1. a) **Task:** Prove that $RSS = SYY - \frac{(SXY)^2}{SXX}$

$$\begin{aligned}
RSS &= \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2 \\
&= \sum_{i=1}^{n} \left[y_i - (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i) \right]^2 \\
&= \sum_{i=1}^{n} \left[y_i - \bar{y} + \frac{SXY}{SXX}\bar{x} - \frac{SXY}{SXX} x_i \right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \frac{SXY}{SXX}(x_i - \bar{x}) \right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \bar{y})^2 - 2\frac{SXY}{SXX}(y_i - \bar{y})(x_i - \bar{x}) + \left(\frac{SXY}{SXX}\right)^2 (x_i - \bar{x})^2 \right] \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\frac{SXY}{SXX} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + \left(\frac{SXY}{SXX}\right)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= SYY - 2\frac{SXY}{SXX} SXY + \left(\frac{SXY}{SXX}\right)^2 SXX \\
&= SYY - \frac{SXY^2}{SXX}
\end{aligned}$$

   b) **Task:** Use Cauchy-Schwarz to prove that RSS is nonnegative.

$$\begin{aligned}
RSS &= SYY - \frac{SXY^2}{SXX} \\
&= \|y - \bar{y}\|^2 - \frac{\langle x - \bar{x}, y - \bar{y} \rangle^2}{\|x - \bar{x}\|^2} \\
&\geq \|y - \bar{y}\|^2 - \frac{\|x - \bar{x}\|^2 \|y - \bar{y}\|^2}{\|x - \bar{x}\|^2} \\
&\geq \|y - \bar{y}\|^2 - \|y - \bar{y}\|^2 \\
&\geq 0
\end{aligned}$$

2. **Task:** Show that linear regression and k-nearest neighbor regression are both of the class $\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i$.
   Linear regression is a model of the form:

$$
\begin{aligned}
\hat{f}(x_0) &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\
&= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 \\
&= \bar{y} + (x_0 - \bar{x}) \frac{SXY}{SXX} \\
&= \bar{y} + (x_0 - \bar{x}) \frac{\sum_{i=1}^{N} (x_i - \bar{x}) y_i}{SXX} \\
&= \bar{y} + \sum_{i=1}^{N} \frac{(x_0 - \bar{x})(x_i - \bar{x})}{SXX} y_i \\
&= \sum_{i=1}^{N} \left[\frac{(x_0 - \bar{x})(x_i - \bar{x})}{SXX} y_i + \frac{\bar{y}}{N} \right] \\
&= \sum_{i=1}^{N} \left[\frac{(x_0 - \bar{x})(x_i - \bar{x})}{SXX} + \frac{1}{N} \right] y_i
\end{aligned}
$$

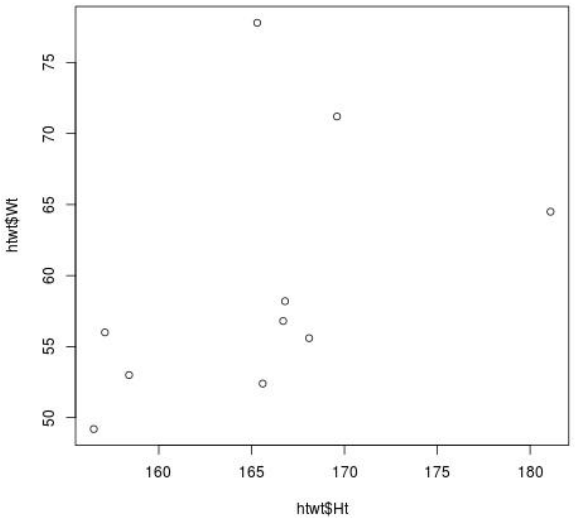
Accordingly, it is of the desired form for

$$
l_i(x_0; \mathcal{X}) = \frac{(x_0 - \bar{x})(x_i - \bar{x})}{SXX} + \frac{1}{N}
$$

Nearest neighbors regression is trivially of that form, with $N_{k,i}$ being the set of k x_j closest to x_0:

$$
l_i(x_0; \mathcal{X}) = \frac{I_{N_{k,x_0}}(x_i)}{k}
$$

3. Please note that all work is in the R supplement

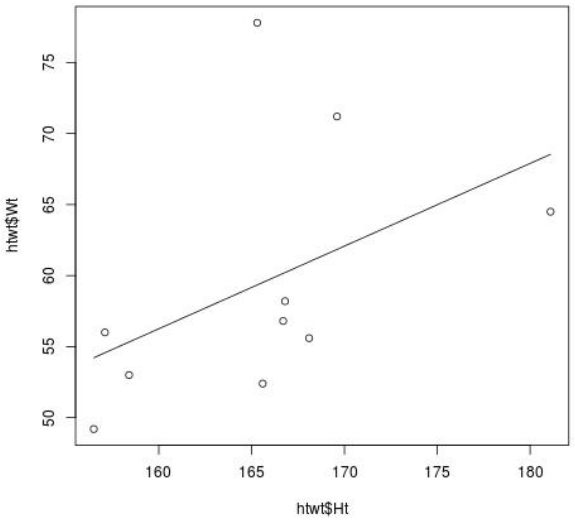


2.1.1

I would argue that this data, though seeming to indicate some sort of correlation bet ween height and weight befitting a linear model, is too small considering the amount of noise for a linear regression to make sense. The estimates would be far too noisy to read anything into it. That said, if there was more data it looks like a linear model would probably make sense.

2.1.2 See the R output for the calculation and confirmation of most of the values.

$$\hat{\beta}_1 = .582, \hat{\beta}_0 = -36.876,$$



2.1.3

$$\sigma^2 = 71.502, se(\hat{\beta}_1) = .389, se(\hat{\beta}_0) = 64.473, Cov(\beta_0, \beta_1) = -25.070$$

$$t_{\beta_1} = 1.496, t_{\beta_0} = -94.752, p_{\beta_1} = .173, p_{\beta_0} = 1.718e-13$$

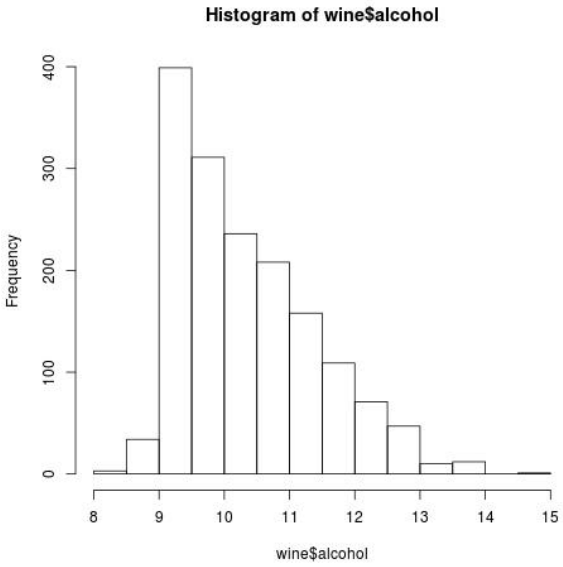
2.1.4

Regression	df	ss	ms	F	pval
Regression	1	159.947	159.947	2.237	≈ 0
Residual	8	572.014	71.502		
Total	731.961				

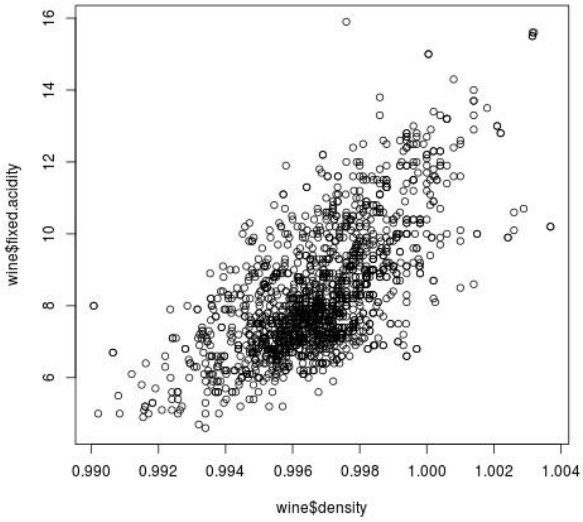
$t^2_{\beta_1} = 2.239 = F$

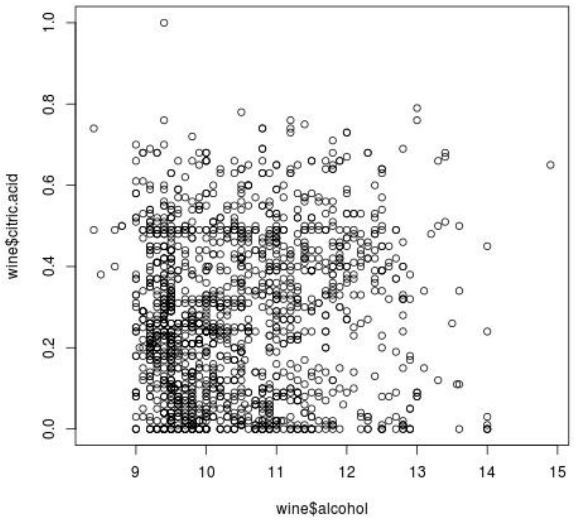
4. Please note hat all work is in the R supplement

   b) While certain variables seem to be approximately normally distributed, such as pH, others seem to be very distinctly not normally distributed, such as alcohol.



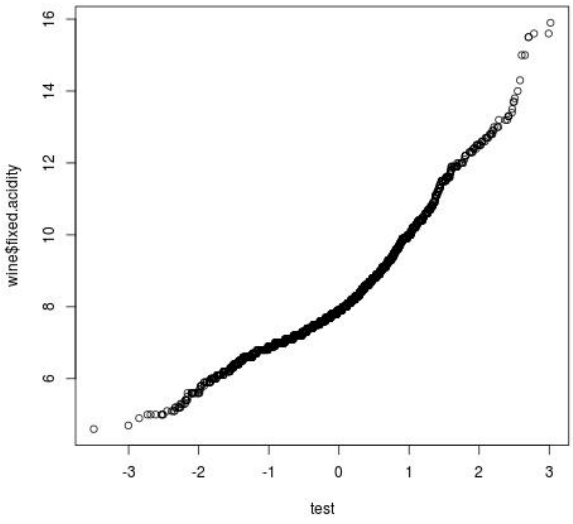
We also see some variables that clearly have a strong correlation, such as density and fixed.acidity, while with others like alcohol and citric.acid, if there is a correlation is hardly obvious.

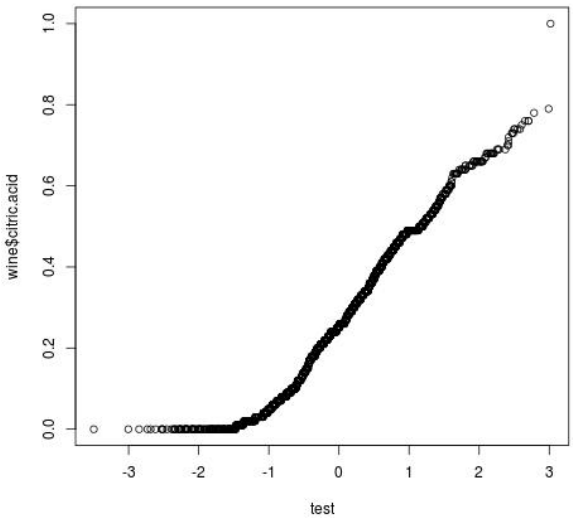


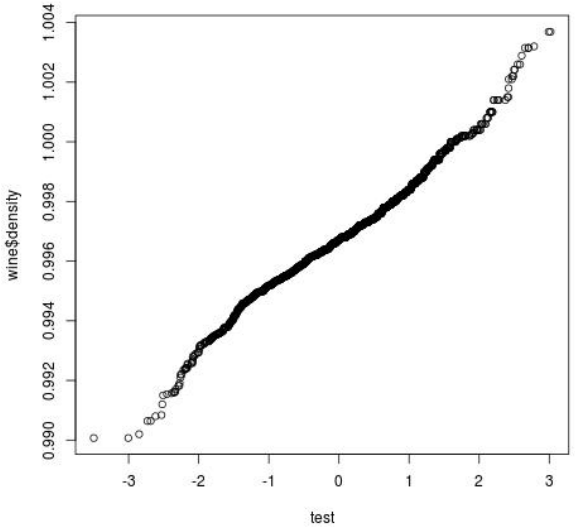


Also, quality is clearly a categorical, rather than truly continuous.

c) Most of these pairwise relationships at least seem to indicate some kind of correlation based on eyeballing the plot. pH and density look the closest to being independent, though there may be a very weak negative correlation. citric.acid and density look like they at best have a weak positive correlation. Its hard to tell if the correlations are linear, but they look to be at worst a little bit curved, such as in the case of fixed.acidity and pH, which can be modeled with a linear model

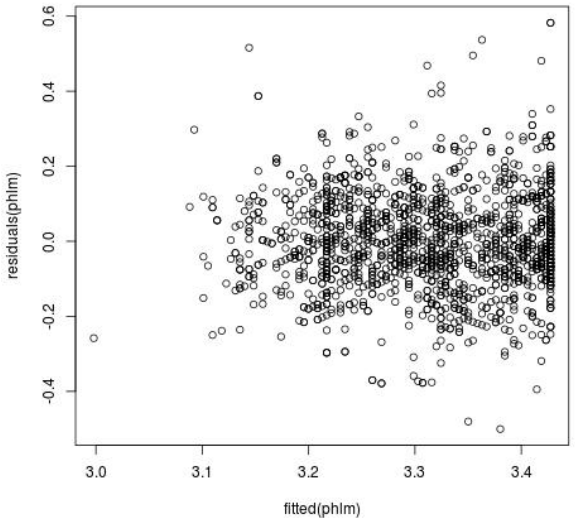

wine$fixed.acidity
16
14
12
10
8
6
-3
-2
-1
0
1
2
3
test


wine$citric.acid
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
test

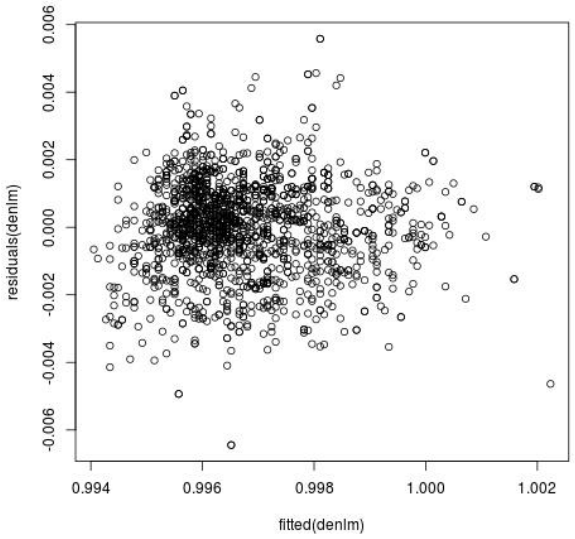


However, from comparing the quantile plots with a vector of normals, it looks like possibly fixed.acidity and definitely citric.acid are not normally distributed. This means are standard error estimates will be off with a linear regression.

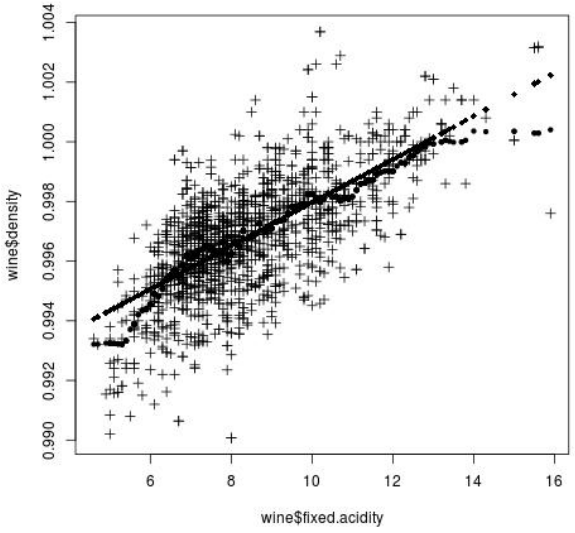
d) Both models seem appropriate based on the usuals statistical tests. For the pH regression, we have an R^2 of .2937, with the F-test for the model and T-tests for the βs essentially yielding zero (smaller than the smallest non-zero number r can represent). Similarly, we get an R^2 of .4463 and p-values of essentially 0 for the density regression as well. However, looking at the residuals plots, it is a little less clear.



I would say the residuals for the pH model look fine, they look symmetrically distributed around 0 throughout the range of fitted values, indicating there isn't bias and that the model is alright.



However, for the density model this doesn't quite seem to be true. We seem to get a negative bias on the tails of the range of fitted values and a positive bias in the middle. This indicates that the linear model, as is, isn't appropriate.



e)

f) It is clear for the pH model that the linear model (diamonds) is a better fit than the nearest neighbor model (circles), since its global assumption of linearity appears to hold. The nearest neighbor model is essentially bouncing around the linear estimate, indicating that it merely has higher variance without capturing the underlying relationship better (it doesn't seem to have less bias).

   On the other hand, the nearest neighbor model does seem to be doing a better job on the density model. Since it is a local model that introduces less bias, it has picked up on the not quite linear (without transforming fixed.acidity anyway) relationship. The linear regression, since its global and based on a strong assumption, has introduced bias since the assumption failed.

g) Considering that citric.acid is an acid, and pH is a measure for acidity, with lower values indicating higher acidity, it would stand to reason that a decrease in citric.acid would cause lower pH, as we have seen.
   Fixed acids seem to, in general, have much higher density than alcohol and water, which are the two main ingredients in wine. Thus, fixed acids increase the density of wine, and all else equal, one would expect higher concentrations of them would correspond to higher density.