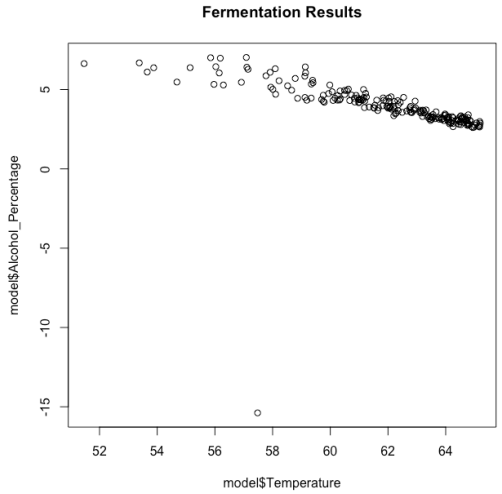
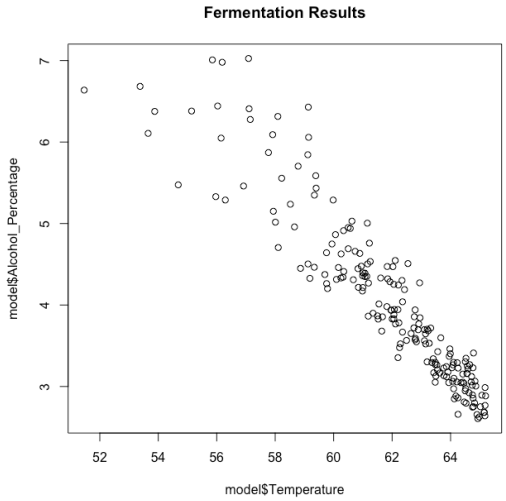
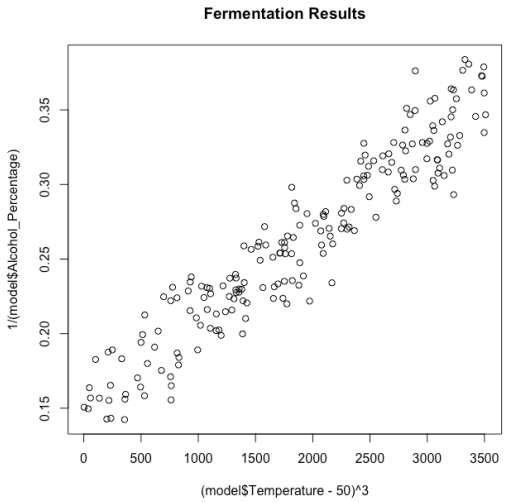
Luke Beebe
STAT 463 - Assignment 3 [code is in R file]

1) There is an datapoint with a negative alcohol percentage This outlier makes no rational sense (Unless they now make beers that turn you sober), so I removed it.



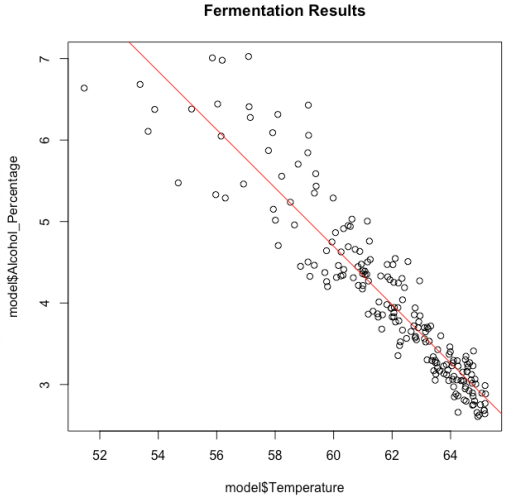

2) The new plot shows a clearer picture of the data which seems to meet the assumptions for a simple linear model. However, we see that the explanatory variable predicts the variance of the data which means that if we were to look for intervals, we'd have to make the heteroscedastic data homoscedastic by transforming the variables. The transformations I chose are shown on the x and y labels. I managed to stabilize the variance.




3) The parameter estimates is alc%=26.174-0.358t. The $R^2$ is 86.89%.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.173742   0.613251   42.68   <2e-16 ***
Temperature -0.357968   0.009907  -36.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.38 on 197 degrees of freedom
Multiple R-squared:  0.8689,    Adjusted R-squared:  0.8682
F-statistic:  1306 on 1 and 197 DF,  p-value: < 2.2e-16
```

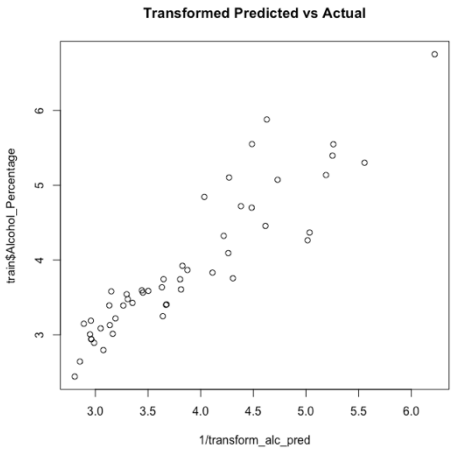


4) The sample correlation between predicted and actual values of the validation dataset is 91.70591%.

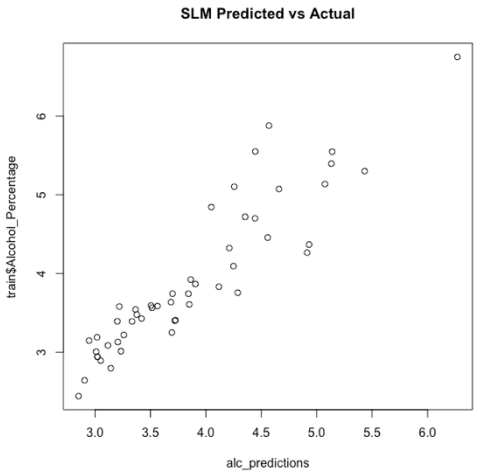
5) The steps I took in part two were enough to warrant the use of simple linear regression for this goal, so long as I transform the new values back, so they are useful for the original plot.

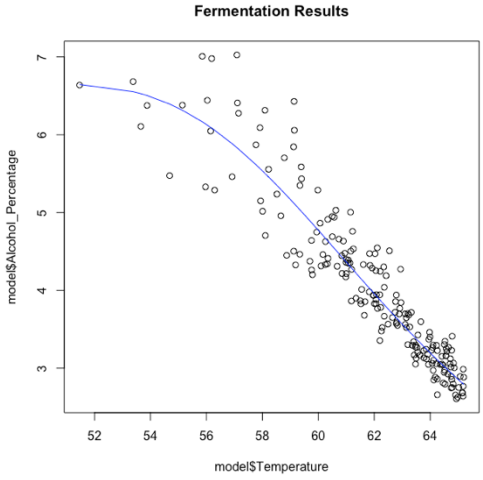
6) I will restate what I said in number two, that I managed to stabilize the variance. I did this using two transformations:
$y_t=1/y_o$ $x_t=(x_o-50)^3$
$y_t=1.503*10^{-1}+(5.926*10^{-5})x_t$
$y_o=(1.503*10^{-1}+(5.926*10^{-5})(x_0-50)^3)^{-1}$
I found the $R^2$ was slightly better with this model. Here is the regression curve plotted.

7) The sample correlation between predicted and actual values of the transformed model and validation dataset was lower at 91.51226%




8) The plot with bands is shown to the right. You can see the prediction interval adjusts for the variance. The formula for the upper and lower endpoints for these bands is given in this picture above the chart, where $x_p$ ($x_t$) and $y_p$ ($y_o$) are transformed values (from #6). You then find $y_t$ by taking the reciprocal of the number this equation produces, giving us the upper and lower bounds of the prediction interval.




$$\hat{y}_p \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}.$$