Name: Chris Jerylle Vargas Oidem
Student ID: 45476624
Unit Code: STAT270 Applied Statistics

<u>Question 1</u>

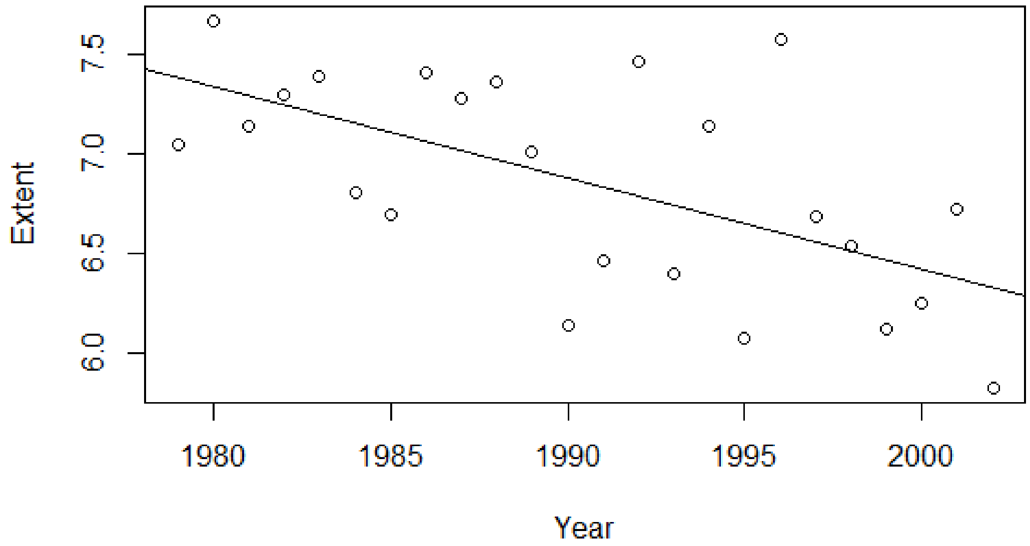
a) **State the statistical model for a simple linear regression of Extent explained by Year.**

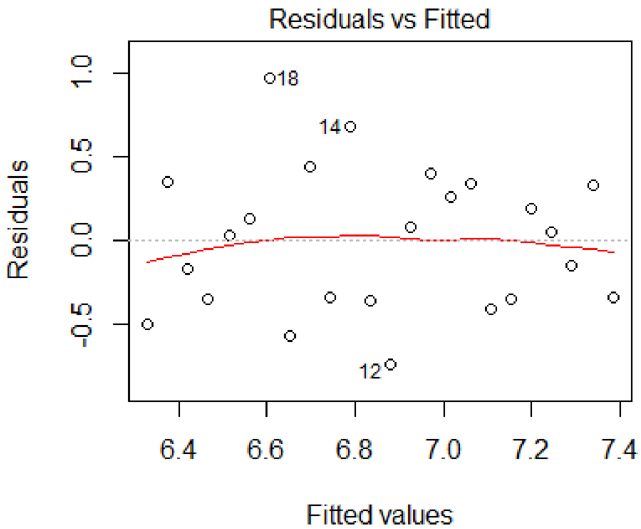
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

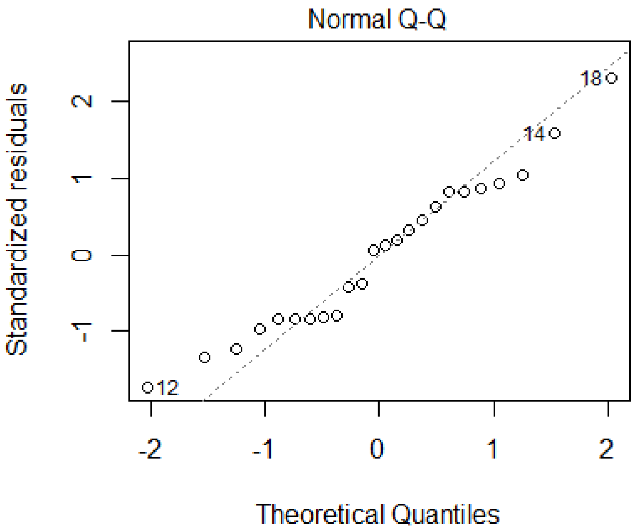
$Extent_i = \beta_0 + \beta_1 Year_i + \varepsilon_i$; $\varepsilon_i \sim N(0,\sigma^2)$

b) **Fit a simple linear regression to the 1979-2002 data. Explain why there is a linear relationship.**

Based on the scatterplot below, it is reasonable to assume a linear relationship as the data looks like it follows a linear trend. Q-Q plot verifies normal residuals as it follows a relatively straight line and the residual vs fitted plot confirms constant variance in residuals, deeming the linear regression model appropriate.







c)

```
Call:
lm(formula = Extent ~ Year, data = dat1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.74002 -0.34571  0.03998  0.33513  0.97518 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 98.15180   25.65735   3.825 0.000922 ***
Year        -0.04587    0.01289  -3.558 0.001760 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4371 on 22 degrees of freedom
Multiple R-squared:  0.3653,	Adjusted R-squared:  0.3364 
F-statistic: 12.66 on 1 and 22 DF,  p-value: 0.00176
```

There is a linear relationship as the slope is not zero and effect of year is statistically significant.

$\hat{Y}_i = b_0 + b_1X_i \rightarrow \widehat{\text{Extent}}_i = 98.1518 - 0.04587\text{Year}_i$

For every 1-year increase, the extent of the sea ice is expected to decrease by 0.04587km^2, on average.

d) Is this a strong relationship? Explain your answer in the context of this data.

Correlation Coefficient = 0.6044005

No, it is not a strong linear relationship. Based on the output in part c), only 60% of the variation in Extent is explained by the linear regression on Year.

e) Predict the extent of the sea ice (in km^2) for the year 2000.

X = 2000

$\hat{Y}_i = 98.1518 - 0.04587(2000)$

$= 6.4118$

≈ 6.412

The expected extent of the sea ice is 6.412km^2 for the year 2000.

f) 95% prediction band for the extent of sea ice in the year 2000:

6.412 ± 2.073873 x 0.4371 x 1.058369

= (5.452599, 7.371401)

```
RStudio Output:
#      lwr      upr
# 5.461929 7.380798
```

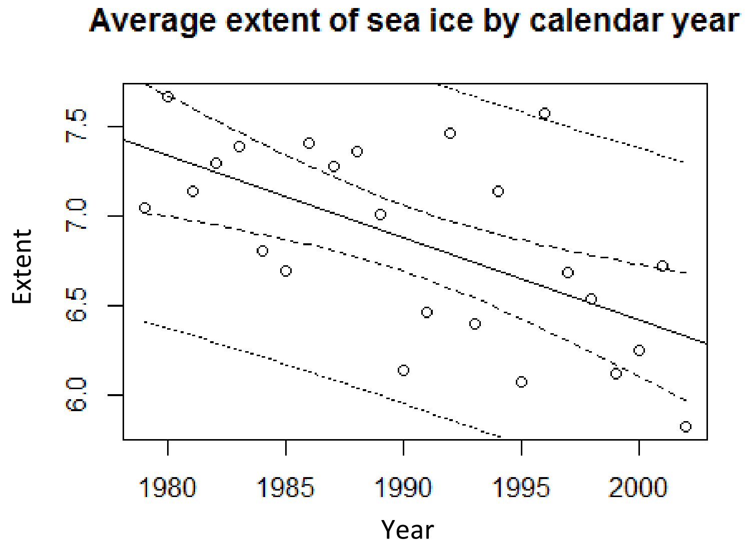
g) 95% confidence band for the extent of sea ice in the year 2000:

6.412 ± 2.073873 x 0.4371 x 0.3466192

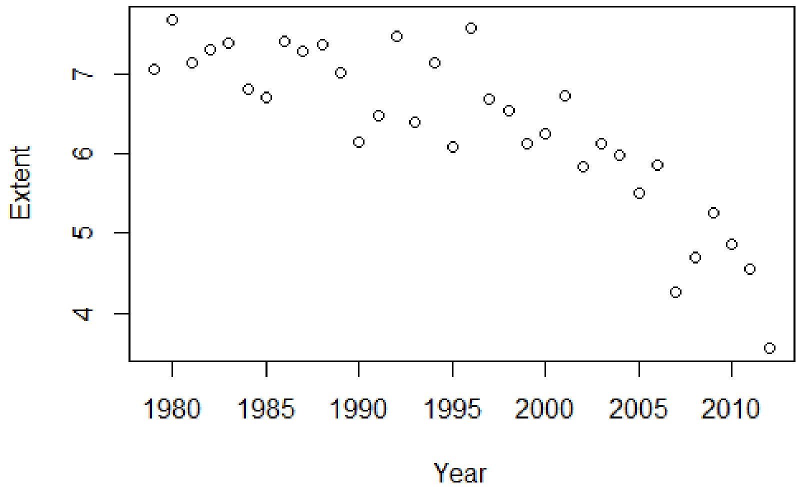
= (6.097793, 6.726207)

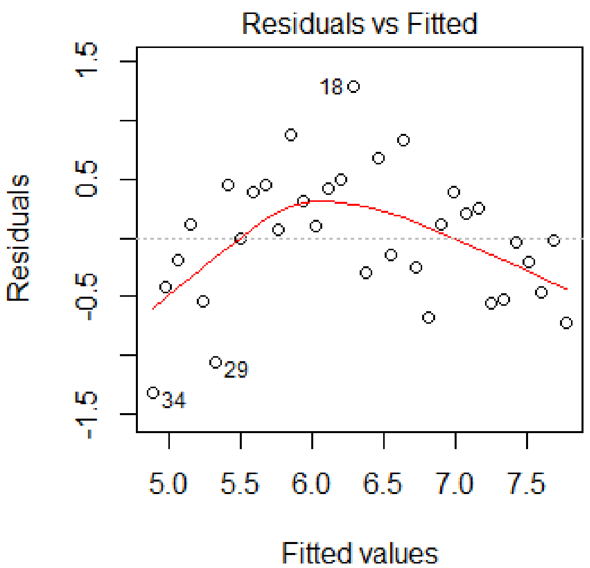
```
RStudio Output:
#      lwr      upr
# 6.107146 6.735582
```

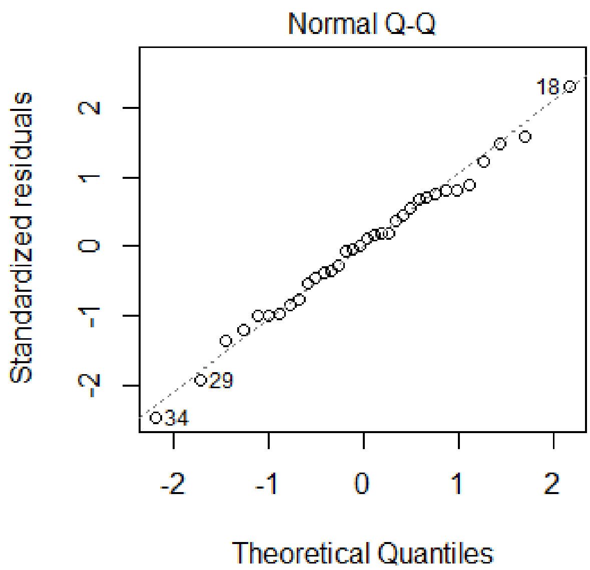
h) Given the data, we have a 95% prediction interval for the extent of sea ice in the year 2000 to be (5.452599, 7.371401) and a 95% confidence interval for the true mean at Year = 2000 to be (6.097793, 6.726207).



i) **Justify why a simple linear regression is inappropriate for the 1979-2012 data:**
There is no linear relationship between Year and Extent. A curvature pattern can be seen, thus a simple linear regression is inappropriate and will not fit. The residuals vs fitted plot shows a negative quadratic trend and parabola shape. We should fit a polynomial model and validate.







j) Fit a second order polynomial regression model to the data and validate the model:

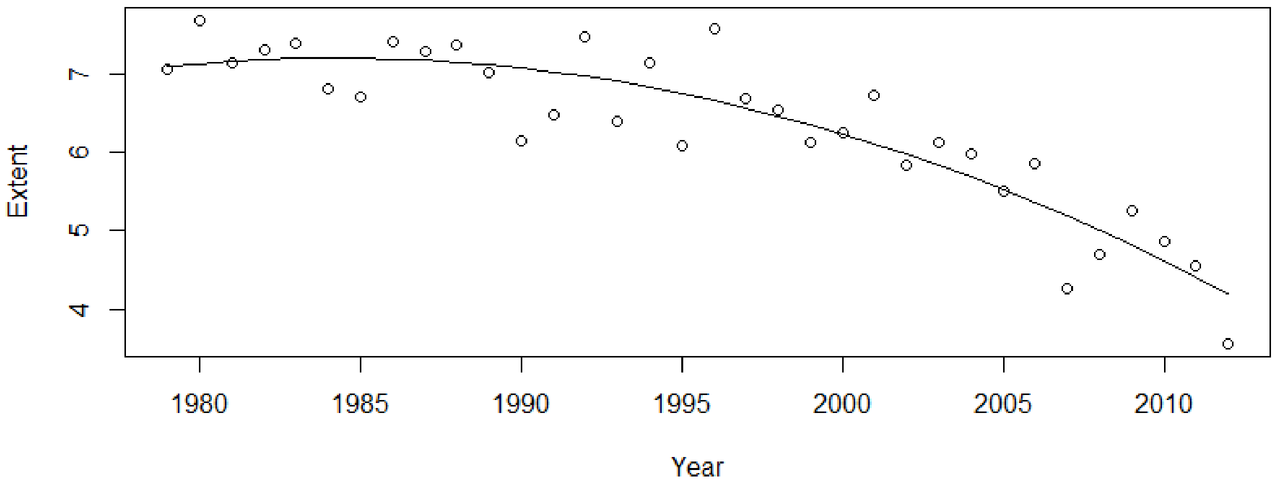
```
Call:
lm(formula = Extent ~ Year + I(Year^2), data = seaice)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9300 -0.2932  0.0938  0.2796  0.9173 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.540e+04  3.626e+03  -4.247 0.000183 ***
Year         1.553e+01  3.634e+00   4.273 0.000170 ***
I(Year^2)   -3.912e-03  9.105e-04  -4.297 0.000159 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4565 on 31 degrees of freedom
Multiple R-squared:  0.8171,	Adjusted R-squared:  0.8053 
F-statistic: 69.27 on 2 and 31 DF,  p-value: 3.653e-12
```

$\hat{Y}_i = b_0 + b_1X_i + b_2X_i^2$

$= -15{,}400 + 15.53X - 0.003912X^2$

k) Plot the fitted polynomial to your data:

$\widehat{Extent}_i = -15{,}400 + 15.53Year_i - 0.003912Year_i{}^2$



l) Using the second order model you fitted, predict the extent of the sea ice for the year 2000:

X = 2000

$\hat{Y}_i = -15{,}400 + 15.53(2000) - 0.003912(2000)^2$

$= 12$

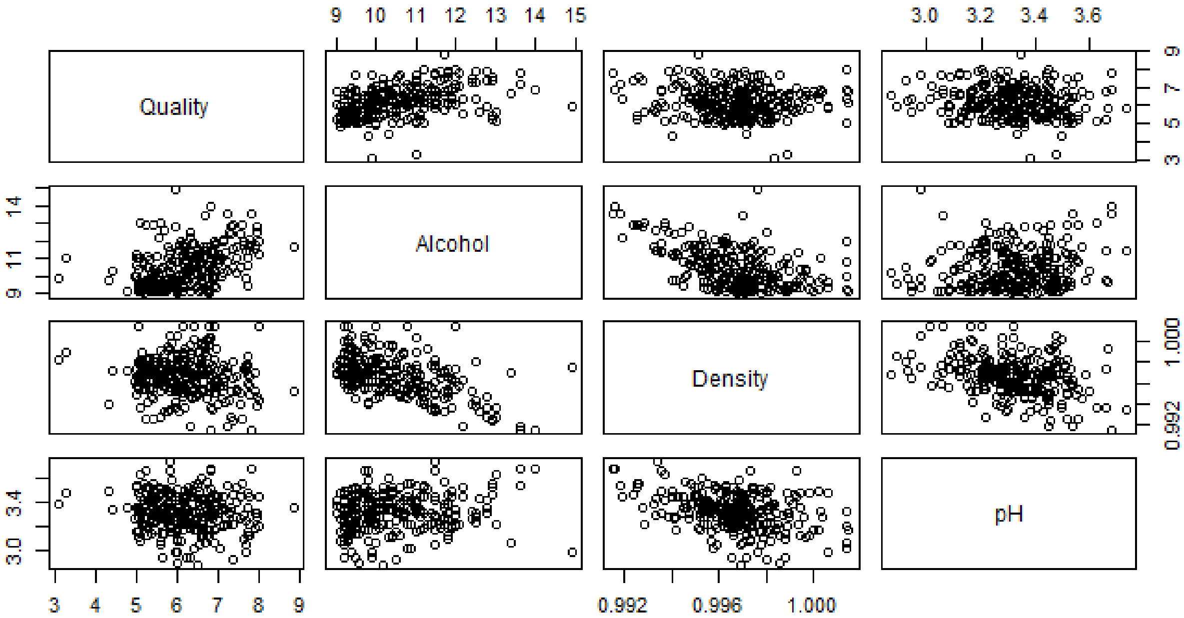
The expected extent of the sea ice is 12km^2 for the year 2000.

m) Compare your answers in part e) and part l). Which prediction value do you recommend and why?

I would recommend the prediction value in part e) as the value in part l) is a result of overfitting a model. The R-squared is also unusually high, which means that the model has begun to describe random error in the data rather than relationships between variables.

Question 2

a) **State the statistical model for a multiple regression with Quality as the response using all other variables as predictors, defining any parameters as necessary.**



$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i + \beta_3 X_i + \varepsilon_i$

$\text{Quality}_i = \beta_0 + \beta_1 \text{Density}_i + \beta_2 \text{pH}_i + \beta_3 \text{Alcohol}_i + \varepsilon_i \,;\ \varepsilon_i \sim N(0,\sigma^2)$

b) **Fit this multiple regression model and write down the fitted model.**

$\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i + b_3 X_i$

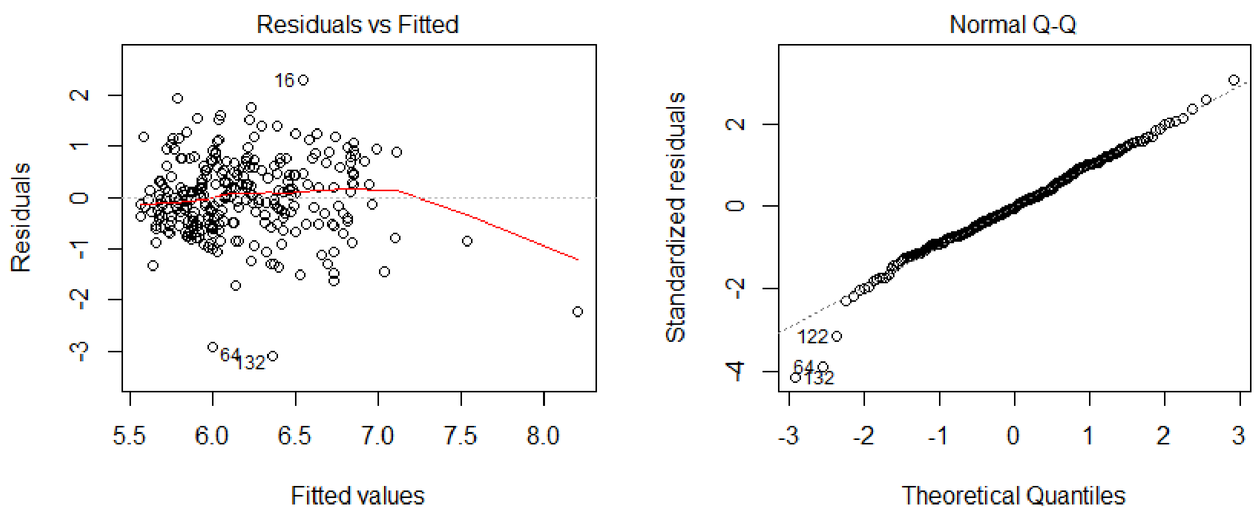
→ $\widehat{\text{Quality}}_i = -46.18943 + 51.33496\text{Density}_i - 0.84169\text{pH}_i + 0.38190\text{Alcohol}_i$

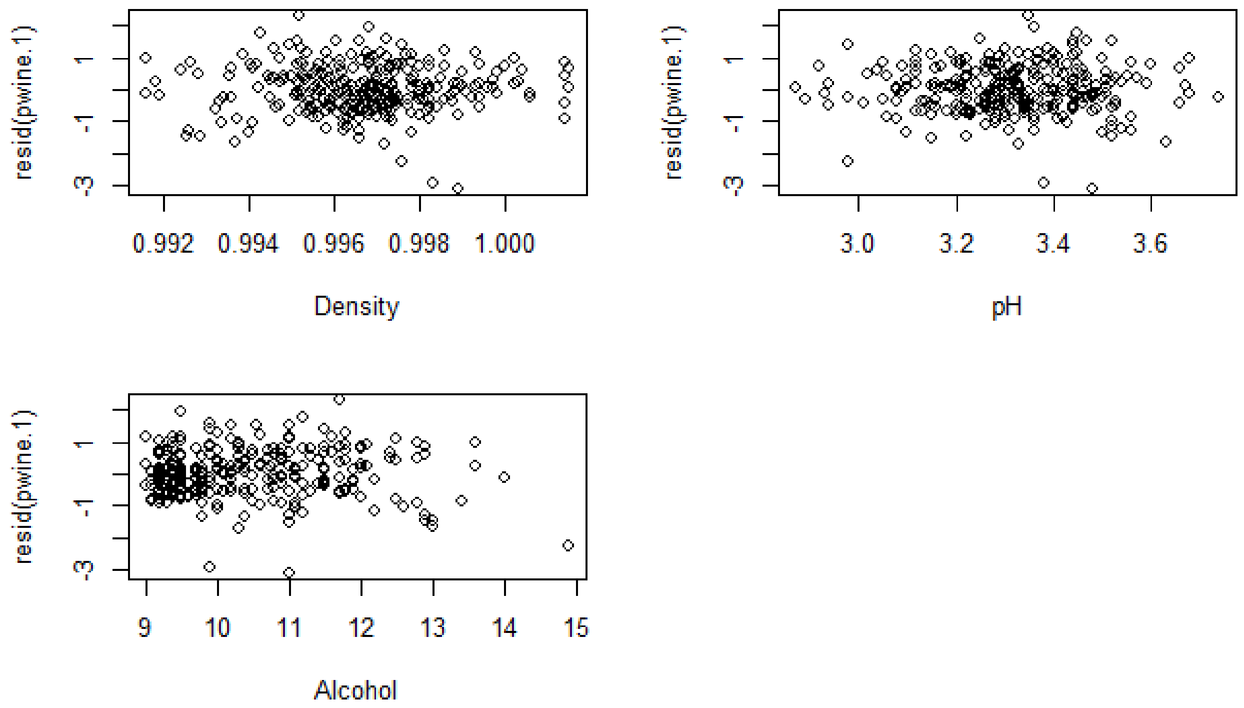
c) **What are the assumptions required for a multiple regression analysis? If possible, validate those assumptions for the multiple regression model you fitted in part b**.

$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon_i$

Where $\varepsilon_i \sim N(0,\sigma^2)$

- The normal Q-Q plot of residuals is linear, implying errors close to normally distributed and better adherence to model requirements.
- The residuals vs fitted has no discernable pattern besides the outlier dragging the slope down.
- Residuals vs predictor plots shows no obvious pattern.





d) Conduct an F-test for the overall regression i.e. is there any relationship between the response and the predictors. Write your answer as a formal hypothesis test and include the ANOVA table (one combined regression SS source is sufficient)

```
Analysis of Variance Table

Response: Quality
           Df  Sum Sq Mean Sq F value   Pr(>F)
Density     1   1.956   1.956  3.4791 0.063187 .
pH          1   5.503   5.503  9.7858 0.001943 **
Alcohol     1  36.772  36.772 65.3905 1.83e-14 ***
Residuals 282 158.583   0.562
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Full Regression SS = 1.956 + 5.503 + 36.772 = 44.231

Regression MS = 44.231/3 = 14.74367

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$; H_1: at least one β_i is **not** $= 0$;

Test statistic: F_{obs} = Regression MS/Residual MS = 14.74367/0.56235 = 26.21796

P-value: $P(F_{3,282} \geq 26.22) = 5.47e\text{-}15 < 0.01$

- Reject at the 5% level

  There is a significant linear relationship between the response and at least one of the three predictor variables.

e) From the analysis in part b, determine the 95% CI for the Alcohol slope parameter and comment on its meaning in this context.

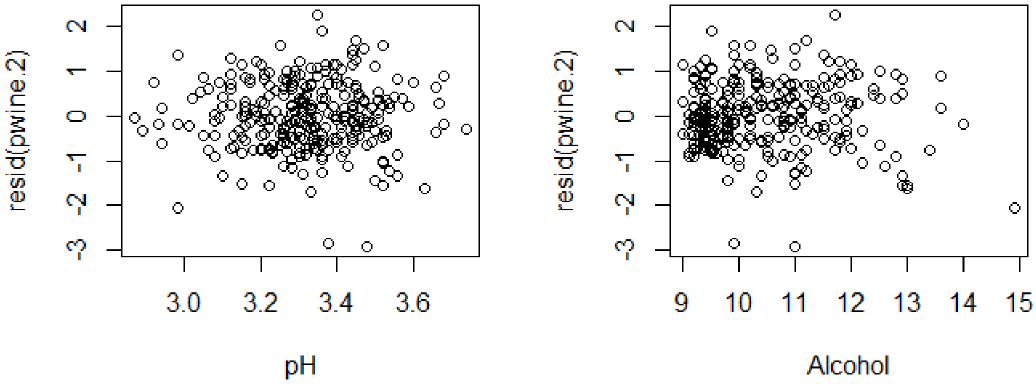
0.38190 ± 1.968 x 0.04723

= (0.2889514, 0.4748486)

It is found that 0.5 is contained in the interval between 0.289 and 0.475. Thus, we have data consistent with the claim of 0.38 increase in Quality for each 1% increase in Alcohol, on average.

f) Using the model selection procedures used in this course, find the best multiple regression model that explains the data giving reasons for your choice(s).

$\text{Quality}_i = \beta_0 + \beta_1 \text{pH}_i + \beta_2 \text{Alcohol}_i + \varepsilon_i$; $\varepsilon_i \sim N(0,\sigma^2)$



Density has the largest non-statistic p-value and does not contribute statistically significant and additional information to predicting the response after controlling for the other predictors in the model – pH and Alcohol. Thus, dropping it from the model.

g) State the final fitted regression model and comment on its interpretation.

$\widehat{\text{Quality}}_i = 6.0141 - 1.0260\text{pH}_i + 0.3409\text{Alcohol}_i$

For a unit increase in pH there is a 1.0260 decrease in Quality and for every unit increase in Alcohol, there is a 0.3409 increase in Quality. All of these are while holding all other parameters constant.