MA4710 PROJECT 4

DANIEL HENDERSON

Introduction

We will explore a few transformation options to address violations of our linearity assumption that was identified last week. To aid in our investigation, years climbing is transformed back to a numeric predictor. The explanatory value that it added when treating it categorically with 30-plus levels is minimal and it greatly increases the model's complexity.

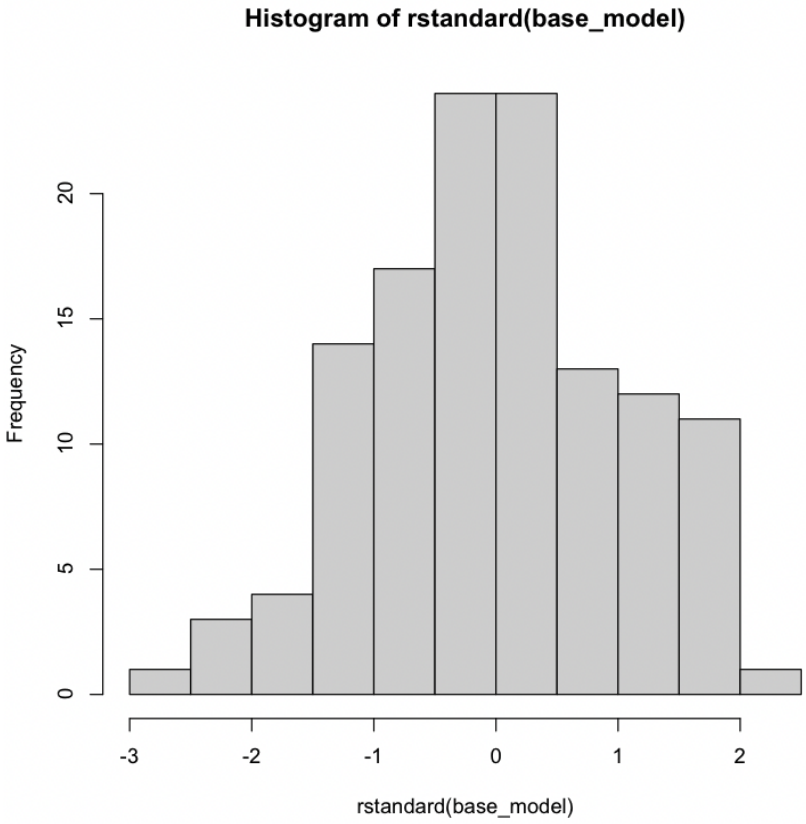
There where a couple of changes to the data this week. The biggest being the removal of 2 observations that showed obvious errors in their reported weight values. (observation 109 and 120 in last week's csv). This addressed some underlying issues in our residuals verse bmi plot that should have been caught last week. Second, I removed all of the protocol categorical variables from the csv file to aid in my analysis. After investigation, these columns appeared to play no important role in explaining the variance in the observed climbing grade in our sample data. This is due to all of the levels and lack of observations in some of the levels, and should not be taken as evidence that they play no role in explaining the variance of hardest climbing grade of all climbers.

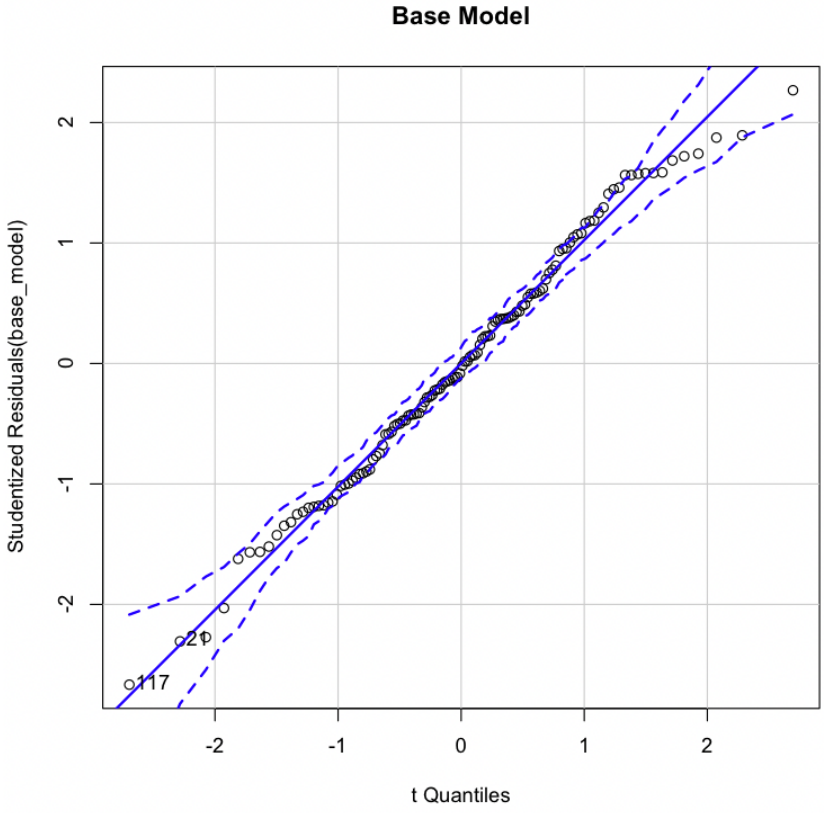
The starting model is shown below

```
Call:
lm(formula = grade ~ years_climbing + hangboard_freq + sex *
    bmi + max_pull, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4978 -1.0962 -0.0723  0.9259  3.1836

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.728e+00  4.011e+00  -2.176 0.031577 *
years_climbing  1.970e-01  3.038e-02   6.485 2.21e-09 ***
hangboard_freq  5.469e-01  1.431e-01   3.820 0.000215 ***
sexMale         1.591e+01  4.252e+00   3.742 0.000285 ***
bmi             5.733e+03  1.923e+03   2.982 0.003490 **
max_pull        1.160e-01  1.425e-02   8.138 4.85e-13 ***
sexMale:bmi    -7.321e+03  2.026e+03  -3.613 0.000448 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.486 on 117 degrees of freedom
Multiple R-squared:  0.594,     Adjusted R-squared:  0.5732
F-statistic: 28.53 on 6 and 117 DF,  p-value: < 2.2e-16
```

Initial Assumption Evaluation

Our normality of residuals assumption is holding quite well as displayed by our histogram and q-q plot.

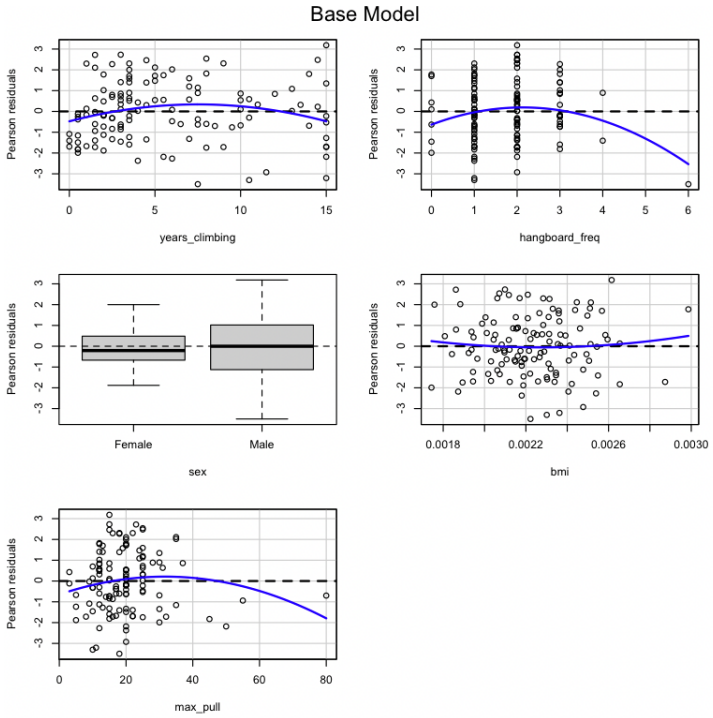




Additionally the Shapiro-Wilkes test reports a p-value of 0.4226 which supports the assumption of normality.

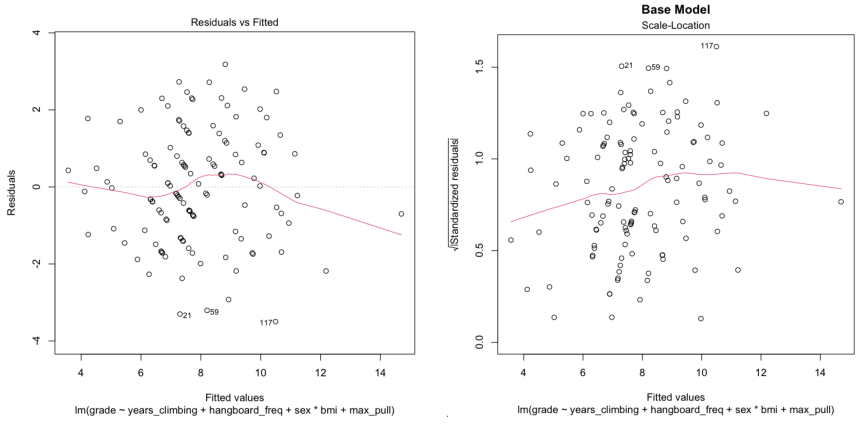
Initial Assumption Evaluation Continued

Next we look at our linearity and homoscedasticity assumption. First observe the residuals plots against each predictor in our model



The plots above suggest that we may be able to improve our linearity assumption of the model via transformations of the max pull up reps and hangboard frequency predictors. Additionally, there is one potential violation of constant variance in our set of predictors, specifically in years climbing. There isn't an obvious funnel shape to the residuals when look at the plot, but there are fewer residuals close to zero as years climbing increases. This subtle violation will become apparent when we attempt a log transformation to this variable. But one should note that this makes a lot of sense because some climbers plateau while others excel.

Next observe the residual verse fitted value plots shown below



When looking at the scale location it is evident that our standardized residuals are getting slightly larger as the fitted values increase, which is evidence to suggest that our constant variance assumption is violated. When viewing the left hand side plot, we see a slight decrease to the trend line for larger fitted values. This implies we are having some issues modeling elite climbers and in violation of our linearity assumption. Note, this was also apparent when we looked at residuals verses hangboard frequency and max pull up rep predictors.

In summary, the more pressing matter is to deal with issues of model misspecification. After finding a better model, we will reassess our constant variance assumption.

Revisiting Initial Data Exploration

By analyzing the scatter plot matrix of our entire data set the following observations occurs:

1. If hangboard frequency is a good predictor, then climbing sessions certainly should be better
2. Max push seems to have a linear relationship with grade but the reciprocal of max push looks even better.
3. Similarly we might try to transform max pull by raising it to the power of -1 or -0.5.
4. Lack of observations of years climbing in upper left of plot may suggest a log transform is needed

The scatter plot matrix is not included in this report due to it's size. However, it can be found in Rplots.pdf upon running the prog4.R script.

Addition of Climbing Sessions

It is evident that there is a trend between climbing sessions and hardest grade - a rather good one. This should have been identified earlier, nonetheless, the new model summary is shown below.

```
Call:
lm(formula = grade ~ years_climbing + hangboard_freq + sex +
    bmi + max_pull + climbing_sessions + sex:bmi, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3580 -1.0079 -0.1071  1.0671  3.3287

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -8.108e+00  3.905e+00  -2.076 0.040089 *
years_climbing     2.028e-01  2.960e-02   6.850 3.76e-10 ***
hangboard_freq     4.061e-01  1.480e-01   2.744 0.007035 **
sexMale            1.450e+01  4.163e+00   3.483 0.000699 ***
bmi                4.956e+03  1.890e+03   2.623 0.009892 **
max_pull           1.044e-01  1.445e-02   7.226 5.68e-11 ***
climbing_sessions  3.872e-01  1.384e-01   2.797 0.006038 **
sexMale:bmi       -6.562e+03  1.988e+03  -3.300 0.001284 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.444 on 116 degrees of freedom
Multiple R-squared:  0.6197,    Adjusted R-squared:  0.5967
F-statistic:    27 on 7 and 116 DF,  p-value: < 2.2e-16
```

As you can see, the addition of this new predictor reduced $\hat{\sigma}$ and increased both R^2 and the adjusted-R^2. For further confirmation, an analysis of variance between the two models reports a rather significant p-value as shown below.

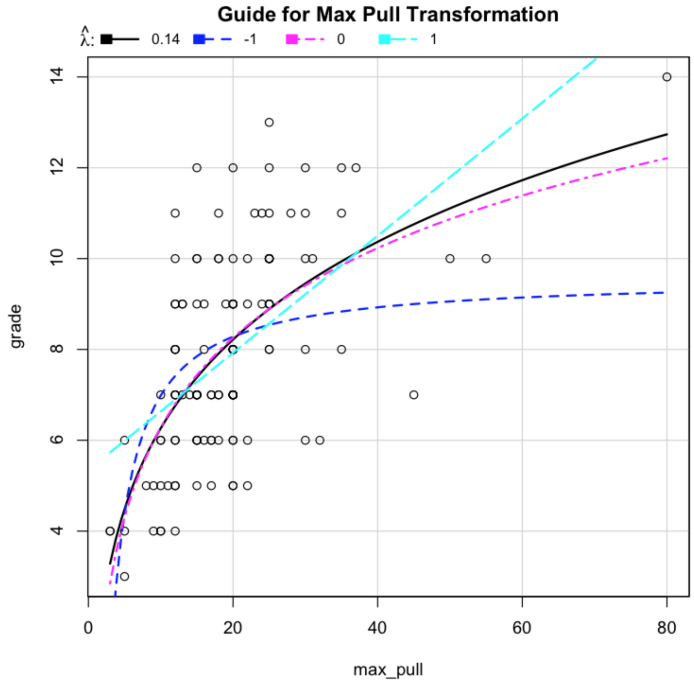
```
Analysis of Variance Table

Model 1: grade ~ years_climbing + hangboard_freq + sex * bmi + max_pull
Model 2: grade ~ years_climbing + hangboard_freq + sex + bmi + max_pull +
    climbing_sessions + sex:bmi
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    117 258.19
2    116 241.88  1    16.314 7.8239 0.006038 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For breavity, I will not perform another diagnostic analysis on this new model but the plots are included in my script (titled "Base Model Plus Climbing Sessions"). The addition of this predictor did not address the previously identified linearity issues. There where slight improvements to the flattening of the trend-line in the scale-location plot, which suggests an improvement in our assumption that our residuals originate from a population of constant variance. To support this conclusion, I performed a Breusch-Pagan test on the two models and the p-value increased from 0.18 to 0.36 Lastly, the relationship between climbing sessions and hardest grade is rather linear. There is no evidence to suggest that a transformation of this variable is needed to improve its linear relationship with our response variable.

This will be our new working model to try and address the still prevailing linearity issues through data transformations. We will be referring to it as "the working model" from now on.

Transformation of Max Pull Up Reps

To guide us towards an elegant transformation of the max pull up rep predictor we consider the invTranPlot.



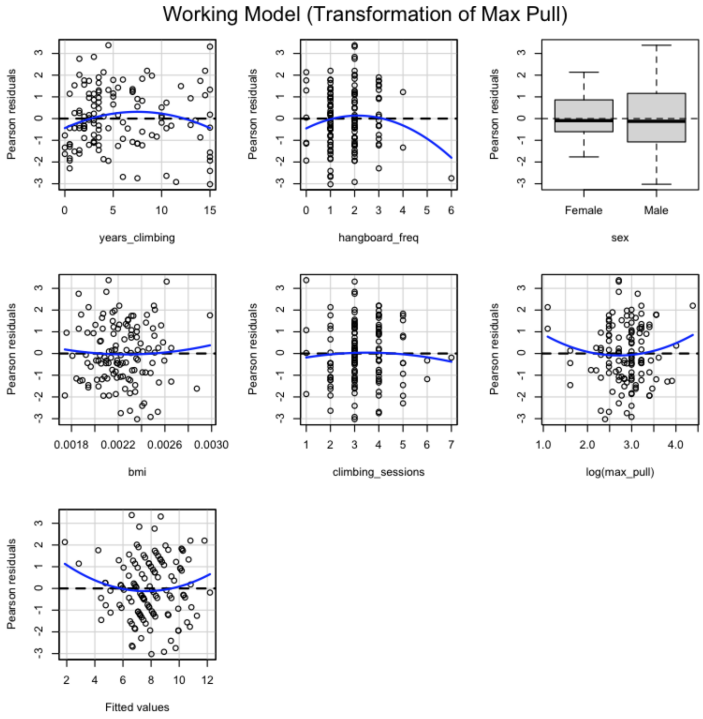
Note that a $\hat{\lambda}$ so close to 0 suggests that we should consider a log transformation. Below is the summary of applying such transformation to our working model.

```
Call:
lm(formula = grade ~ years_climbing + hangboard_freq + sex +
    bmi + climbing_sessions + log(max_pull) + sex:bmi, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-3.023 -1.069 -0.094  1.141  3.377

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.184e+01  3.930e+00  -3.013  0.00318 **
years_climbing     1.882e-01  2.951e-02   6.376 3.84e-09 ***
hangboard_freq     2.696e-01  1.456e-01   1.852  0.06654 .
sexMale            1.311e+01  4.113e+00   3.189  0.00184 **
bmi                4.880e+03  1.865e+03   2.617  0.01006 *
climbing_sessions  3.958e-01  1.361e-01   2.908  0.00436 **
log(max_pull)      2.422e+00  3.222e-01   7.515 1.30e-11 ***
sexMale:bmi       -6.293e+03  1.962e+03  -3.207  0.00173 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.426 on 116 degrees of freedom
Multiple R-squared:  0.629,     Adjusted R-squared:  0.6067
F-statistic:  28.1 on 7 and 116 DF,  p-value: < 2.2e-16
```

This model certainly provides us with more explanatory power. However, there is a substantial loss in significance of the hangboard frequency coefficient. We will be sticking with the rule of dropping any terms that have a p-value greater than .1, so for now it stays. However, this predictor showed signs of linearity violations thus a better treatment of it in our model may yield the lost significance.

Transformation of Max Pull Up Reps Continued

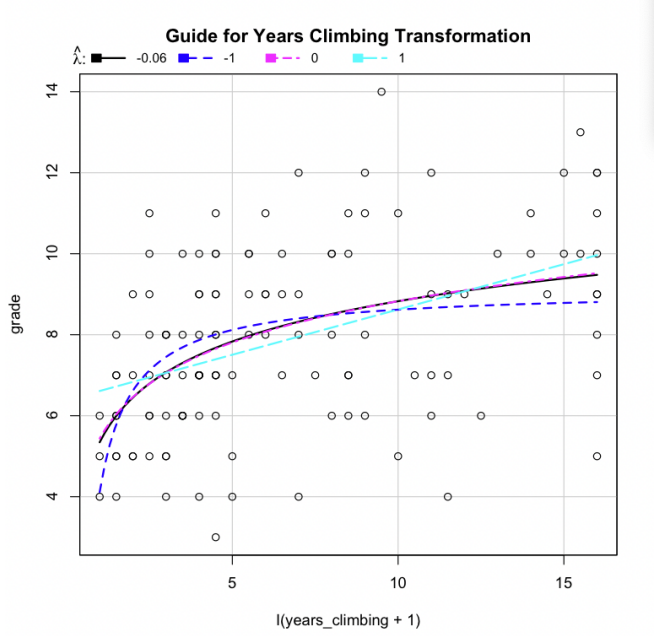
Next we perform a brief diagnostic check to ensure that this transformation did'nt grossly violate any of our assumptions. The associated plots can be found in Rplots.pdf under the title "Working Model (Transformation of Max Pull)". However, we do show the residual plots below.



As you can see the linearity violation between hardest grade and max pull up reps has indeed improved. This is probably the best we can hope for and this model will be our new working model.

Transformation of Years Climbing (Final Transformation)

To guide us towards an elegant transformation of the years climbing predictor we consider the invTranPlot. However, since we are using the lower bound of the categorical intervals to treat this predictor numerically - the value 0 occurs in our data. Thus, we must shift it horizontally by one unit to the right.



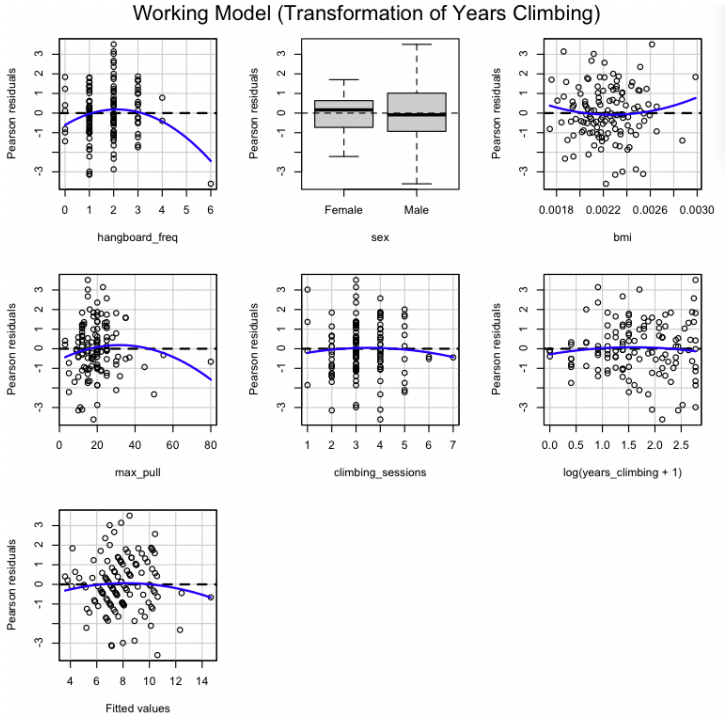
Note this $\hat{\lambda}$ is even closer to 0, suggesting that we should consider a log transformation on years climbing plus 1 - because $\log(0) = -\infty$, so 0 can not be an element of years climbing.

Transformation of Years Climbing Continued

Below is the summary of applying such transformation to our working model

```
Call:
lm(formula = grade ~ hangboard_freq + sex + bmi + max_pull +
    climbing_sessions + log(years_climbing + 1) + sex:bmi, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6089 -0.9273 -0.0808  0.9941  3.5071

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)             -9.872e+00  3.758e+00  -2.627 0.009782 **
hangboard_freq           4.353e-01  1.426e-01   3.053 0.002813 **
sexMale                  1.465e+01  4.014e+00   3.650 0.000394 ***
bmi                      5.254e+03  1.817e+03   2.891 0.004589 **
max_pull                 1.043e-01  1.394e-02   7.488 1.49e-11 ***
climbing_sessions        3.763e-01  1.335e-01   2.820 0.005657 **
log(years_climbing + 1)  1.352e+00  1.761e-01   7.677 5.61e-12 ***
sexMale:bmi             -6.630e+03  1.917e+03  -3.460 0.000758 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.394 on 116 degrees of freedom
Multiple R-squared:  0.6458,    Adjusted R-squared:  0.6244
F-statistic: 30.21 on 7 and 116 DF,  p-value: < 2.2e-16
```

Note that the effect of this transformation definitely increased the explanatory value of our model. Additionally, I find it quite interesting that it restored some of the significance to the hangboard frequency coefficient. This is the inherent issue of trusting such p-values when in violation of our standard assumptions.

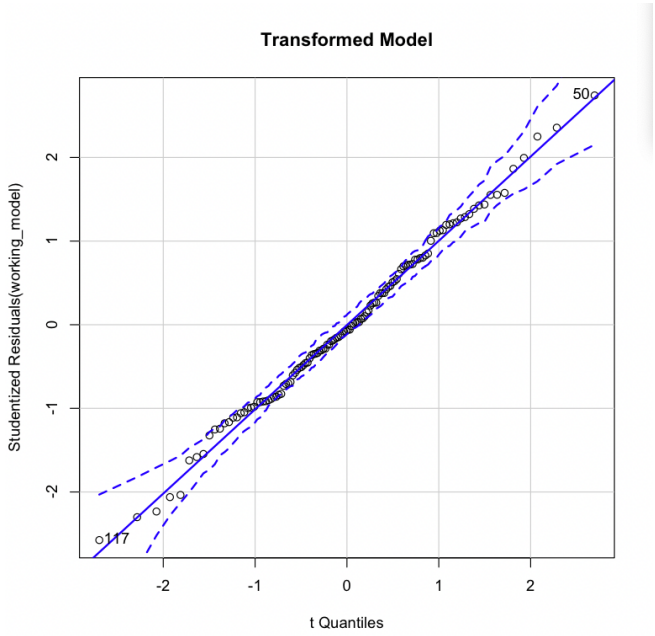


Our issue of linearity on the predictor of years climbing is certainly solved. However, this sheds much light onto the non constant variance issue between years climbing and hardest grade - as shown by the obvious funnel in the plot. This will be investigated in our final conclusion.

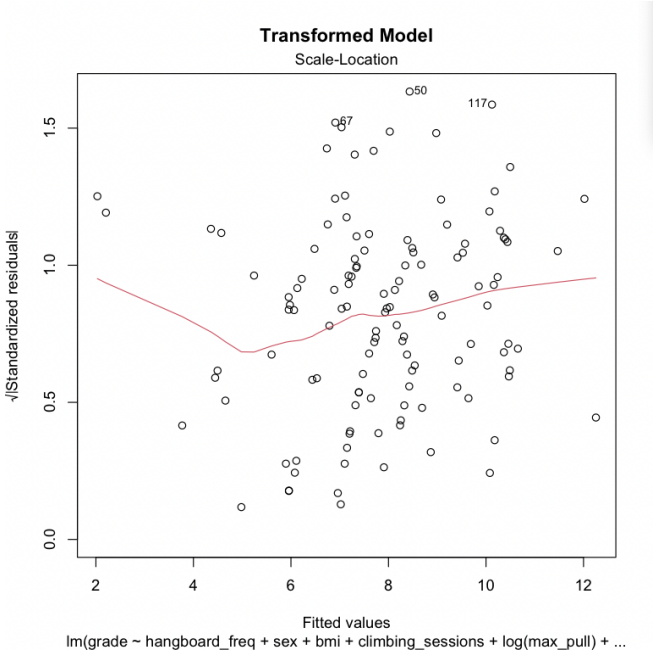
Note that I have investigated max push up reps inclusion, with and without transformations, and at various stages of the model. It appears to be quite collinear with other predictors. Therefore, it never found a place in the final model. Additionally, I have tried various transformations on hangboard frequency and they all drastically reduce the explantory power and significance of it's coefficient. Therefore it remains as is and plays a role in our model.

Conclusion and More Diagnostics of Transformed Model

Note that our adjusted-R^2 increased by 0.0512 units while R^2 increased by 0.0519 units, with only a loss of one degree of freedom. Consequently, our estimate of the standard deviation of our population also improved - from 1.486 to 1.394. It should be noted that our final model has no overly influential observations, though some outliers. We are in much better standing with all of our assumptions. The normality of residuals has improved as displayed by our updated qq-plot



Additonally, the p-value of our Shapiro-Wilkes test nearly doubled to 0.89.
Our assumption of constant variance appears to be more believable when considering the results of Breusch-Pagan test, which initially had a p-value of 0.18 and in our final model the test reported a p-value of 0.46. Graphically, this is displayed by the smoothing of the trend line in the scale location plot. It should be noted that there is obviously an issue in assuming that the variance is constant when using years climbing as predictor of hardest grade - as previously discussed in the years climbing transformation section.



In conclusion, we made vast improvements to our model misspecification issues via our three transformations. This is likely the best we can hope for.