
```

```
##
## Call:
## lm(formula = HEATRATE ~ RPM + INLET + EXHAUST + AIRFLOW, data = turbines)
##
## 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       -9.186e+00  7.704e-01 -11.923  < 2e-16 ***
## EXHAUST      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
```

Since the adjusted R-squared value of the model without CPRATIO but with AIRFLOW, we will keep the AIRFLOW variable in our model. The P-value of AIRFLOW is close enough to 0.05 that it is significant enough to include in the model going forward.

```
turbines_model_reduced1$coefficients
```

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

The reduced model now becomes:

$$\widehat{Heatrate} = 13618 + 0.09 * RPM - 9.2 * Inlet + 14.4 * Exhaust - 0.85 * Airflow$$

d

Test all possible interaction terms for the best model in part (c) at $\alpha = 0.05$ *What is the final model would you suggest to use for predicting y? Explain.*

```
turbines_model_interact <- lm(data = turbines, HEATRATE ~ (RPM + INLET + EXHAUST + AIRFLOW)**2)
summary(turbines_model_interact)
```

```
## 
## Call:
## lm(formula = HEATRATE ~ (RPM + INLET + EXHAUST + AIRFLOW)^2, 
##     data = turbines)
## 
## 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           -2.366e+01  7.364e+00  -3.213 0.002180 ** 
## EXHAUST         -4.555e+00  1.795e+01  -0.254 0.800610    
## AIRFLOW          1.021e+01  6.279e+00   1.627 0.109455    
## RPM:INLET       -1.133e-04  8.720e-05  -1.299 0.199266    
## RPM:EXHAUST      1.656e-04  3.116e-04   0.531 0.597314    
## RPM:AIRFLOW     -8.257e-04  4.653e-04  -1.775 0.081414 .  
## INLET:EXHAUST    2.417e-02  1.457e-02   1.659 0.102791    
## INLET:AIRFLOW    1.418e-02  3.852e-03   3.681 0.000523 ***
## EXHAUST: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
```

Using the individual coefficients test (t-test) above, we can see that two interaction terms, $Exhaust * Airflow$ and $Inlet * Airflow$, are significant as they have P-values less than 0.05. The suggested model will drop the other interaction terms.

```
turbines_model_interact_reduced <- lm(data = turbines, HEATRATE ~ RPM + INLET + EXHAUST + AIRFLOW + INLET * AIRFL
OW + EXHAUST * AIRFLOW )
summary(turbines_model_interact_reduced)
```

```
## 
## Call:
## lm(formula = HEATRATE ~ RPM + INLET + EXHAUST + AIRFLOW + INLET * 
##     AIRFLOW + EXHAUST * AIRFLOW, data = turbines)
## 
## 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           -1.280e+01  1.090e+00 -11.741  < 2e-16 ***
## EXHAUST          2.327e+01  2.901e+00   8.024 4.46e-11 ***
## AIRFLOW          1.347e+00  3.496e+00   0.385 0.701414    
## INLET:AIRFLOW    1.613e-02  3.640e-03   4.432 4.03e-05 ***
## EXHAUST: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
```

```
coefficients(turbines_model_interact_reduced)
```

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

The final model for predicting heat rate becomes:

$$\begin{gathered}\widehat{Heatrate} = 13603 + 0.046 * RPM - 12.8 * Inlet + 23.3 * Exhaust + 1.3 * Airflow \\ + 0.016 * Inlet * Airflow - 0.04 * Exhaust * Airflow\end{gathered}$$

This final model does not include other interaction terms because they are not close enough to 0.05 to remain significant.

e

Give practical interpretations of the β_i estimates.

Based on the coefficient estimates calculated in part d, we can conclude that:

```
* Heat rate will increase when RPM increases and all other independent variables are held constant
* Heat value will decrease when inlet temperature increases and all other independent variables are held constant
    * Heat value will only decrease if the ratio of airflow to inlet temperature is less than 800:1
        * This is due to the interaction term of inlet temperature and airflow
* Heat value will increase if exhuast temperature increases
    * Heat value will only increase if the ratio of airflow to exhaust temperature is less than 582.5:1
        * This is due to the interaction term of exhaust temperature and airflow
```

Additionally, the model can be simplified as:

$$\begin{aligned}\widehat{Heatrate} &= 13603 + 0.046 * RPM + \phi_1 * Inlet + \phi_2 * Exhaust \\ \phi_1 &= 0.16 * Airflow - 12.8 \\ \phi_2 &= 23.3 - 0.04 * Airflow\end{aligned}$$

Looking at the ϕ terms closer:

```
* An increase in airflow will offset the reduction in heat rate as the inlet temperature increases
* An increase in airflow will offset the increase in heat rate as the exhuast temperature increases
```

f

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

From the summary of the `turbines_model_interact_reduced` model in part d, the RMSE is 401.4.

g

Find the adjusted R^2 value from the model in part (d) and interpret it.

From the summary of the `turbines_model_interact_reduced` model in part d, the Adjusted R-Squared is 0.9367. This means that 93.67% of the variation in the heat rate of a turbine with high pressure inlet fogging can be explained by the model described in part d.

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.

```
predict_df <- data.frame(RPM = 273145, INLET = 1240, EXHAUST = 920, AIRFLOW = 25)
predict(turbines_model_interact_reduced, predict_df, interval = "predict")
```

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

Based on the results of the predict function above, we can make a point estimate of heat rate of 31,228 kilojoules per kilowatt per hour.

Problem 4

The file tires.csv provides the results of an experiment on tread wear per 160 km and the driving speed in km/hour. The researchers looked at 2 types of tires and tested 20 random sample tires. The response variable is the tread wear per 160 km in percentage of tread thickness and the quantitative predictor is average speed in km/hour.

a

Define the dummy variable that explains the two types of tires.

The dummy variable is the “type” variable. It is defined as “A” or “B”. For the purpose of this problem, A = 0 and B = 1.

b

Test the additive model at $\alpha = 0.05$ *and write a first-order model for the tread wear per 160 km as a function of average speed and type of tires.*

```
tires_model_full <- lm(data = tires, wear ~ ave + type)
summary(tires_model_full)
```

```
##
## Call:
## lm(formula = wear ~ ave + type, data = tires)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.092858 -0.033451 -0.000953  0.039404  0.116668
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.6445083  0.0525675  -12.26   <2e-16 ***
## ave          0.0113094  0.0005155   21.94   <2e-16 ***
## typeB        0.1725006  0.0093544   18.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05384 on 137 degrees of freedom
## Multiple R-squared:  0.8861, Adjusted R-squared:  0.8844
## F-statistic: 532.8 on 2 and 137 DF,  p-value: < 2.2e-16
```

```
coefficients(tires_model_full)
```

```
## (Intercept)         ave       typeB
## -0.64450834  0.01130937  0.17250064
```

Since the P-value for all coefficients are less than 0.05, we will keep all independent variables in the model. The First Order Model is as follows:

$$\widehat{Wear} = -0.64 + 0.01 * Speed + 0.17 * Type$$

Or, if the tire is Type B:

$$\widehat{Wear} = -0.64 + 0.01 * Speed + 0.17$$

And if the tire is Type A:

$$\widehat{Wear} = -0.64 + 0.01 * Speed$$

Where:

$$\begin{aligned}\beta_0 &= \ -0.64 \\ \beta_1 &= \ 0.01 \\ \beta_2 &= \ 0.17\end{aligned}$$

c

Interpret all possible regression coefficient estimates.

Since the β_0 regression coefficient estimate is -0.64, we can assume that if driving speed were 0 and the Type of the tire was Type A, the tire would have a negative wear of -0.64 or -64% of tread thickness. This makes sense as a tire should not have any wear if it has not been used (average speed = 0).

Additionally, if driving speed were 0, and the Type of the tire was Type B, the tire would have slightly more wear of -0.47 or -47% of tread thickness. β_2 in this case indicates that Type B tires have slightly more wear than Type A tires. In other words, Type B tires would have a wear increase of 17% of tread thickness.

In regards to β_1, an average speed increase of 1 will increase the tire wear by 0.01 or 1% of tread thickness.

d

Test the interaction term between the 2 variables at $\alpha = 0.05$. What model would you suggest to use for predicting y? Explain.

```
tires_model_interact <- lm(data = tires, wear ~ (ave + type)**2)
summary(tires_model_interact)
```

```
## 
## Call:
## lm(formula = wear ~ (ave + type)^2, data = tires)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.070158 -0.016493 -0.003643  0.024086  0.063703 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.3888744  0.0347705  -11.18   <2e-16 ***
## ave          0.0087833  0.0003415   25.72   <2e-16 ***
## typeB       -1.0800050  0.0779442  -13.86   <2e-16 ***
## ave:typeB    0.0119840  0.0007439   16.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.03169 on 136 degrees of freedom
## Multiple R-squared:  0.9608, Adjusted R-squared:   0.96 
## F-statistic:  1112 on 3 and 136 DF,  p-value: < 2.2e-16
```

```
coefficients(tires_model_interact)
```

```
##  (Intercept)          ave        typeB    ave:typeB 
## -0.388874431  0.008783344 -1.080004985  0.011984013 
```

The summary above shows the interaction term between average speed and Type of tire is significant (P-value < 0.05) and therefore, we should use this interaction term for predicting y (tire wear).

The model in this case would be:

$$\widehat{Wear} = -0.39 + 0.01 * Speed - 1.08 * Type + 0.01 * Speed * Type$$

Or, if the tire is Type B:

$$\widehat{Wear} = -1.47 + 0.02 * Speed$$

And if the tire is Type A:

$$\widehat{Wear} = -0.39 + 0.01 * Speed$$

e

From the model in part (d) Find the adjusted-R^2 value and interpret it.

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

```
## [1] 0.9599663
```

From the adjusted r-squared calculation above, we can say that approximately 96% of a tires tread wear can be explained by the interaction model described in part d.

f

Predict the tread wear per 160 km in percentage of tread thickness for a car that has type A with an average speed 100 km/hour.

```
predict_df_tires <- data.frame(ave = 100, type = "A")
predict(tires_model_interact, predict_df_tires, interval = "predict")
```

```
##       fit       lwr       upr
## 1 0.48946 0.4263475 0.5525725
```

The tread wear is approximately 49% of tread thickness with type A tires with an average speed of 100 km/hr.

Problem 5

A team of mental health researchers wishes to compare three methods (A, B, and C) of treating severe depression. They would also like to study the relationship between age and treatment effectiveness as well as the interaction (if any) between age and treatment. Each member of a simple random sample of 36 patients, comparable with respect to diagnosis and severity of depression, was randomly assigned to receive treatment A, B, or C. The data are given in MentalHealth.csv.

a

Which is the dependent variable?

The dependent variable is “EFFECT”.

b

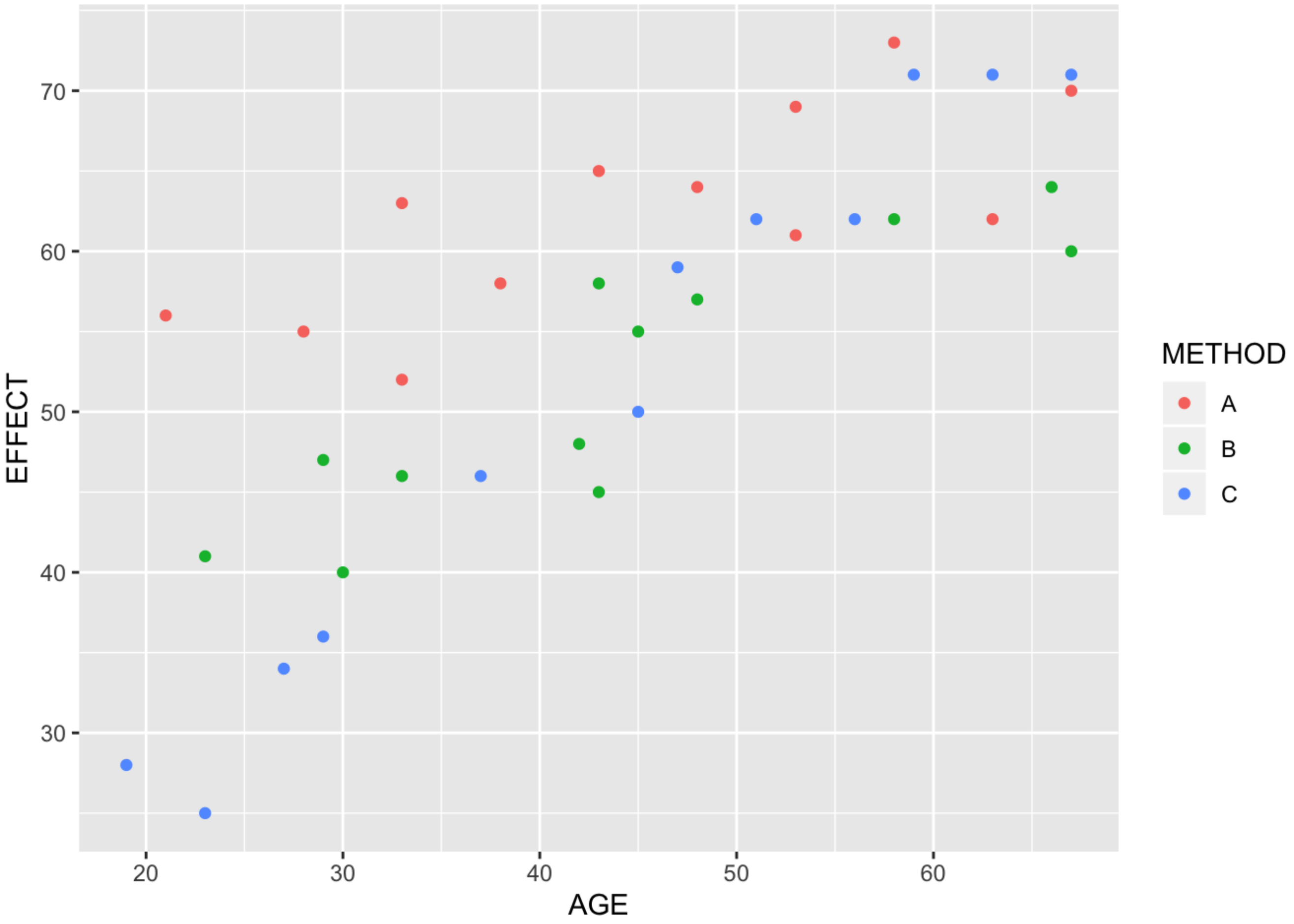
What are the independent variables?

The independent variables are “AGE” and “METHOD”.

c

Draw a scatter diagram of the sample data with EFFECT on the y-axis and AGE on the x-axis using different symbols/colors for each of the three treatments. Comment.

```
health %>%
  ggplot(aes(x = AGE, y = EFFECT, col = METHOD)) +
  geom_point()
```



For all three METHODs, there appears to be a general increase in EFFECT as AGE increases.

d

Is there any interaction between age and treatment? [Hint: Use dummy variable coding, the least square method and* $\alpha = 0.05$*]

The dummy coding for the following model will be as follows:

```
* Method A = 0, 0
* Method B = 1, 0
* Method C = 0, 1
```

```
health_full <- lm(data = health, EFFECT ~ METHOD + AGE)
summary(health_full)
```

```
## 
## Call:
## lm(formula = EFFECT ~ METHOD + AGE, data = health)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.5732  -3.3922   0.9829   3.9613   9.5062 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  32.54335    3.58105   9.088 2.23e-10 ***
## METHODB      -9.80758    2.46471  -3.979 0.000371 ***
## METHODC     -10.25276    2.46542  -4.159 0.000224 ***
## AGE           0.66446    0.06978   9.522 7.42e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.035 on 32 degrees of freedom
## Multiple R-squared:  0.784,  Adjusted R-squared:  0.7637 
## F-statistic: 38.71 on 3 and 32 DF,  p-value: 9.287e-11
```

```
coefficients(health_full)
```

```
## (Intercept)     METHODB     METHODC         AGE 
##  32.5433481  -9.8075777 -10.2527575   0.6644606 
```

Each variable appears to be significant. Test the interaction:

```
health_interact <- lm(data = health, EFFECT ~ (METHOD + AGE)**2)
summary(health_interact)
```

```
##
## Call:
## lm(formula = EFFECT ~ (METHOD + AGE)^2, data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4366 -2.7637  0.1887  2.9075  6.5634
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  47.51559    3.82523  12.422 2.34e-13 ***
## METHODB     -18.59739    5.41573  -3.434 0.001759 **
## METHODC     -41.30421    5.08453  -8.124 4.56e-09 ***
## AGE           0.33051    0.08149   4.056 0.000328 ***
## METHODB:AGE   0.19318    0.11660   1.657 0.108001
## METHODC:AGE   0.70288    0.10896   6.451 3.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.925 on 30 degrees of freedom
## Multiple R-squared:  0.9143, Adjusted R-squared:  0.9001
## F-statistic: 64.04 on 5 and 30 DF,  p-value: 4.264e-15
```

Based on the summary above, it appears as though the interaction of Method and Age is significant (P-value less than 0.05 for Method C and Age).

The suggested model would then become:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{3i} X_{2i} + \beta_5 X_{3i} X_{1i} + \epsilon \\
\beta_0 &= 47.51 \\
\beta_1 &= -18.60 \\
\beta_2 &= -41.30 \\
\beta_3 &= 0.33 \\
\beta_4 &= 0.70 \\
\beta_5 &= 0.19
\end{aligned}
$$

$$
Effect_i = \begin{cases} 47.5 + 0.33 * Age & \text{if } i^{th} \text{ person receives treatment A} \\ 28.9 + 0.52 * Age & \text{if } i^{th} \text{ person receives treatment B} \\ 6.2 + 1.03 * Age & \text{if } i^{th} \text{ person receives treatment C} \end{cases}
$$

e

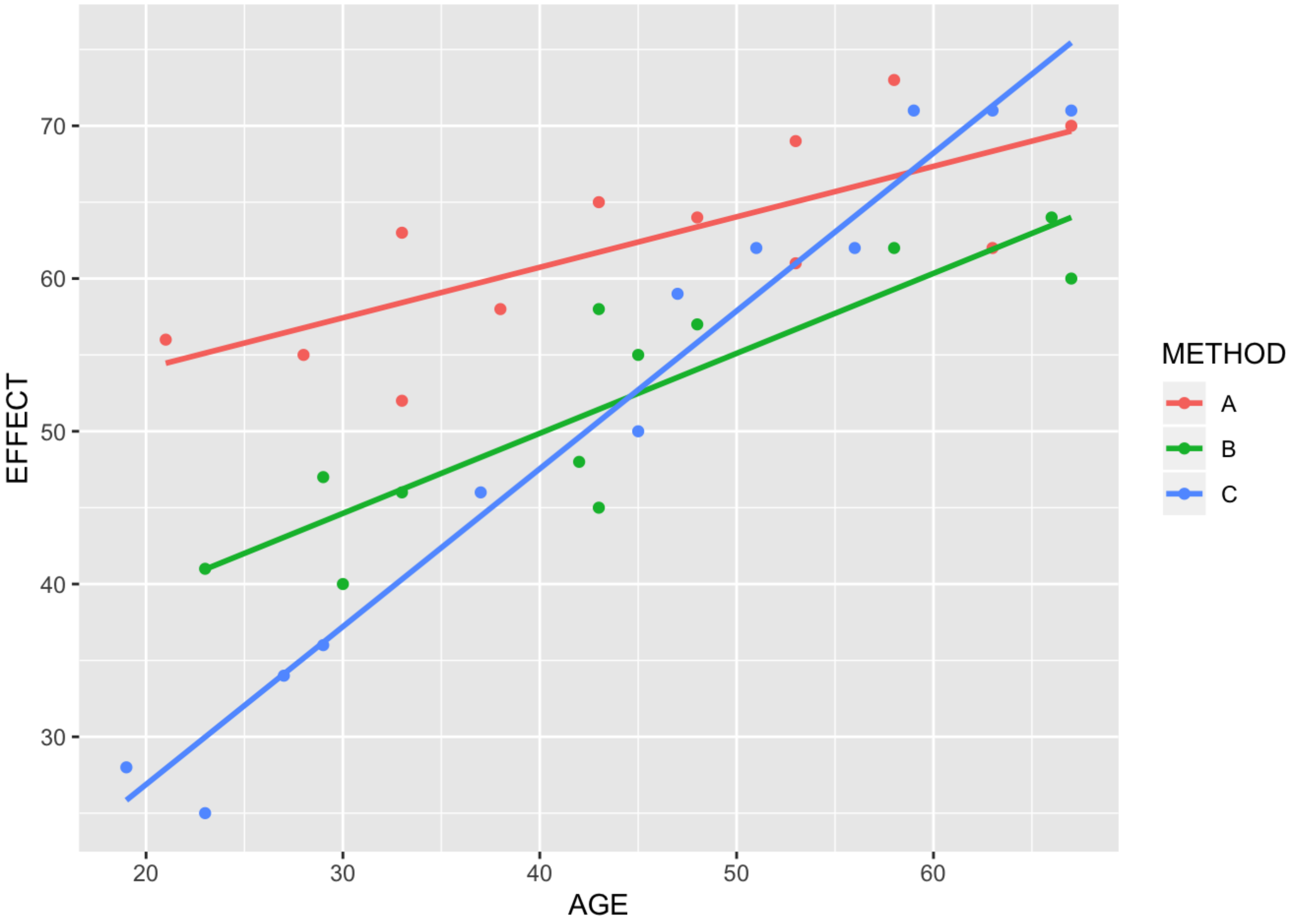
How would you interpret the effect of treatment?

At lower ages, treatment A out performs treatment B and treatment C in terms of effectiveness. As age increases, the effect of treatment C approaches and eventually exceeds the effectiveness of both treatment A and treatment B. Additionally, at higher ages, the effect of treatment B would eventually out perform treatment A as the slope on the linear regression equation is higher on treatment B than treatment A.

f

Plot the three regression lines on the scatter diagram obtained in c. May one have the same conclusion as in question d.?

```
health %>%
  ggplot(aes(x = AGE, y = EFFECT, col = METHOD)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)
```



Based on the differing slope of the regression lines, we can confirm that there is an interaction between age and treatment as the slope of the blue line is much more significant than the green and red lines.

Problem 6

Problem 6 [Optional]. Erecting boiler drums In a production facility, an accurate estimate of hours needed to complete a task is crucial to management in making such decisions as the proper number of workers to hire, an accurate deadline to quote a client, or cost-analysis decisions regarding budgets. A manufacturer of boiler drums wants to use regression to predict the number of hours needed to erect the drums in future projects. To accomplish this, data for 35 boilers were collected. In addition to hours (y), the variables measured were boiler capacity ($x1$ =lb/hr), boiler design pressure ($x2$ =pounds per square inch, or psi), boiler type ($x3$ =1 if industry field erected, 0 if utility field erected), and drum type (x4 =1 if steam, 0 if mud).The data are saved in the BOILERS.csv file.

a

Write the first order model for hours.

```
boilers <- boilers %>%
  mutate(Boiler = relevel(Boiler, ref = "utility"))
contrasts(boilers$Boiler)
```

```
##          industry
## utility         0
## industry        1
```

```
contrasts(boilers$Drum)
```

```
##       steam
## mud       0
## steam     1
```

```
boilers_full <- lm(data = boilers, Manhrs ~ Capacity + Pressure + Boiler + Drum)
coefficients(boilers_full)
```

```
##    (Intercept)       Capacity       Pressure Boilerindustry      Drumsteam
##   -2021.253808       7.576243       1.052940    2401.490997    1971.480481
```

$$\widehat{Hours} = -2021 + 7.576 * Capacity + 1.05 * Pressure + 2401 * Boiler + 1971 * Drum$$

Or for Boiler = industry (1) and Drum = steam (1):

$$\widehat{Hours} = 2352 + 7.576 * Capacity + 1.05 * Pressure$$

For Boiler = utility (0) and Drum = steam (1):

$$\widehat{Hours} = -49.77 + 7.576 * Capacity + 1.05 * Pressure$$

For Boiler = industry (1) and Drum = mud (0)

$$\widehat{Hours} = 380.2 + 7.576 * Capacity + 1.05 * Pressure$$

For Boiler = utility (0) and Drum = mud (0):

$$\widehat{Hours} = -2021 + 7.576 * Capacity + 1.05 * Pressure$$

b

Construct the Anova table for the first order model (the additive model).

```
boilers_intercept <- lm(data = boilers, Manhrs ~ 1)
anova(boilers_intercept, boilers_full)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	34	133403056	NA	NA	NA	NA
2	30	12709156	4	120693900	71.22458	7.042914e-15

2 rows

```
# Create anova table
header <- c("Source of Variation", "Df", "Sum of Squares", "Mean Squares", "F-Statistic")
anova_table_6 <- data.frame(rbind(c("Regression", p_6, SSR_6, MSR_6, F_6),
                                  c("Residual", n_6 - p_6 - 1, SSE_6, MSE_6, ""),
                                  c("Total", n_6 - 1, SSR_6 + SSE_6, "", "")))
names(anova_table_6) <- header
anova_table_6
```

Source of Variation <fctr>	Df <fctr>	Sum of Squares <fctr>	Mean Squares <fctr>	F-Statistic <fctr>
Regression	4	120693900	30173475	71.2245762031719
Residual	30	12709156	423638.533333333	
Total	34	133403056		

3 rows

c

Use the Anova table from part b to conduct a test for the full model (Use $\alpha = 0.01$).

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_A \text{ : at least one } \beta_i \text{ is not zero}$$

```
anova(boilers_intercept, boilers_full)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	34	133403056	NA	NA	NA	NA
2	30	12709156	4	120693900	71.22458	7.042914e-15

2 rows

The P-value from the anova function above is less than 0.01. The full model is therefore significant and we can reject H_0.

d

Would you drop any predictors out of the full model? Explain.

Using the individual t-test:

```
summary(boilers_full)
```

```
## 
## Call:
## lm(formula = Manhrs ~ Capacity + Pressure + Boiler + Drum, data = boilers)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1408.45  -423.07    52.93   340.24  1454.85 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -2021.2538   924.6578  -2.186  0.03676 *  
## Capacity           7.5762     0.6981  10.853 6.58e-12 ***
## Pressure           1.0529     0.4560   2.309  0.02800 *  
## Boilerindustry  2401.4910   691.7933   3.471  0.00159 ** 
## Drumsteam       1971.4805   225.4166   8.746 9.41e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 650.9 on 30 degrees of freedom
## Multiple R-squared:  0.9047, Adjusted R-squared:  0.892 
## F-statistic: 71.22 on 4 and 30 DF,  p-value: 7.043e-15
```

Each of the predictors is significant as each P-value is below 0.05. Therefore, we should not drop any predictors.

e

Test individually the interaction terms $\alpha = 0.05$. *What model would you suggest to use for predicting y? Explain.*

```
boilers_interact <- lm(data = boilers, Manhrs ~ (Capacity + Pressure + Boiler + Drum)**2)
summary(boilers_interact)
```

```
##
## Call:
## lm(formula = Manhrs ~ (Capacity + Pressure + Boiler + Drum)^2,
##     data = boilers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -909.00 -302.63   13.51  313.22  920.88
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               9.684e+05  7.275e+05   1.331   0.1956
## Capacity                  2.395e+03  1.745e+03   1.373   0.1825
## Pressure                 -1.602e+03  1.188e+03  -1.348   0.1903
## Boilerindustry           -9.683e+05  7.278e+05  -1.330   0.1959
## Drumsteam                 1.575e+03  1.734e+03   0.908   0.3728
## Capacity:Pressure        -2.510e-02  2.105e-02  -1.193   0.2447
## Capacity:Boilerindustry  -2.380e+03  1.741e+03  -1.367   0.1843
## Capacity:Drumsteam        3.278e+00  1.297e+00   2.528   0.0185 *
## Pressure:Boilerindustry   1.605e+03  1.189e+03   1.350   0.1897
## Pressure:Drumsteam       -3.800e-02  8.433e-01  -0.045   0.9644
## Boilerindustry:Drumsteam -4.954e+02  1.236e+03  -0.401   0.6921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 521.3 on 24 degrees of freedom
## Multiple R-squared:  0.9511, Adjusted R-squared:  0.9307
## F-statistic:  46.7 on 10 and 24 DF,  p-value: 2.817e-13
```

From the summary above, it appears as though the interaction between Capacity and Drum type is the only significant interaction. Therefore, we will include it in the final model without any other interaction.

```
boilers_interact_reduced <- lm(data = boilers, Manhrs ~ Capacity + Pressure + Boiler + Drum + Capacity*Drum)
summary(boilers_interact_reduced)
```

```
##
## Call:
## lm(formula = Manhrs ~ Capacity + Pressure + Boiler + Drum + Capacity *
##     Drum, data = boilers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1178.8  -356.9   105.8   271.0   927.2
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -752.9769   777.9738  -0.968 0.341113
## Capacity              5.1603     0.7695   6.706 2.34e-07 ***
## Pressure              0.7892     0.3620   2.180 0.037501 *
## Boilerindustry     1901.5687   553.2722   3.437 0.001798 **
## Drumsteam          1054.0013   271.0174   3.889 0.000540 ***
## Capacity:Drumsteam    3.3546     0.7518   4.462 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 509.8 on 29 degrees of freedom
## Multiple R-squared:  0.9435, Adjusted R-squared:  0.9338
## F-statistic: 96.87 on 5 and 29 DF,  p-value: < 2.2e-16
```

```
coefficients(boilers_interact_reduced)
```

```
##        (Intercept)           Capacity           Pressure
##       -752.9769314          5.1602655          0.7891761
##     Boilerindustry          Drumsteam Capacity:Drumsteam
##       1901.5686971       1054.0013388          3.3545588
```

The final model to predict hours becomes:

$$\widehat{Hours} = -753 + 5.16 * Capacity + 0.789 * Pressure + 1902 * Boiler + 1054 * Drum + 3.35 * Capacity * Drum$$

f

Write all possible submodels for two categorical variables (do not have to substitute values of β_i)

For Boiler = industry (1) and Drum = steam (1):

$$\widehat{Hours} = -753 + 5.16 * Capacity + 0.789 * Pressure + 1902 + 1054 + 3.35 * Capacity$$

For Boiler = utility (0) and Drum = steam (1):

$$\widehat{Hours} = -753 + 5.16 * Capacity + 0.789 * Pressure + 1054 + 3.35 * Capacity$$

For Boiler = industry (1) and Drum = mud (0)

$$\widehat{Hours} = -753 + 5.16 * Capacity + 0.789 * Pressure + 1902$$

For Boiler = utility (0) and Drum = mud (0):

$$\widehat{Hours} = -753 + 5.16 * Capacity + 0.789 * Pressure$$