MA4710 PROJECT 5

DANIEL HENDERSON

Introduction

This week we look at a variety of ways to detect collinearity between the numerical variables of my data set. The data set contains the same number of observations as last submission, though, a couple of the variables have been removed. Specifically, height and weight are no longer present in the 'FullData.csv'. This aids in keeping the code in model_selection.R clean. This is not a cause of major concern because body mass index is still a present column. However, there have been some uninvestigated interactions with these terms, which may add value to the model.

After investigating collinearity, we will determine a final model based upon on an exhaustive selection procedure.

Collinearity Investigation on Full Model

After parsing the "FullData.csv" file into a dataframe, we apply the following transformations:

$$\text{years\_climbing} \leftarrow \ln(\text{years\_climbing} + 1)$$
$$\text{max\_pull} \leftarrow \ln(\text{max\_pull})$$
$$\text{max\_push} \leftarrow \text{max\_push}^{-0.5}$$

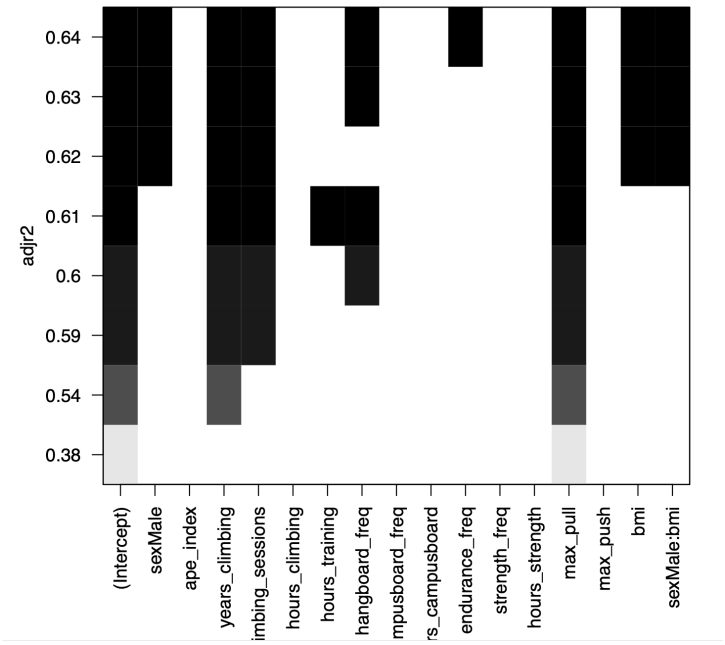
These transformations address previously identified problems of our linearity assumption. They where initially determined by an elegant rounding of the reported $\hat{\lambda}$ value from the invTransPlot function. I determined that the condition number of the scaled covariance matrix of predictors to be $\kappa = 4.3$, which supports that claim that there are no significant signs of collinearity between our predictors. For further evidence, consider the Variance Inflation Factors shown below

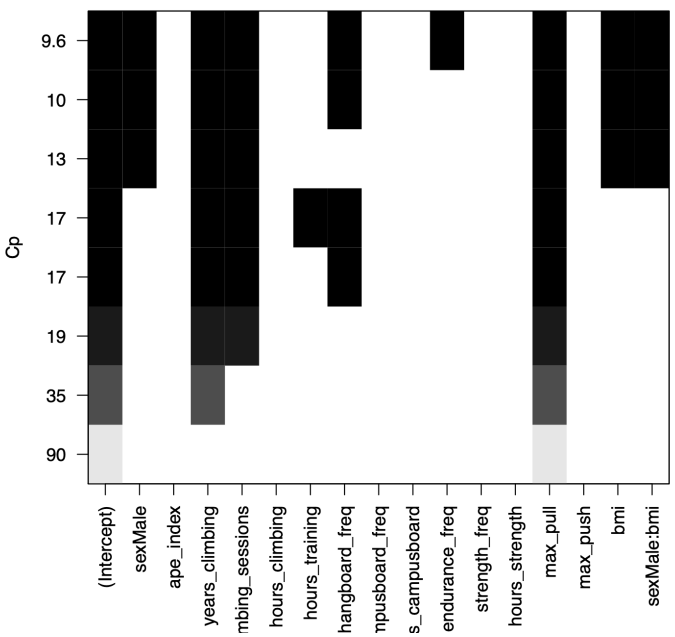
sex	ape_index	years_climbing	climbing_sessions
1.708048	1.044264	1.193404	2.033235
hours_climbing	hours_training	hangboard_freq	campusboard_freq
1.635883	1.704772	1.351770	3.623048
hours_campusboard	endurance_freq	strength_freq	hours_strength
3.643525	1.239905	1.940075	1.993203
max_pull	max_push	bmi	
2.070233	1.830645	1.301739	

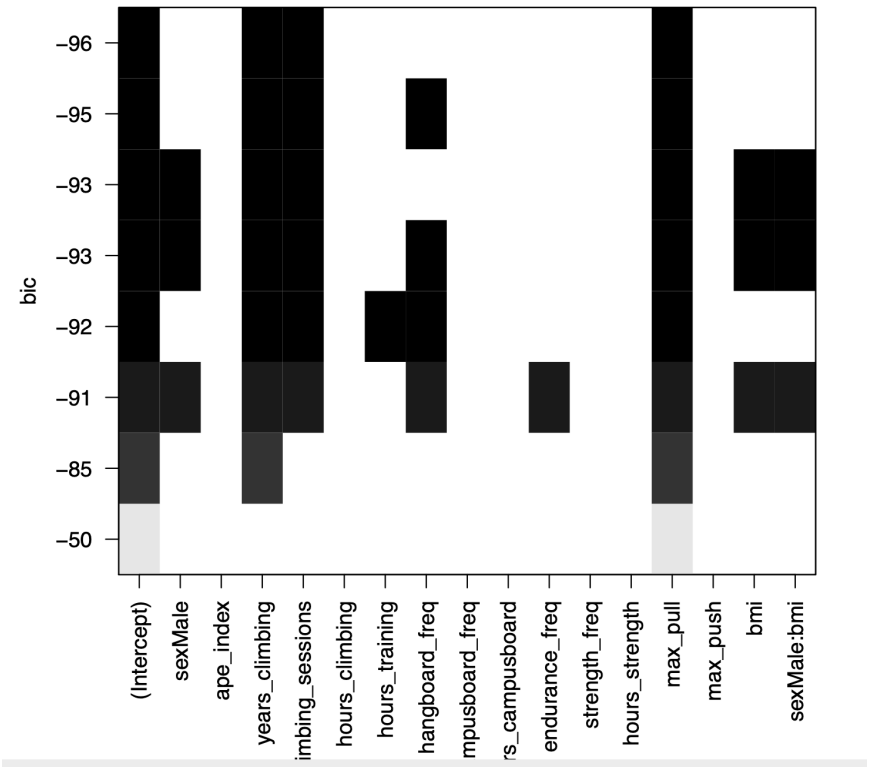
Notice that campusboard_freq and hours_campusboard have the highest variance inflation factors. Additionally, they are nearly identical. This makes sense - there out to be relationship between the amount of times you engage in an exercise per week and the total number of hours doing the exercise per week. The covariance between the two variables is 0.43, which aligns with the above reasoning. However, both of these values are not a major concern according the crude rule of thumb that we should flag a VIF exceeding 10.

Variable Selection Proceduresl

In the model_selection.R script, we perform an exhaustive search over the columns of our dataframe to determine the "best" models of various size. We let out intuition guide us and allow the sex variable to interacte pairwise with bmi as an option when we determine the best subset models. Below are some diagnostic plots of this procedure:







Note that I also performed a forward selection, backward elimination, and a stepwise selection algorithm to the data. These criteria's didn't uncover anything novel - the exhaustive search was much more encompassing.

Variable Selection Procedures Continued

The utility of the model under investigation would be to make an unbiased prediction for training purposes, thus, we seek to maximize the adjusted R^2 value and minimize $\hat{\sigma}$. Due to the various transformations of the best predictors, our interpretability of the model coefficients is limited. Additionally, I believe we are missing some predictors that should be included in our model - assuming there even is a "true" regression model - so we will not try and make an inference on the individual coefficients. After investigating the results that are suggested by the plots on the previous page, I settled on the following model:

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

Residuals:
    Min      1Q  Median      3Q     Max
-3.2496 -1.0059 -0.0299  0.9255  3.5834

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -14.5198     3.8168  -3.804 0.000229 ***
sexMale              14.1891     3.9888   3.557 0.000546 ***
bmi                5674.8326  1814.5856   3.127 0.002235 **
max_pull              2.4221     0.3096   7.824 2.73e-12 ***
endurance_freq       -0.2357     0.1435  -1.643 0.103074
hangboard_freq        0.3342     0.1415   2.361 0.019894 *
climbing_sessions     0.4258     0.1327   3.209 0.001724 **
years_climbing        1.2592     0.1748   7.204 6.53e-11 ***
sexMale:bmi       -6827.0162  1905.0840  -3.584 0.000498 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.37 on 115 degrees of freedom
Multiple R-squared:  0.6605,    Adjusted R-squared:  0.6369
F-statistic: 27.97 on 8 and 115 DF,  p-value: < 2.2e-16
```

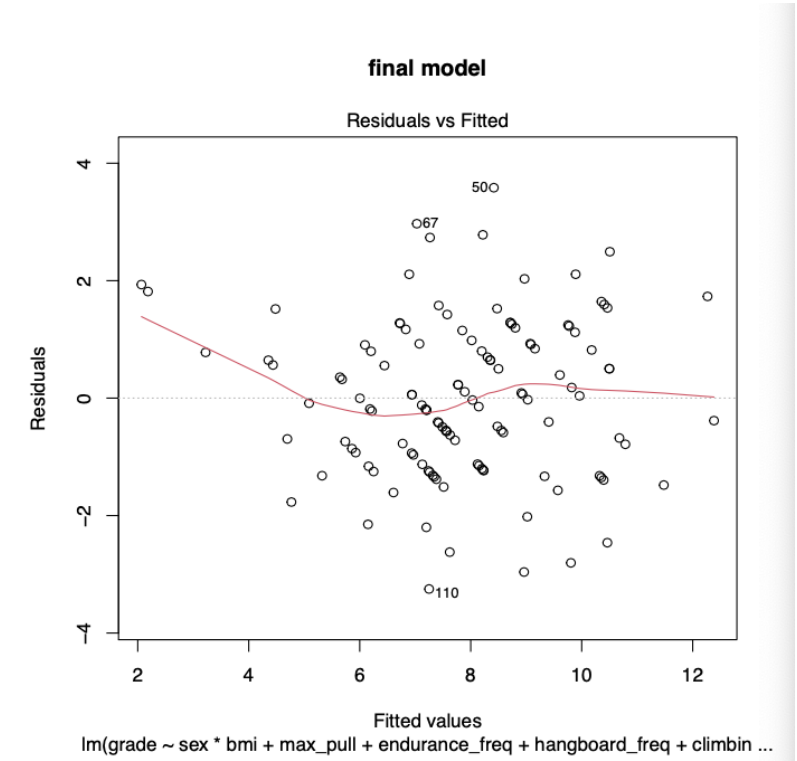
Recall that a couple of the variables are transformed, so the actual linear model Rish formula is

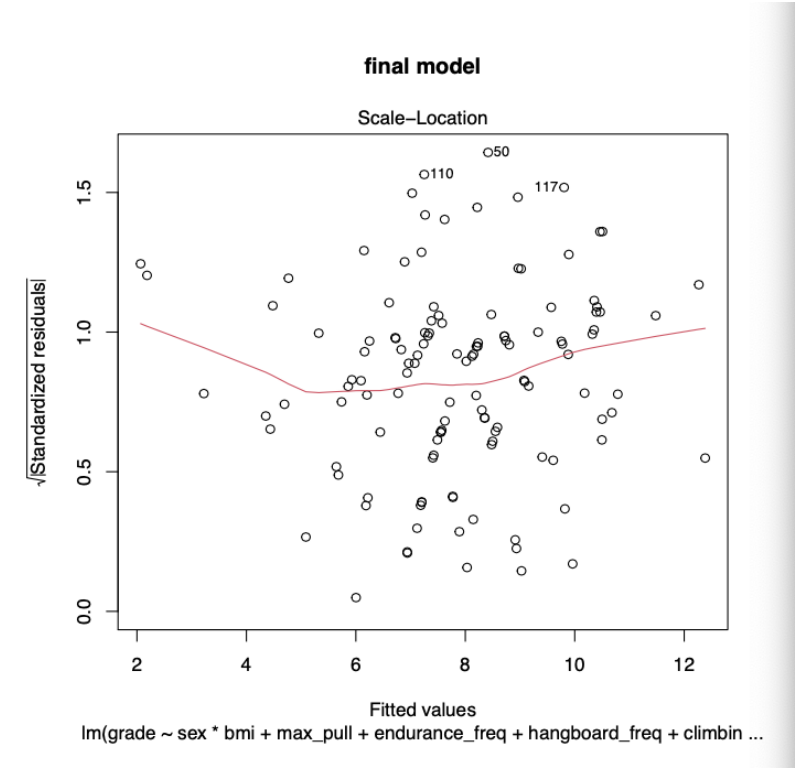
$$\text{grade} \to \ln(\text{years\_climbing} + 1) + \ln(\text{max\_pull}) + \text{max\_push}^{-0.5} + \text{sex} + \text{bmi} + \text{sex:bmi} + \text{endurance\_freq} + \text{hangboard\_freq} + \text{climbing\_sessions}.$$

This model is appealing because it adequently meets our assumptions and has the smallest mallows C_p value. Specifically, $C_p = 9.6$ which is approximately equal to the number of columns in it's corresponding model matrix. Though, it is slightly larger which may be an indication that our model is not taking into account a needed predictor. That is, it may be an underfitted model.

Final Model Assumption Diagnostics

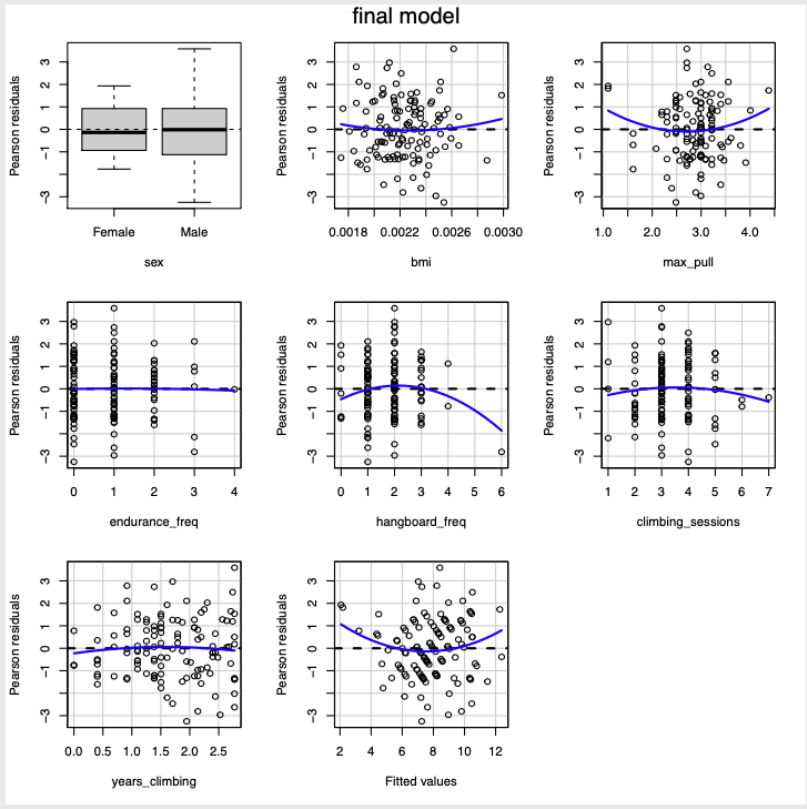
Next we take a look at our standard assumptions on this model. The Shapiro test for normallity of the standardized residuals reports a p-value of 0.85, which strongly supports our assumption that the residuals are normally distributed. However, much stronger evidence is shown in the histogram and normal q-q plot. Both of these plots can be found in Rplots.pdf and are omitted because they are great examples of plots generated from normally distributed residuals. The Breusch-Pagan test reports a p-value of 0.47, which is evidence that our constant variance assumption of the residuals is not violated. The homoscedasticty assumption is strengthened by the lack of a funnel pattern in a plot of normal residuals verse the models fitted values. This can also be concluded from the lack of an increasing or decreasing trend when we standardize the residuals to obtain the scale-location plot.



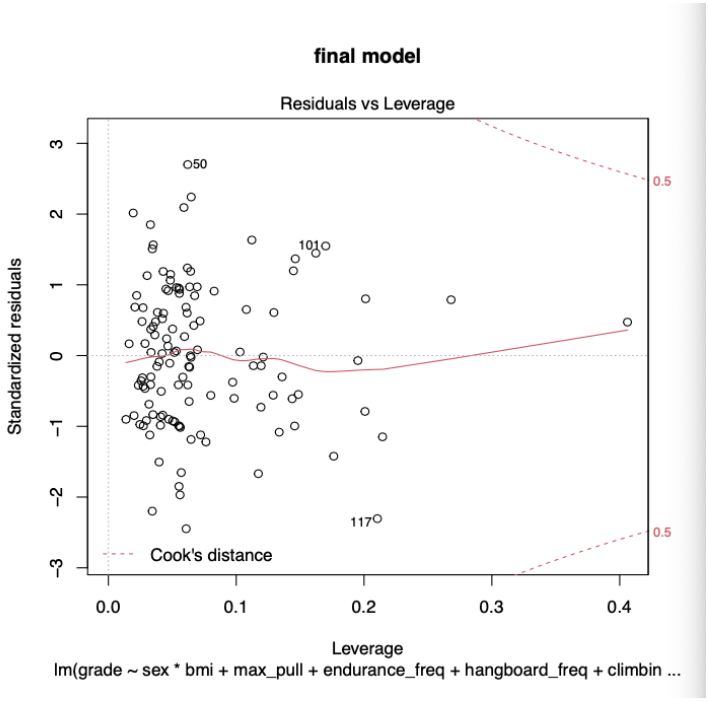


Final Model Assumption Diagnostics Continued

Furthermore, when looking at both the plots on the previous page there is no obvious violation of our linearity assumption. The next step is to look at our residual plots verse each predictor, where a linearity violation (or constant variance assumption) may become apparent.



There is an outlier in the hangboard_freq varible, but a transformation to address the problem is not warranted in my opinion. The Tukey's test of additivity does not report a highly significant p-value to support the inclusion of any quadratic terms. Also, there is a potential violation of our constant variance assumption by the slight funneling in the years_climbing plot - nothing we can really do. Next note that there are no high-leverage/influential observationns in our data, as shown by the plot below.



Final Model Assumption Diagnostics Continued

In conclusion, this model specification and consequent assumptions about the residuals are probably as good as we can get. And recall that we didn't uncover a presence of collinearity earlier, so we safely assume that there is not a strong linear relationship between any of the predictor's in our final model matrix. That is, (I believe) it is safe to assume that the conditioning of this least-squares problem is not of concern soley based upon my previous work.