PROBLEM 1

a. Multicollinearity occurs when two or more variables are highly correlated. This causes the standard errors of the coefficients to be inflated. This makes them sensitive to even the slightest changes in the dataset. Multicollinearity also makes it difficult to figure out which individual predictors have the most impact on the response variable.

b. Assessing the VIF values of the coefficients can give insight into multicollinearity, as VIF measures the inflation in variance of a coefficient due to multicollinearity. A VIF greater than 10 may indicate a problematic case of multicollinearity. Additionally, creating a correlation matrix can be wise. A high pairwise correlation between predictors can also indicate multicollinearity.

c. Removing predictors which are not very impactful can help combat multicollinearity. Combining variables under specific conditions can also be helpful. Lastly, using techniques such as LASSO can be key as they are less sensitive to multicollinearity.

PROBLEM 2

a. Causation does not imply correlation, regardless of r = 0.983, which suggests a strong positive linear association. This is due to the fact that there might be a hidden confounding variable which is not being taken into account in this observational data. Other factors such as industry trends or experience levels can influence both variables, which aren't been taken into account as well. In order to establish the correlation, further studies and experiments need to be done.

b. A r = 0.722 suggests a positive linear relationship between males with a college degree and males in managerial positions. In a regression analysis, it can create a multicollinearity issue if both of these are included as predictors. As mentioned in the Problem 1 answer, creating a correlation matrix or assessing the VIF values will be key in this case.

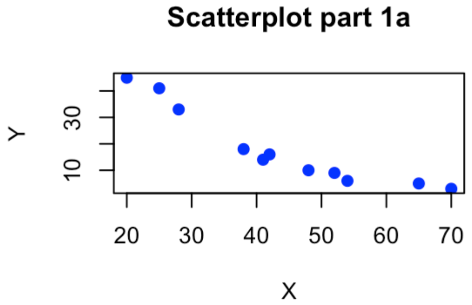
PROBLEM 3

No, I do not agree with this statement. Firstly, the R^2 value is very high (0.93), indicating that the model can explain 93% of the variation in y. This is a very good evidence of how x1, x2 and x3 are contributing strongly to predict carbohydrate amounts.

The statement might have occurred after looking at the individual t-values for the three predictors:
X1: t = 3.2/2.4 = 1.33
X2: t = -0.4/0.6 = -0.67
X3: t = -1.1/0.8 = -1.375
These are small enough to conclude that the predictors are not significant; however, looking at multicollinearity might be a good next step here, especially given the high R^2 value. Multicollinearity might be making the t-test unreliable.

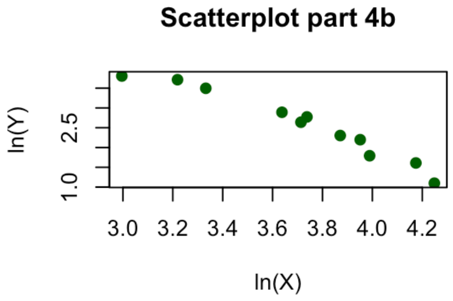
To conclude, looking at the R^2 value gives us direct proof of how the predictors are clearly useful for the model.

PROBLEM 4



The plot shows a downward inverse relationship between x and y.

b.



This plot shows a moderately negative linear relationship between ln(y) and ln(x).

c.

Based on the output from RStudio:

```
> summary(model)

Call:
lm(formula = lnY ~ lnX, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.32942 -0.07912  0.06168  0.11249  0.24640 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.6364     0.6028   17.64 2.73e-08 ***
lnX          -2.1699     0.1614  -13.44 2.91e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2021 on 9 degrees of freedom
Multiple R-squared:  0.9526,	Adjusted R-squared:  0.9473 
F-statistic: 180.7 on 1 and 9 DF,  p-value: 2.911e-07

>
```

- This model summary shows us that the p-value for lnX = -2.1699, which is smaller than 0.05; hence, it tells us that ln(x) is a significant predictor of ln(y).
- The R^2 value is 0.9473, which is very high and means that the model can explain 94.73% of the variation in ln(y) based on ln(x).

PROBLEM 5

a. E(y)= β0 +β1 x+β2*x^2

- As this equation has 3 coefficients, the model needs 3 distinct levels.
- At least 4 observations are needed to have sufficient degrees of freedom for estimating σ^2. Three for the parameters and one used for estimating σ^2.

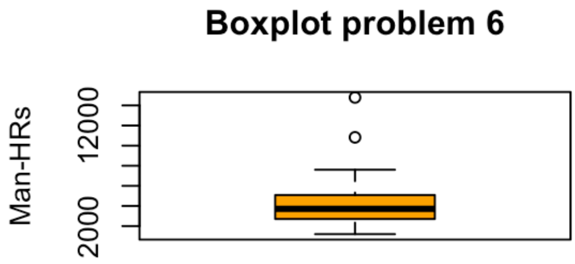
b. E(y)= β0 +β1*x1 +β2*x2 +β3*x1*x2

- This equation has 4 coefficients, the model needs at least 2 levels for both x1 and x2 respectively.
- The sample size needs to be at least 5, where 4 points can be used for coefficient estimation and the last point can estimate σ^2.

c. E(y)=β0 +β1*x1 +β2*x2 +β3*x1*x2 +β4*x1^2 +β5*x2^2

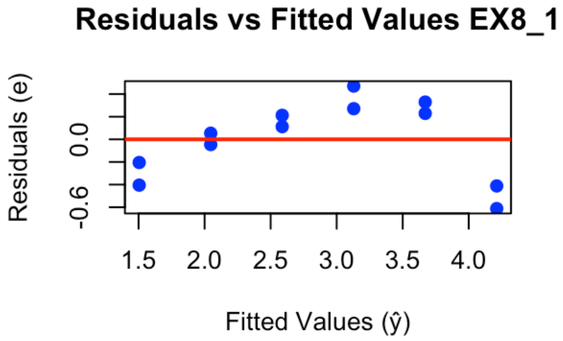
- This equation has 6 coefficients, and the model needs at least 3 levels for both x1 and x2 respectively.
- The sample size needs to be at least 7, where 6 can be used for coefficient estimation and the last point can estimate σ^2.

PROBLEM 6

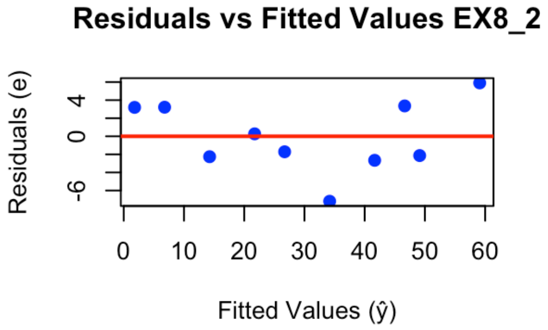


This boxplot, obtained from the BOILERS dataset, suggests that the data is skewed positively. This can be seen by how the median is closer to the bottom of the box.
There seem to be two obvious outliers towards the top of the box, which suggests non-normality.

PROBLEM 7



This shows a trend which is not random, as the line curves downwards. This means that the first-order model is inadequate.

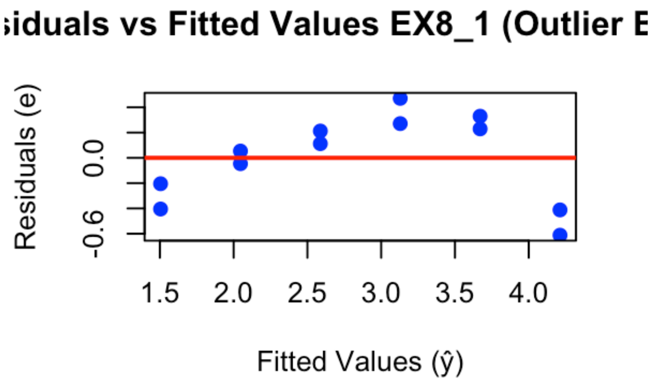


This shows a random pattern about the line x = 0, which suggests that the first-order model is adequate and fits the data well.

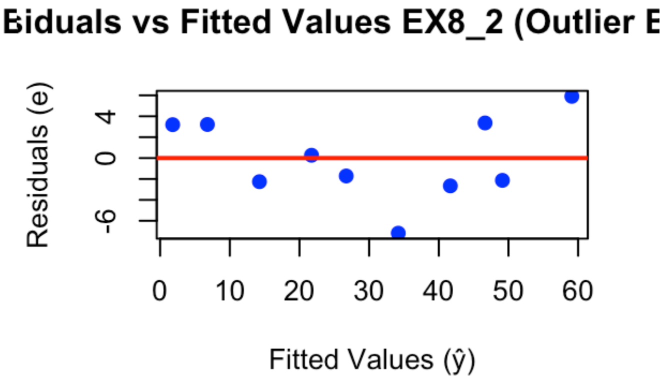
PROBLEM 8

1. Using Leverage helps measure how far an observation's predictor value is from the mean of all predictor values.
2. Standardised Residuals can help measure how far the predicted value is from the actual value of yi. These are scaled in terms of Standard Deviations, and hence, usually, it is assumed that values with large residuals may be outliers.
3. Cook's Distance uses both leverage and Residuals to measure the influence an observation has on the entire regression model. If we remove observations with high cooks' distances would change the model estimates significantly.

No. These three methods will flag different observations based on what they try to assess and how they work. They will not select the same observations.



Based on this graph, you can see that there are no outliers for the dataset EX8_1



Based on this graph, you can see that there are no outliers for the dataset EX8_2

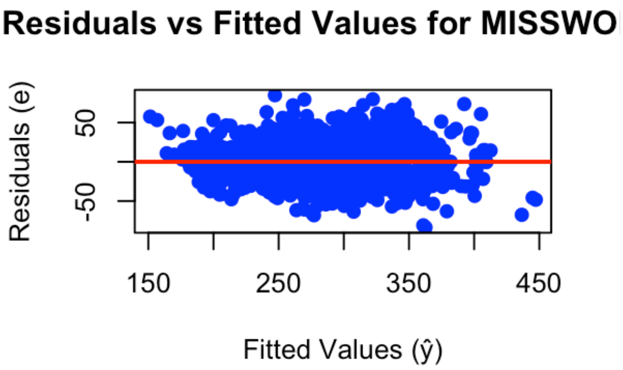
PROBLEM 9

a.

```
> summary(model)

Call:
lm(formula = y ~ x, data = misswork)

Residuals:
    Min      1Q  Median      3Q     Max 
-83.681 -17.308  -0.469  15.791  84.861 

Coefficients: (1 not defined because of singularities)
                                         Estimate Std. Error t value
(Intercept)                             4.594e+02  6.745e+01   6.811
xecon__pct_civilian_labor              -1.005e+02  2.603e+01  -3.860
xecon__pct_unemployment                 6.198e+01  5.636e+01   1.100
xecon__pct_uninsured_adults             6.762e+01  2.963e+01   2.282
xecon__pct_uninsured_children          -2.816e+01  4.865e+01  -0.579
xdemo__pct_female                       5.810e+01  6.607e+01   0.879
xdemo__pct_below_18_years_of_age        1.460e+02  4.694e+01   3.111
xdemo__pct_aged_65_years_and_older     -6.741e+02  5.306e+01 -12.704
xdemo__pct_hispanic                    -3.500e+02  5.937e+01  -5.895
xdemo__pct_non_hispanic_african_american -3.145e+02  6.044e+01  -5.203
xdemo__pct_non_hispanic_white          -3.166e+02  6.006e+01  -5.271
xdemo__pct_american_indian_or_alaskan_native -4.534e+02  6.203e+01  -7.309
```

```
xdemo__pct_asian                              -3.112e+02  7.064e+01  -4.406
xdemo__pct_adults_less_than_a_high_school_diploma  2.736e+01  3.318e+01   0.825
xdemo__pct_adults_with_high_school_diploma        4.443e+01  2.505e+01   1.773
xdemo__pct_adults_with_some_college              -2.446e+01  2.717e+01  -0.900
xdemo__pct_adults_bachelors_or_higher                    NA         NA      NA
xdemo__birth_rate_per_1k                      -2.638e-01  5.855e-01  -0.451
xdemo__death_rate_per_1k                       1.215e+01  8.859e-01  13.710
xhealth__pct_adult_obesity                    -1.355e+01  3.555e+01  -0.381
xhealth__pct_adult_smoking                     5.887e+01  2.670e+01   2.205
xhealth__pct_diabetes                         6.648e+00  7.629e+01   0.087
xhealth__pct_low_birthweight                   9.290e+01  7.782e+01   1.194
xhealth__pct_excessive_drinking                2.521e+01  2.595e+01   0.971
xhealth__pct_physical_inacticity               3.153e+02  3.348e+01   9.420
xhealth__air_pollution_particulate_matter      -2.684e+00  7.263e-01  -3.695
xhealth__homicides_per_100k                    2.281e-01  2.686e-01   0.849
xhealth__motor_vehicle_crash_deaths_per_100k     6.694e-01  1.679e-01   3.987
xhealth__pop_per_dentist                      -5.818e-04  5.638e-04  -1.032
xhealth__pop_per_primary_care_physician         -1.320e-03  7.769e-04  -1.699
                                             Pr(>|t|)
(Intercept)                                   1.64e-11 ***
xecon__pct_civilian_labor                     0.000121 ***
xecon__pct_unemployment                        0.271730
xecon__pct_uninsured_adults                    0.022694 *
xecon__pct_uninsured_children                  0.562863
xdemo__pct_female                             0.379418
xdemo__pct_below_18_years_of_age               0.001915 **
xdemo__pct_aged_65_years_and_older              < 2e-16 ***
xdemo__pct_hispanic                           5.05e-09 ***
xdemo__pct_non_hispanic_african_american       2.36e-07 ***
xdemo__pct_non_hispanic_white                  1.65e-07 ***
xdemo__pct_american_indian_or_alaskan_native     5.35e-13 ***
xdemo__pct_asian                              1.16e-05 ***
xdemo__pct_adults_less_than_a_high_school_diploma 0.409804
xdemo__pct_adults_with_high_school_diploma       0.076453 .
xdemo__pct_adults_with_some_college             0.368244
xdemo__pct_adults_bachelors_or_higher                NA
xdemo__birth_rate_per_1k                       0.652332
xdemo__death_rate_per_1k                        < 2e-16 ***
xhealth__pct_adult_obesity                    0.703105
xhealth__pct_adult_smoking                     0.027640 *
xhealth__pct_diabetes                         0.930583
xhealth__pct_low_birthweight                   0.232841
xhealth__pct_excessive_drinking                0.331583
xhealth__pct_physical_inacticity                < 2e-16 ***
xhealth__air_pollution_particulate_matter       0.000231 ***
xhealth__homicides_per_100k                    0.396026
xhealth__motor_vehicle_crash_deaths_per_100k    7.15e-05 ***
xhealth__pop_per_dentist                      0.302355
xhealth__pop_per_primary_care_physician         0.089629 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.89 on 1040 degrees of freedom
Multiple R-squared:  0.7931,  Adjusted R-squared:  0.7875
F-statistic: 142.3 on 28 and 1040 DF,  p-value: < 2.2e-16

>
```

b.



Based on the random pattern and no obvious trend being present, we can conclude that the first-order linear model is adequate for modelling the relationship between annual wages and missed hours at work.

c. It would be okay to delete this observation from the dataset entirely before conducting the regression analysis. This is due to the fact that, given the context, we know this outlier was an error which would not have happened under natural circumstances.

PROBLEM 10

a. $y=\beta0 +\beta1*x+\beta2 (x-1.45)_+ +\beta3 (x-5.20)_+ +\varepsilon$

b.

For x <=1.45:
Model: y=β0 +β1*x as x1 and x2 = 0 where the β0 is the intercept and β1 is the slope

For 1.45 < x <= 5.20:
Model: y=(β0 –β2*1.45)+(β1 +β2)*x where the Intercept is β0 –1.45β2 and the slope is β1 +β2

For x > 5.20:
y=β0 –1.45*β2 –5.20*β3 +(β1 +β2 +β3)*x where the intercept is β0 –1.45β2 –5.20β3 and the slope is β1 +β2 +β3

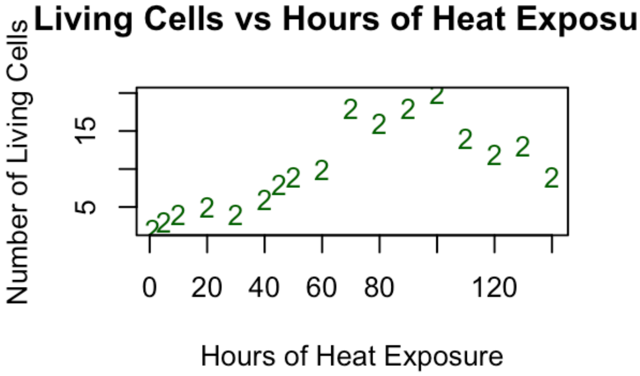
c.

In this case
H0: β2 and β3 both equal to 0
H1: At least one of the slopes (β2, β3) is not equal to 0.

Then a t-test can be performed on β2, β3, where if the p-value <0.05, then there is a change in the slope.

PROBLEM 11

a.



Based on this graph, there seems to be a quadratic relationship between hours and cells. Here, the peak was reached between 80-90 hours.

b.

$y=\beta 0 +\beta 1 * x+\beta 2 \, (x-85)_{+}+\varepsilon$

For the knot value: 90+80/2 = 85 = x

c.

```
> summary(model_piecewise)

Call:
lm(formula = CELLS ~ HOURS + knot, data = growth)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6845 -1.6074 -0.4342  1.4189  4.2142 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.37593    1.12864   0.333    0.744    
HOURS        0.20514    0.02192   9.359 2.11e-07 ***
knot        -0.34030    0.05435  -6.261 2.09e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.196 on 14 degrees of freedom
Multiple R-squared:  0.8714,	Adjusted R-squared:  0.8531 
F-statistic: 47.45 on 2 and 14 DF,  p-value: 5.806e-07

>
```

- The R^2 value of 0.8531 tells us that the model can explain 85.31% of the variation in cell count, which is very high.
- The p-values for HOURS and knot indicate they are statistically significant.

d.

```
> cat("F-statistic:", f_stat_value, "\n")
F-statistic: 47.44714 
> cat("Degrees of freedom:", df1, "and", df2, "\n")
Degrees of freedom: 2 and 14 
> cat("p-value:", p_value, "\n")
p-value: 5.805783e-07 
>
```

- The p-value is less than 0.05, hence the null hypothesis is rejected that both slope coefficients are zero (i.e one of these predictors contributes strongly to explain the variation in cell count).
- This also indicated that the model is statistically sound to fit both datasets.

e.

Estimated rate of growth for x > 85 hours: -0.1352 cells/hour

```
> cat("Rate of growth (x < 70 hours):", round(rate_less_than_70, 4), "cells/hour\n")
Rate of growth (x < 70 hours): 0.2051 cells/hour
> cat("Rate of growth (x > 85 hours):", round(rate_more_than_85, 4), "cells/hour\n")
Rate of growth (x > 85 hours): -0.1352 cells/hour
```

It is for less than 85, as that is the knot value, and after which the cells decrease.

f.

Here the
H0: There is no change in slope after 85h
HA: The slope changes after. 85h

```
> knot_pval <- summary(model_piecewise)$coefficients["knot", "Pr(>|t|)"]
> cat("p-value for test of slope change after 85 hours:", knot_pval, "\n")
p-value for test of slope change after 85 hours: 2.086939e-05
> if (knot_pval < 0.05) {
+   cat("Reject H0\n")
+ } else {
+   cat("Fail to reject H0\n")
+ }
Reject H0
>
```

This output from R Code helps us reject the H0 and helps conclude that the two slopes before and after 85 are statistically different.