MA4710 PROJECT 3

DANIEL HENDERSON

Introduction

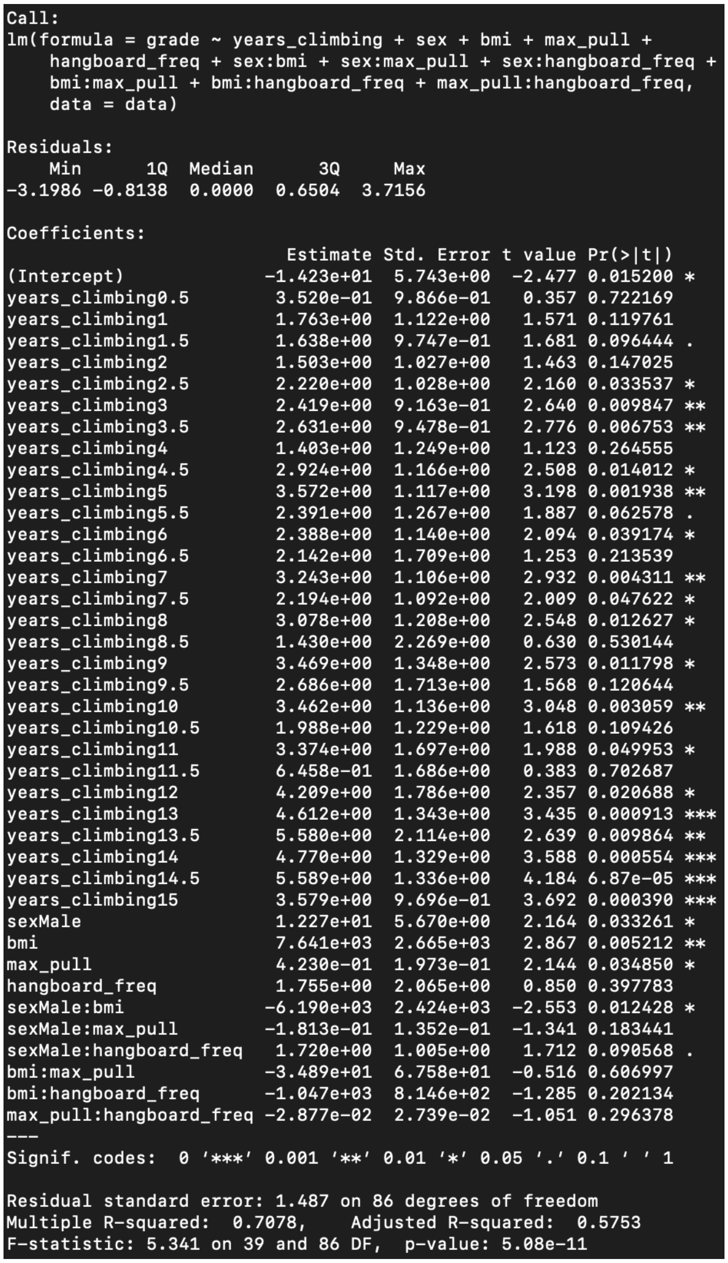
This week I explore adding categorical predictors and interaction terms into my linear model.

As explained in my first report, my data set comes from a survey that was posted on the /climbharder subredit. In the survey, participants where asked to select an interval of how long they have been climbing. The length of the interval was half a year, with the lower bound being 0 years, the upper bound being 15 years, and there was one additional category to select for those who have been climbing for more than 15 years. Thus, in the raw dataset years climbing was a categorical predictor. However, prior to this week I was using the lower bound of each interval to use years climbing as a numeric predictor in my model. There where some linearity issues present when using years climbing as a numeric, which are addressed this week when we treat it as a categorical variable.

Since I had transformed years climbing in my data set to a numeric variable, I started this weeks investigation from scratch. With the use of the na.omit function in R, I removed all observational units that where missing values in a numeric predictor field of the survey. Next, there where various categorical predictors that where missing responses. I eliminated any categorical variable that was left blank using the subset function in R. Lastly, there where numerous survey participants who reported some of the numeric values in improper units. I uncovered these observations in my diagnostic analysis and transformed such observational fields to units of centimeters and kilograms, where appropriate. The resulting data set is found in 'FullData.csv', where Full refers to the dataset of survey participants who completed the questions as asked. (there was no investigation to confirm that this elimination was done so without bias). There are now 126 observational units in this set

Model Reduction

We start this week's investigation by modeling hardest climbing grade as the response, with years climbing, hangboard frequency, sex, body mass index, and max pull-up reps as predictors. Initially, each predictor and all pairwise interactions between the predictors are included in the model. As one could infer from my discussion of years climbing above, it was introduced into my model this week as a categorical predictor with 30 levels, where $0 - 0.5$ years experience as the control. For various reasons, higher levels of years climbing produced many NA coefficients when it interacted pairwise with other predictors in my model. I suspect this is due to a lack of observations in the uppers levels of years climbing. That is, not many people with 5+ years of experience filled out the survey. However the sport of climbing has grown drastically in the last 5 years. So our sample still may adequately reflect our population. Nonetheless, to eliminate the NA coefficients I removed all interactions that included years climbing and I obtained the following model

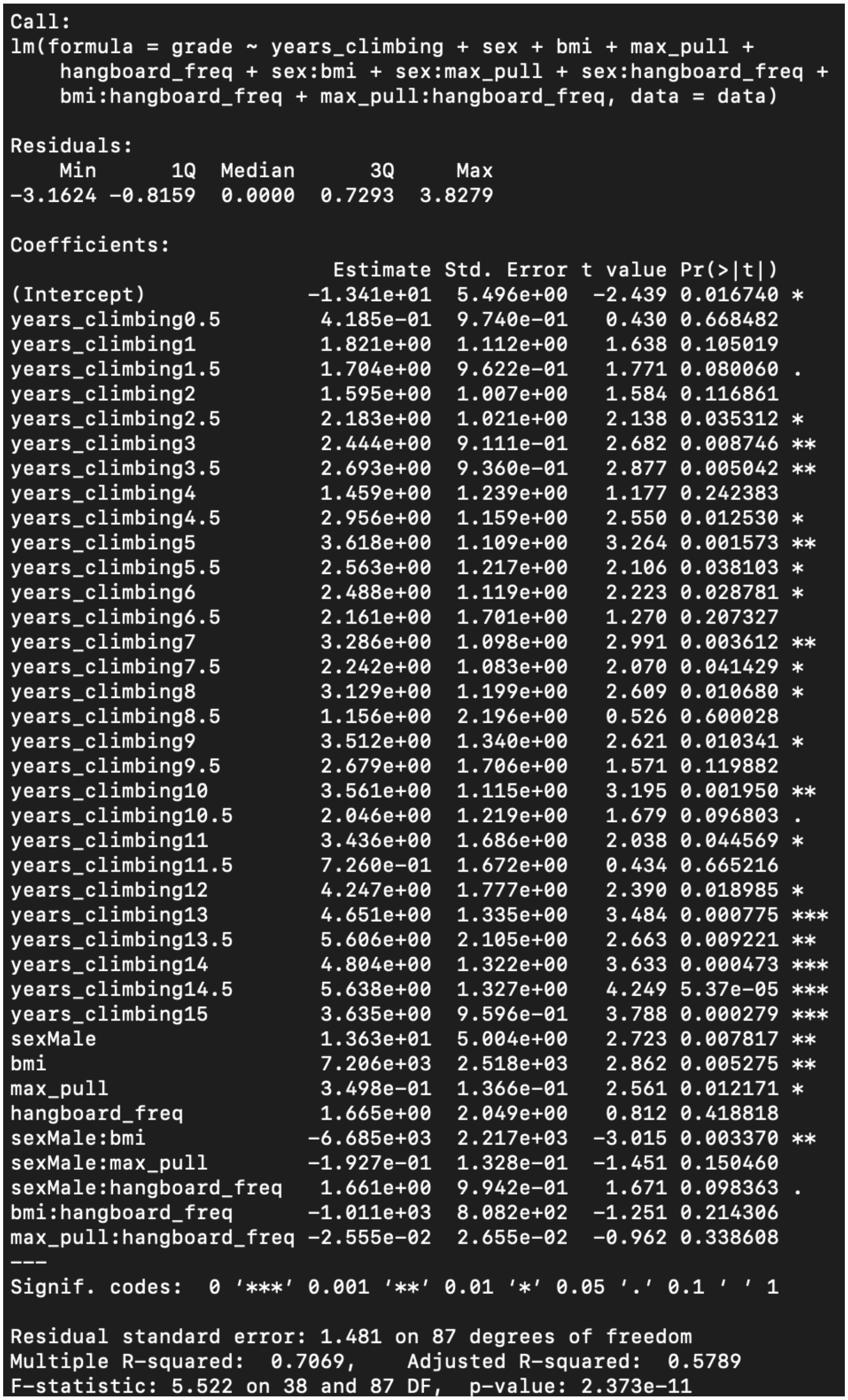


```
Call:
lm(formula = grade ~ years_climbing + sex + bmi + max_pull +
    hangboard_freq + sex:bmi + sex:max_pull + sex:hangboard_freq +
    bmi:max_pull + bmi:hangboard_freq + max_pull:hangboard_freq,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1986 -0.8138  0.0000  0.6504  3.7156

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.423e+01  5.743e+00  -2.477 0.015200 *
years_climbing0.5        3.520e-01  9.866e-01   0.357 0.722169
years_climbing1          1.763e+00  1.122e+00   1.571 0.119761
years_climbing1.5        1.638e+00  9.747e-01   1.681 0.096444 .
years_climbing2          1.503e+00  1.027e+00   1.463 0.147025
years_climbing2.5        2.220e+00  1.028e+00   2.160 0.033537 *
years_climbing3          2.419e+00  9.163e-01   2.640 0.009847 **
years_climbing3.5        2.631e+00  9.478e-01   2.776 0.006753 **
years_climbing4          1.403e+00  1.249e+00   1.123 0.264555
years_climbing4.5        2.924e+00  1.166e+00   2.508 0.014012 *
years_climbing5          3.572e+00  1.117e+00   3.198 0.001938 **
years_climbing5.5        2.391e+00  1.267e+00   1.887 0.062578 .
years_climbing6          2.388e+00  1.140e+00   2.094 0.039174 *
years_climbing6.5        2.142e+00  1.709e+00   1.253 0.213539
years_climbing7          3.243e+00  1.106e+00   2.932 0.004311 **
years_climbing7.5        2.194e+00  1.092e+00   2.009 0.047622 *
years_climbing8          3.078e+00  1.208e+00   2.548 0.012627 *
years_climbing8.5        1.430e+00  2.269e+00   0.630 0.530144
years_climbing9          3.469e+00  1.348e+00   2.573 0.011798 *
years_climbing9.5        2.686e+00  1.713e+00   1.568 0.120644
years_climbing10         3.462e+00  1.136e+00   3.048 0.003059 **
years_climbing10.5       1.988e+00  1.229e+00   1.618 0.109426
years_climbing11         3.374e+00  1.697e+00   1.988 0.049953 *
years_climbing11.5       6.458e-01  1.686e+00   0.383 0.702687
years_climbing12         4.209e+00  1.786e+00   2.357 0.020688 *
years_climbing13         4.612e+00  1.343e+00   3.435 0.000913 ***
years_climbing13.5       5.580e+00  2.114e+00   2.639 0.009864 **
years_climbing14         4.770e+00  1.329e+00   3.588 0.000554 ***
years_climbing14.5       5.589e+00  1.336e+00   4.184 6.87e-05 ***
years_climbing15         3.579e+00  9.696e-01   3.692 0.000390 ***
sexMale                  1.227e+01  5.670e+00   2.164 0.033261 *
bmi                      7.641e+03  2.665e+03   2.867 0.005212 **
max_pull                 4.230e-01  1.973e-01   2.144 0.034850 *
hangboard_freq           1.755e+00  2.065e+00   0.850 0.397783
sexMale:bmi             -6.190e+03  2.424e+03  -2.553 0.012428 *
sexMale:max_pull        -1.813e-01  1.352e-01  -1.341 0.183441
sexMale:hangboard_freq   1.720e+00  1.005e+00   1.712 0.090568 .
bmi:max_pull            -3.489e+01  6.758e+01  -0.516 0.606997
bmi:hangboard_freq      -1.047e+03  8.146e+02  -1.285 0.202134
max_pull:hangboard_freq -2.877e-02  2.739e-02  -1.051 0.296378
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.487 on 86 degrees of freedom
Multiple R-squared:  0.7078,    Adjusted R-squared:  0.5753
F-statistic: 5.341 on 39 and 86 DF,  p-value: 5.08e-11
```

Note that the lower bound of each level of years climbing is used to denote levels. As you can see, the interaction term with the largest p-value is bmi:max_pull. We proceed by removing this term from the model

Model Reduction Continued

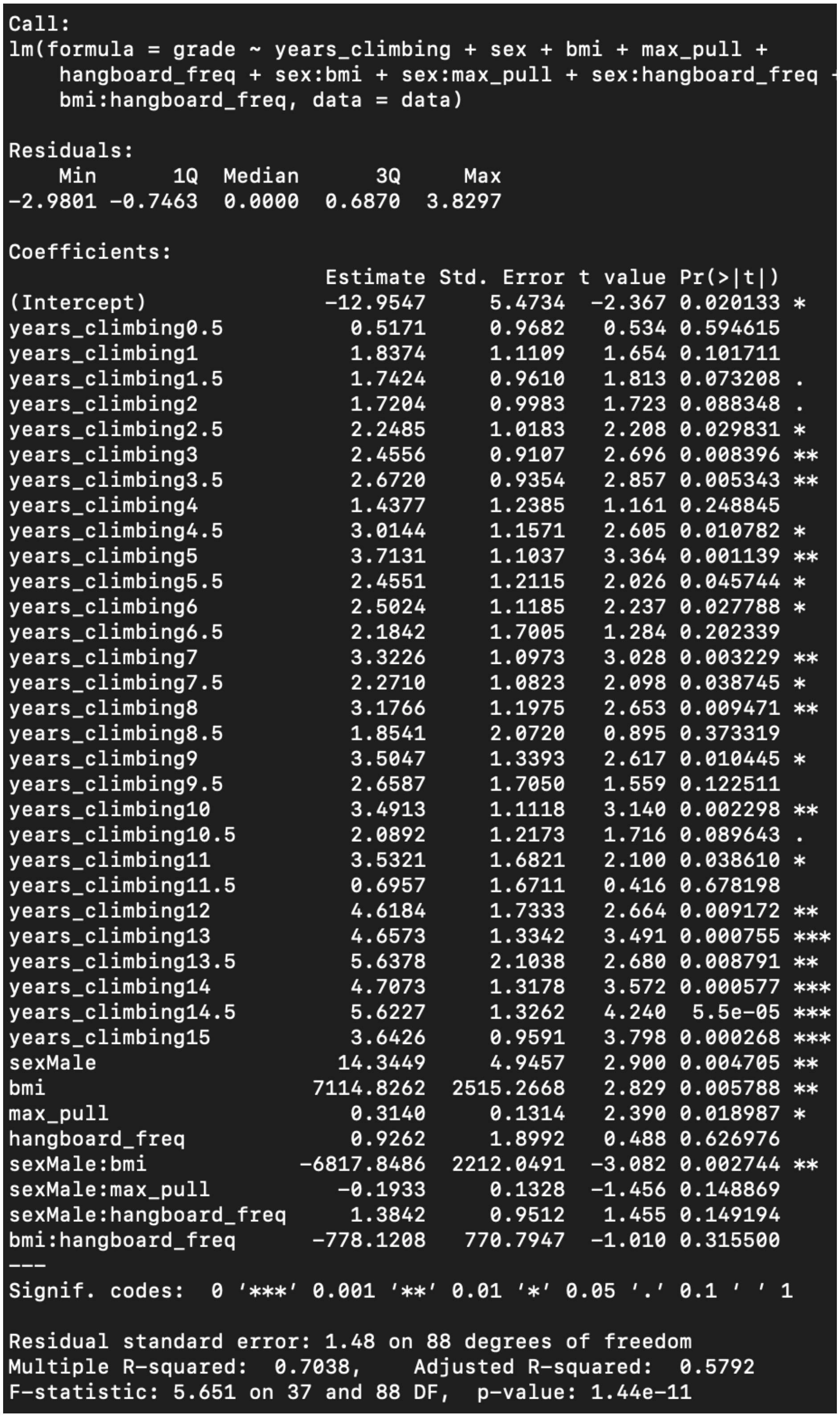
The new model is



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

Residuals:
    Min      1Q  Median      3Q     Max
-3.1624 -0.8159  0.0000  0.7293  3.8279

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -1.341e+01  5.496e+00  -2.439 0.016740 *
years_climbing0.5       4.185e-01  9.740e-01   0.430 0.668482
years_climbing1         1.821e+00  1.112e+00   1.638 0.105019
years_climbing1.5       1.704e+00  9.622e-01   1.771 0.080060 .
years_climbing2         1.595e+00  1.007e+00   1.584 0.116861
years_climbing2.5       2.183e+00  1.021e+00   2.138 0.035312 *
years_climbing3         2.444e+00  9.111e-01   2.682 0.008746 **
years_climbing3.5       2.693e+00  9.360e-01   2.877 0.005042 **
years_climbing4         1.459e+00  1.239e+00   1.177 0.242383
years_climbing4.5       2.956e+00  1.159e+00   2.550 0.012530 *
years_climbing5         3.618e+00  1.109e+00   3.264 0.001573 **
years_climbing5.5       2.563e+00  1.217e+00   2.106 0.038103 *
years_climbing6         2.488e+00  1.119e+00   2.223 0.028781 *
years_climbing6.5       2.161e+00  1.701e+00   1.270 0.207327
years_climbing7         3.286e+00  1.098e+00   2.991 0.003612 **
years_climbing7.5       2.242e+00  1.083e+00   2.070 0.041429 *
years_climbing8         3.129e+00  1.199e+00   2.609 0.010680 *
years_climbing8.5       1.156e+00  2.196e+00   0.526 0.600028
years_climbing9         3.512e+00  1.340e+00   2.621 0.010341 *
years_climbing9.5       2.679e+00  1.706e+00   1.571 0.119882
years_climbing10        3.561e+00  1.115e+00   3.195 0.001950 **
years_climbing10.5      2.046e+00  1.219e+00   1.679 0.096803 .
years_climbing11        3.436e+00  1.686e+00   2.038 0.044569 *
years_climbing11.5      7.260e-01  1.672e+00   0.434 0.665216
years_climbing12        4.247e+00  1.777e+00   2.390 0.018985 *
years_climbing13        4.651e+00  1.335e+00   3.484 0.000775 ***
years_climbing13.5      5.606e+00  2.105e+00   2.663 0.009221 **
years_climbing14        4.804e+00  1.322e+00   3.633 0.000473 ***
years_climbing14.5      5.638e+00  1.327e+00   4.249 5.37e-05 ***
years_climbing15        3.635e+00  9.596e-01   3.788 0.000279 ***
sexMale                 1.363e+01  5.004e+00   2.723 0.007817 **
bmi                     7.206e+03  2.518e+03   2.862 0.005275 **
max_pull                3.498e-01  1.366e-01   2.561 0.012171 *
hangboard_freq          1.665e+00  2.049e+00   0.812 0.418818
sexMale:bmi            -6.685e+03  2.217e+03  -3.015 0.003370 **
sexMale:max_pull       -1.927e-01  1.328e-01  -1.451 0.150460
sexMale:hangboard_freq  1.661e+00  9.942e-01   1.671 0.098363 .
bmi:hangboard_freq     -1.011e+03  8.082e+02  -1.251 0.214306
max_pull:hangboard_freq -2.555e-02  2.655e-02  -0.962 0.338608
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.481 on 87 degrees of freedom
Multiple R-squared:  0.7069,    Adjusted R-squared:  0.5789
F-statistic: 5.522 on 38 and 87 DF,  p-value: 2.373e-11
```

As you can see, the interaction term with the largest p-value is max_pull:hangboard_freq. We proceed by removing this term from the model

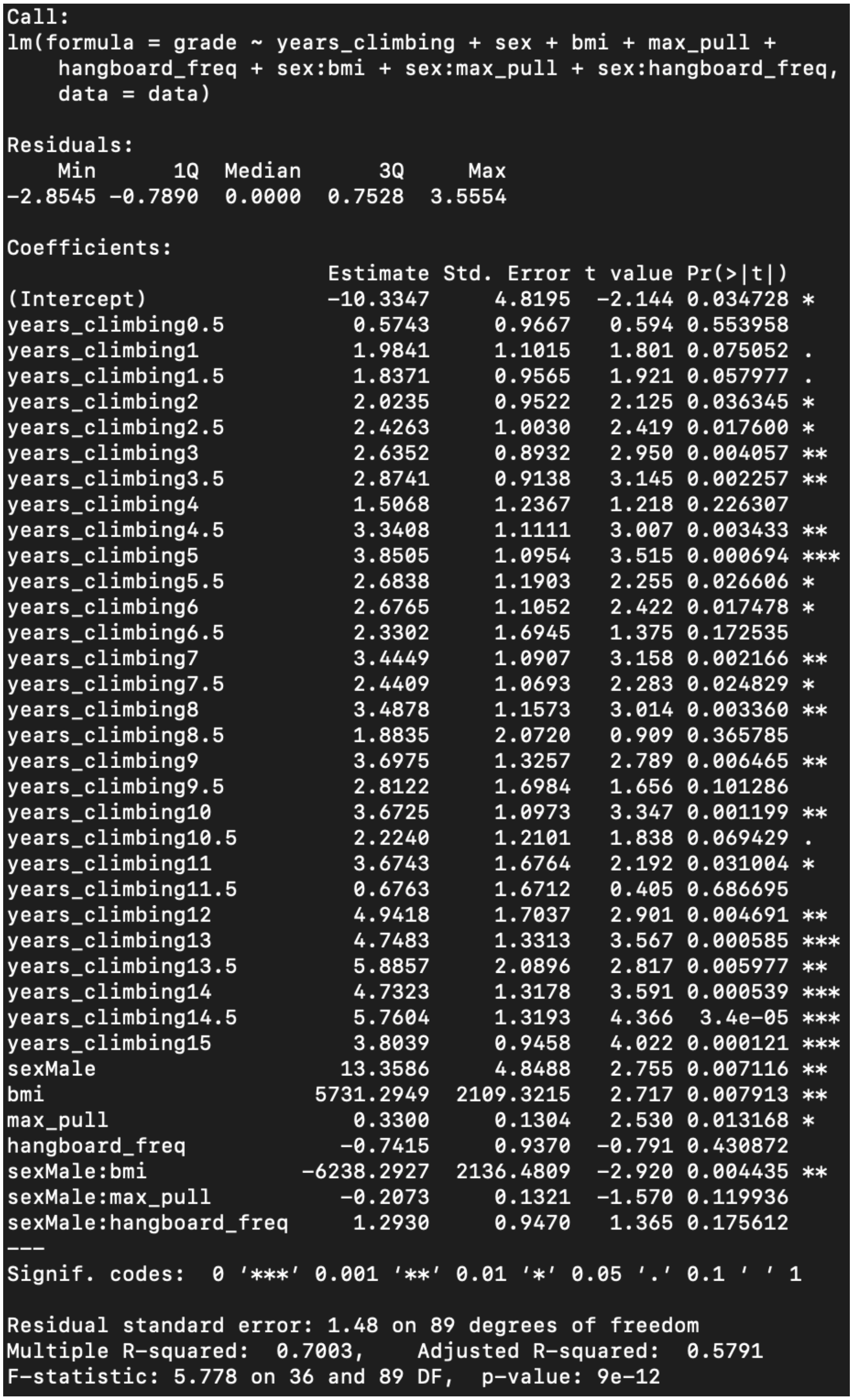
Model Reduction Continued
The new model is



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

Residuals:
    Min      1Q  Median      3Q     Max
-2.9801 -0.7463  0.0000  0.6870  3.8297

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              -12.9547     5.4734  -2.367 0.020133 *
years_climbing0.5          0.5171     0.9682   0.534 0.594615
years_climbing1            1.8374     1.1109   1.654 0.101711
years_climbing1.5          1.7424     0.9610   1.813 0.073208 .
years_climbing2            1.7204     0.9983   1.723 0.088348 .
years_climbing2.5          2.2485     1.0183   2.208 0.029831 *
years_climbing3            2.4556     0.9107   2.696 0.008396 **
years_climbing3.5          2.6720     0.9354   2.857 0.005343 **
years_climbing4            1.4377     1.2385   1.161 0.248845
years_climbing4.5          3.0144     1.1571   2.605 0.010782 *
years_climbing5            3.7131     1.1037   3.364 0.001139 **
years_climbing5.5          2.4551     1.2115   2.026 0.045744 *
years_climbing6            2.5024     1.1185   2.237 0.027788 *
years_climbing6.5          2.1842     1.7005   1.284 0.202339
years_climbing7            3.3226     1.0973   3.028 0.003229 **
years_climbing7.5          2.2710     1.0823   2.098 0.038745 *
years_climbing8            3.1766     1.1975   2.653 0.009471 **
years_climbing8.5          1.8541     2.0720   0.895 0.373319
years_climbing9            3.5047     1.3393   2.617 0.010445 *
years_climbing9.5          2.6587     1.7050   1.559 0.122511
years_climbing10           3.4913     1.1118   3.140 0.002298 **
years_climbing10.5         2.0892     1.2173   1.716 0.089643 .
years_climbing11           3.5321     1.6821   2.100 0.038610 *
years_climbing11.5         0.6957     1.6711   0.416 0.678198
years_climbing12           4.6184     1.7333   2.664 0.009172 **
years_climbing13           4.6573     1.3342   3.491 0.000755 ***
years_climbing13.5         5.6378     2.1038   2.680 0.008791 **
years_climbing14           4.7073     1.3178   3.572 0.000577 ***
years_climbing14.5         5.6227     1.3262   4.240  5.5e-05 ***
years_climbing15           3.6426     0.9591   3.798 0.000268 ***
sexMale                   14.3449     4.9457   2.900 0.004705 **
bmi                     7114.8262  2515.2668   2.829 0.005788 **
max_pull                   0.3140     0.1314   2.390 0.018987 *
hangboard_freq             0.9262     1.8992   0.488 0.626976
sexMale:bmi            -6817.8486  2212.0491  -3.082 0.002744 **
sexMale:max_pull          -0.1933     0.1328  -1.456 0.148869
sexMale:hangboard_freq     1.3842     0.9512   1.455 0.149194
bmi:hangboard_freq      -778.1208   770.7947  -1.010 0.315500
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.48 on 88 degrees of freedom
Multiple R-squared:  0.7038,    Adjusted R-squared:  0.5792
F-statistic: 5.651 on 37 and 88 DF,  p-value: 1.44e-11
```

As you can see, the interaction term with the largest p-value is bmi:hangboard_freq. We proceed by removing this term from the model

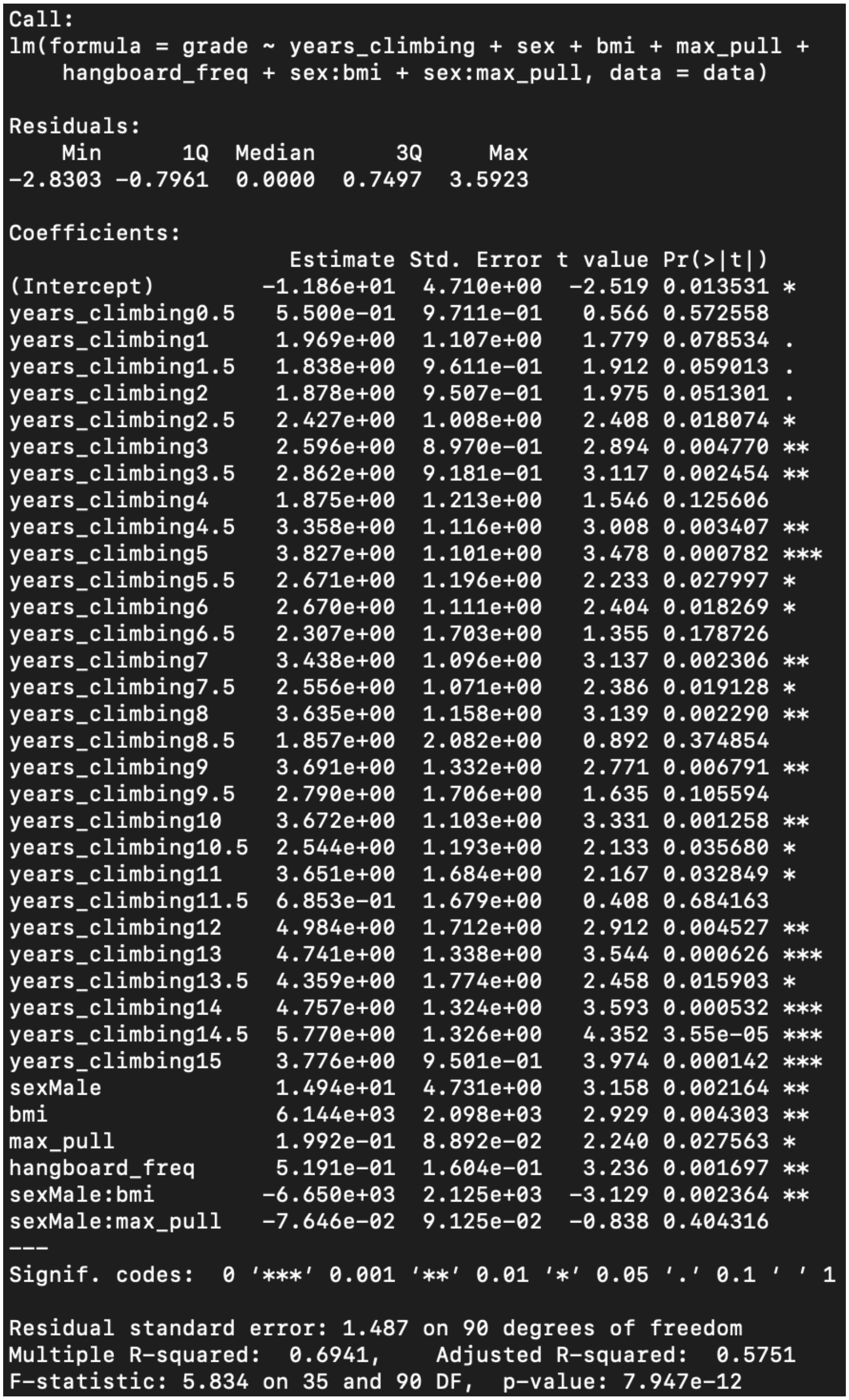
Model Reduction Continued
The new model is



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

Residuals:
    Min      1Q  Median      3Q     Max
-2.8545 -0.7890  0.0000  0.7528  3.5554

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              -10.3347     4.8195  -2.144 0.034728 *
years_climbing0.5          0.5743     0.9667   0.594 0.553958
years_climbing1            1.9841     1.1015   1.801 0.075052 .
years_climbing1.5          1.8371     0.9565   1.921 0.057977 .
years_climbing2            2.0235     0.9522   2.125 0.036345 *
years_climbing2.5          2.4263     1.0030   2.419 0.017600 *
years_climbing3            2.6352     0.8932   2.950 0.004057 **
years_climbing3.5          2.8741     0.9138   3.145 0.002257 **
years_climbing4            1.5068     1.2367   1.218 0.226307
years_climbing4.5          3.3408     1.1111   3.007 0.003433 **
years_climbing5            3.8505     1.0954   3.515 0.000694 ***
years_climbing5.5          2.6838     1.1903   2.255 0.026606 *
years_climbing6            2.6765     1.1052   2.422 0.017478 *
years_climbing6.5          2.3302     1.6945   1.375 0.172535
years_climbing7            3.4449     1.0907   3.158 0.002166 **
years_climbing7.5          2.4409     1.0693   2.283 0.024829 *
years_climbing8            3.4878     1.1573   3.014 0.003360 **
years_climbing8.5          1.8835     2.0720   0.909 0.365785
years_climbing9            3.6975     1.3257   2.789 0.006465 **
years_climbing9.5          2.8122     1.6984   1.656 0.101286
years_climbing10           3.6725     1.0973   3.347 0.001199 **
years_climbing10.5         2.2240     1.2101   1.838 0.069429 .
years_climbing11           3.6743     1.6764   2.192 0.031004 *
years_climbing11.5         0.6763     1.6712   0.405 0.686695
years_climbing12           4.9418     1.7037   2.901 0.004691 **
years_climbing13           4.7483     1.3313   3.567 0.000585 ***
years_climbing13.5         5.8857     2.0896   2.817 0.005977 **
years_climbing14           4.7323     1.3178   3.591 0.000539 ***
years_climbing14.5         5.7604     1.3193   4.366  3.4e-05 ***
years_climbing15           3.8039     0.9458   4.022 0.000121 ***
sexMale                   13.3586     4.8488   2.755 0.007116 **
bmi                     5731.2949  2109.3215   2.717 0.007913 **
max_pull                   0.3300     0.1304   2.530 0.013168 *
hangboard_freq            -0.7415     0.9370  -0.791 0.430872
sexMale:bmi            -6238.2927  2136.4809  -2.920 0.004435 **
sexMale:max_pull          -0.2073     0.1321  -1.570 0.119936
sexMale:hangboard_freq     1.2930     0.9470   1.365 0.175612
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.48 on 89 degrees of freedom
Multiple R-squared:  0.7003,	Adjusted R-squared:  0.5791
F-statistic: 5.778 on 36 and 89 DF,  p-value: 9e-12
```

As you can see, the interaction term with the largest p-value is sexMale:hangboard_freq. We proceed by removing this term from the model

Model Reduction Continued
The new model is

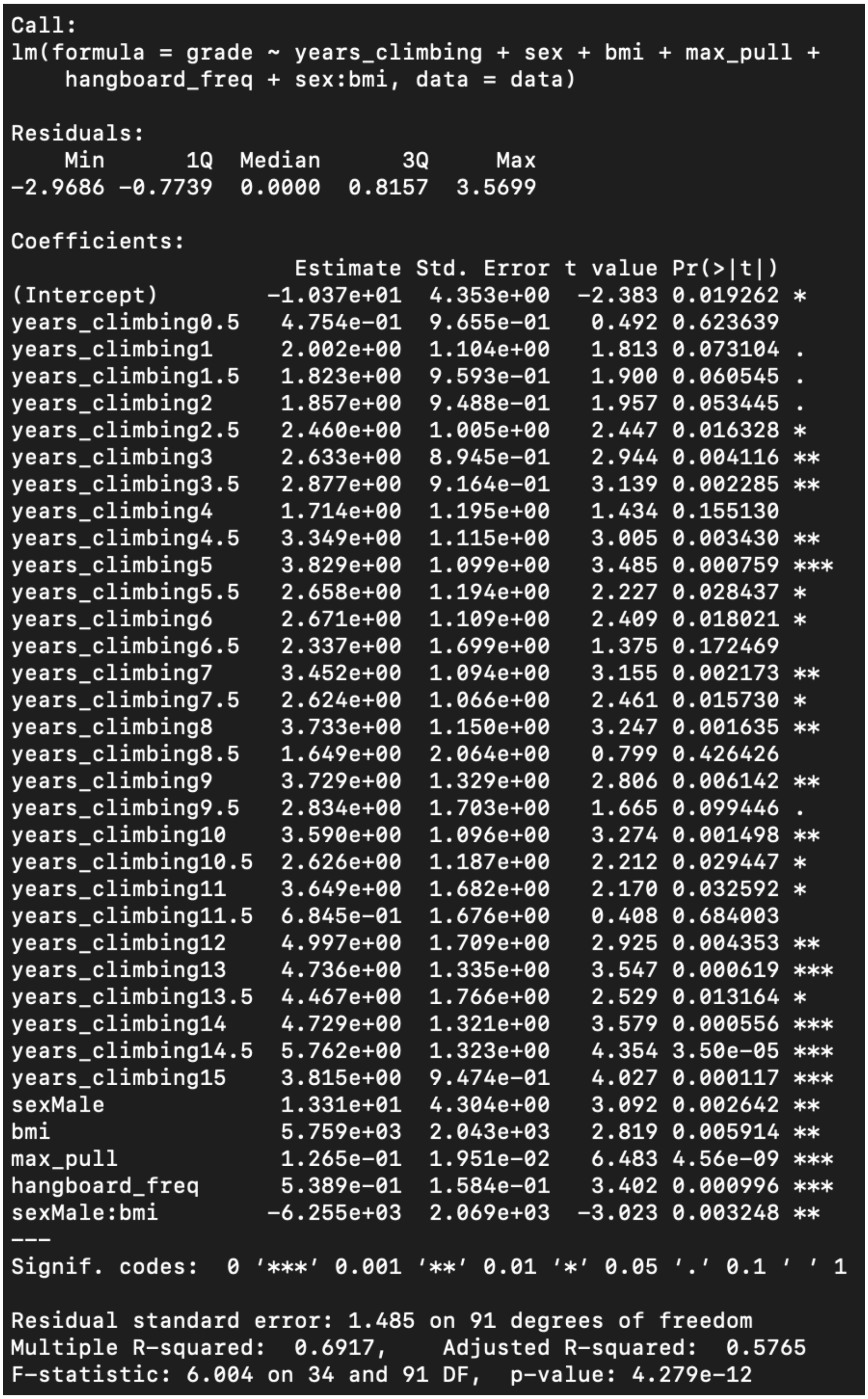


```
Call:
lm(formula = grade ~ years_climbing + sex + bmi + max_pull +
    hangboard_freq + sex:bmi + sex:max_pull, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8303 -0.7961  0.0000  0.7497  3.5923 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -1.186e+01  4.710e+00  -2.519 0.013531 *  
years_climbing0.5   5.500e-01  9.711e-01   0.566 0.572558    
years_climbing1     1.969e+00  1.107e+00   1.779 0.078534 .  
years_climbing1.5   1.838e+00  9.611e-01   1.912 0.059013 .  
years_climbing2     1.878e+00  9.507e-01   1.975 0.051301 .  
years_climbing2.5   2.427e+00  1.008e+00   2.408 0.018074 *  
years_climbing3     2.596e+00  8.970e-01   2.894 0.004770 ** 
years_climbing3.5   2.862e+00  9.181e-01   3.117 0.002454 ** 
years_climbing4     1.875e+00  1.213e+00   1.546 0.125606    
years_climbing4.5   3.358e+00  1.116e+00   3.008 0.003407 ** 
years_climbing5     3.827e+00  1.101e+00   3.478 0.000782 ***
years_climbing5.5   2.671e+00  1.196e+00   2.233 0.027997 *  
years_climbing6     2.670e+00  1.111e+00   2.404 0.018269 *  
years_climbing6.5   2.307e+00  1.703e+00   1.355 0.178726    
years_climbing7     3.438e+00  1.096e+00   3.137 0.002306 ** 
years_climbing7.5   2.556e+00  1.071e+00   2.386 0.019128 *  
years_climbing8     3.635e+00  1.158e+00   3.139 0.002290 ** 
years_climbing8.5   1.857e+00  2.082e+00   0.892 0.374854    
years_climbing9     3.691e+00  1.332e+00   2.771 0.006791 ** 
years_climbing9.5   2.790e+00  1.706e+00   1.635 0.105594    
years_climbing10    3.672e+00  1.103e+00   3.331 0.001258 ** 
years_climbing10.5  2.544e+00  1.193e+00   2.133 0.035680 *  
years_climbing11    3.651e+00  1.684e+00   2.167 0.032849 *  
years_climbing11.5  6.853e-01  1.679e+00   0.408 0.684163    
years_climbing12    4.984e+00  1.712e+00   2.912 0.004527 ** 
years_climbing13    4.741e+00  1.338e+00   3.544 0.000626 ***
years_climbing13.5  4.359e+00  1.774e+00   2.458 0.015903 *  
years_climbing14    4.757e+00  1.324e+00   3.593 0.000532 ***
years_climbing14.5  5.770e+00  1.326e+00   4.352 3.55e-05 ***
years_climbing15    3.776e+00  9.501e-01   3.974 0.000142 ***
sexMale             1.494e+01  4.731e+00   3.158 0.002164 ** 
bmi                 6.144e+03  2.098e+03   2.929 0.004303 ** 
max_pull            1.992e-01  8.892e-02   2.240 0.027563 *  
hangboard_freq      5.191e-01  1.604e-01   3.236 0.001697 ** 
sexMale:bmi        -6.650e+03  2.125e+03  -3.129 0.002364 ** 
sexMale:max_pull   -7.646e-02  9.125e-02  -0.838 0.404316    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.487 on 90 degrees of freedom
Multiple R-squared:  0.6941,	Adjusted R-squared:  0.5751 
F-statistic: 5.834 on 35 and 90 DF,  p-value: 7.947e-12
```

As you can see, the interaction term with the largest p-value is sexMale:max_pull. We proceed by removing this term from the model

Model Reduction Continued

The new model is

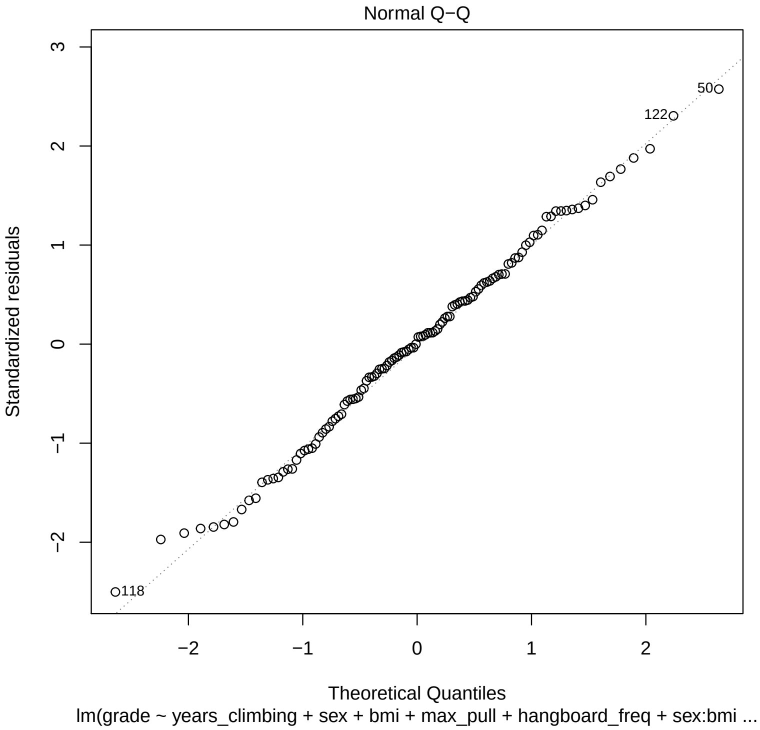


```
Call:
lm(formula = grade ~ years_climbing + sex + bmi + max_pull +
    hangboard_freq + sex:bmi, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9686 -0.7739  0.0000  0.8157  3.5699 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -1.037e+01  4.353e+00  -2.383 0.019262 *  
years_climbing0.5   4.754e-01  9.655e-01   0.492 0.623639    
years_climbing1     2.002e+00  1.104e+00   1.813 0.073104 .  
years_climbing1.5   1.823e+00  9.593e-01   1.900 0.060545 .  
years_climbing2     1.857e+00  9.488e-01   1.957 0.053445 .  
years_climbing2.5   2.460e+00  1.005e+00   2.447 0.016328 *  
years_climbing3     2.633e+00  8.945e-01   2.944 0.004116 ** 
years_climbing3.5   2.877e+00  9.164e-01   3.139 0.002285 ** 
years_climbing4     1.714e+00  1.195e+00   1.434 0.155130    
years_climbing4.5   3.349e+00  1.115e+00   3.005 0.003430 ** 
years_climbing5     3.829e+00  1.099e+00   3.485 0.000759 ***
years_climbing5.5   2.658e+00  1.194e+00   2.227 0.028437 *  
years_climbing6     2.671e+00  1.109e+00   2.409 0.018021 *  
years_climbing6.5   2.337e+00  1.699e+00   1.375 0.172469    
years_climbing7     3.452e+00  1.094e+00   3.155 0.002173 ** 
years_climbing7.5   2.624e+00  1.066e+00   2.461 0.015730 *  
years_climbing8     3.733e+00  1.150e+00   3.247 0.001635 ** 
years_climbing8.5   1.649e+00  2.064e+00   0.799 0.426426    
years_climbing9     3.729e+00  1.329e+00   2.806 0.006142 ** 
years_climbing9.5   2.834e+00  1.703e+00   1.665 0.099446 .  
years_climbing10    3.590e+00  1.096e+00   3.274 0.001498 ** 
years_climbing10.5  2.626e+00  1.187e+00   2.212 0.029447 *  
years_climbing11    3.649e+00  1.682e+00   2.170 0.032592 *  
years_climbing11.5  6.845e-01  1.676e+00   0.408 0.684003    
years_climbing12    4.997e+00  1.709e+00   2.925 0.004353 ** 
years_climbing13    4.736e+00  1.335e+00   3.547 0.000619 ***
years_climbing13.5  4.467e+00  1.766e+00   2.529 0.013164 *  
years_climbing14    4.729e+00  1.321e+00   3.579 0.000556 ***
years_climbing14.5  5.762e+00  1.323e+00   4.354 3.50e-05 ***
years_climbing15    3.815e+00  9.474e-01   4.027 0.000117 ***
sexMale             1.331e+01  4.304e+00   3.092 0.002642 ** 
bmi                 5.759e+03  2.043e+03   2.819 0.005914 ** 
max_pull            1.265e-01  1.951e-02   6.483 4.56e-09 ***
hangboard_freq      5.389e-01  1.584e-01   3.402 0.000996 ***
sexMale:bmi        -6.255e+03  2.069e+03  -3.023 0.003248 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.485 on 91 degrees of freedom
Multiple R-squared:  0.6917,	Adjusted R-squared:  0.5765 
F-statistic: 6.004 on 34 and 91 DF,  p-value: 4.279e-12
```

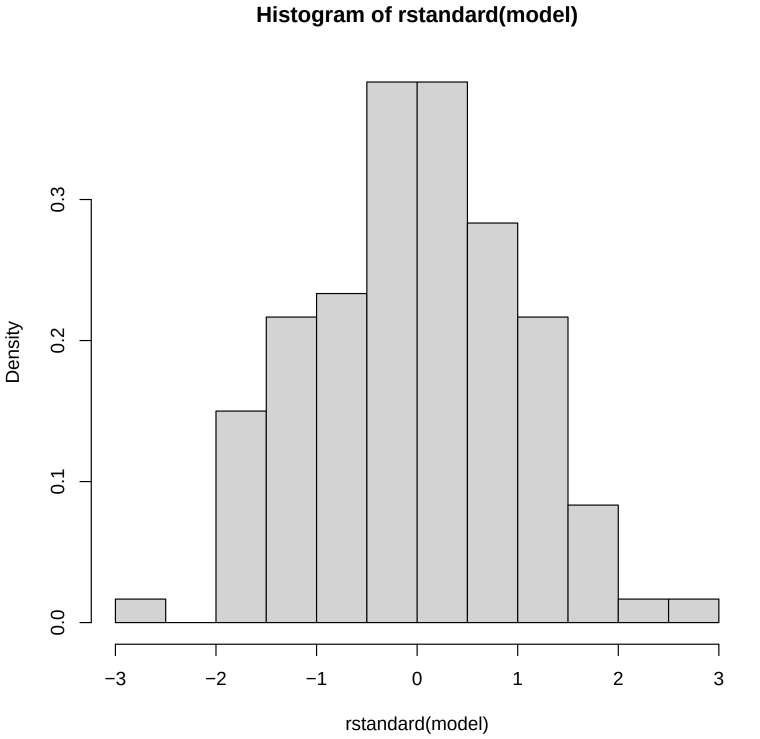
This concludes our systematic reduction of interaction terms as specified by the assignment instructions.

Normality of Residuals

The normal q-q plot, rather nicely, supports our assumption that the residuals are from a normally distributed population



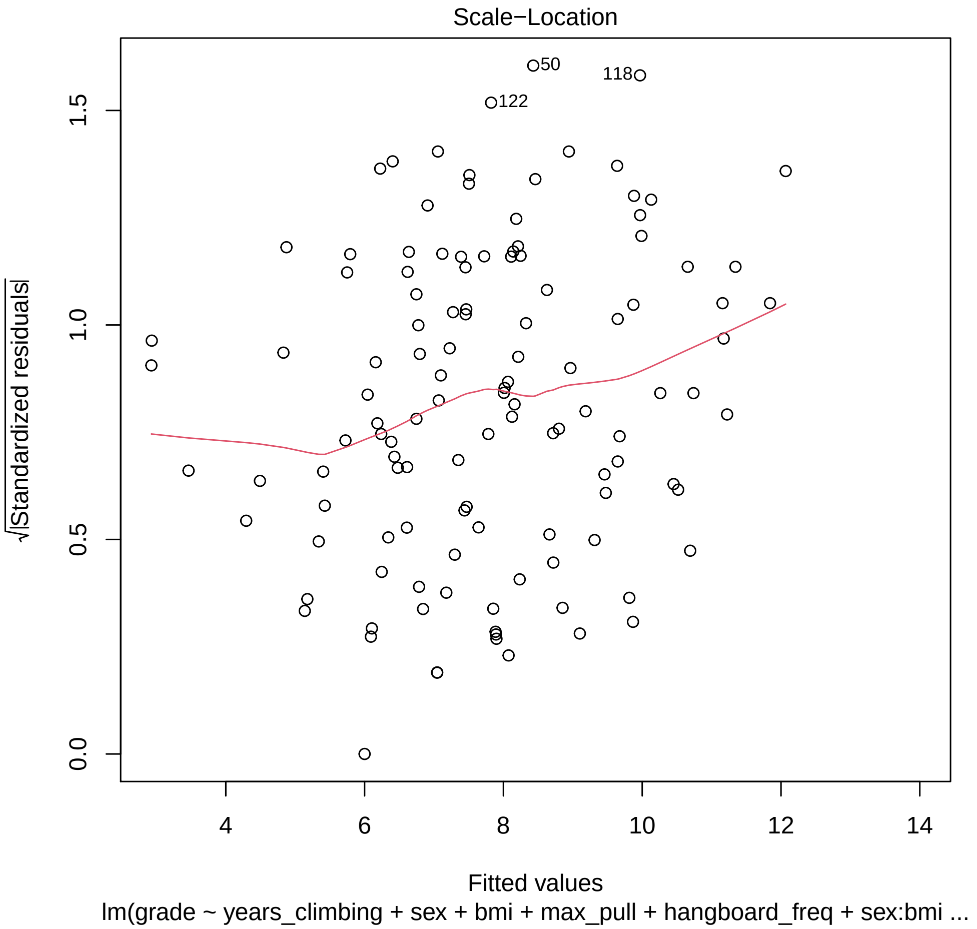
This can also be seen from viewing the histogram



For further confirmation we find a p-value of 0.9083 in a Shapiro-Wilkes test.

Linearity and Constant Variance of Residuals

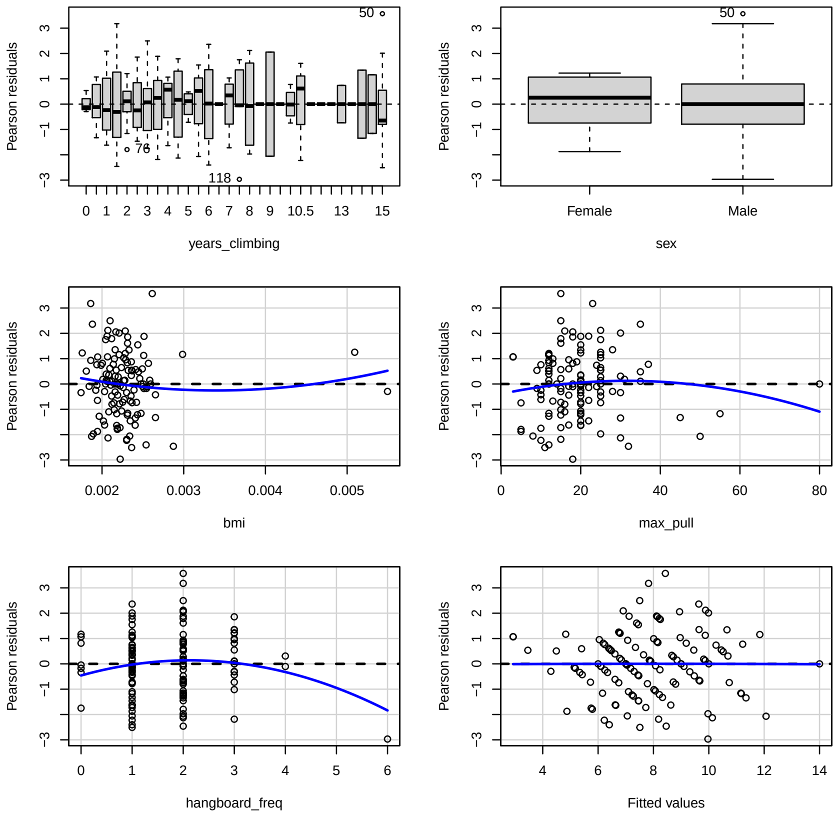
First observe the scale-location plot

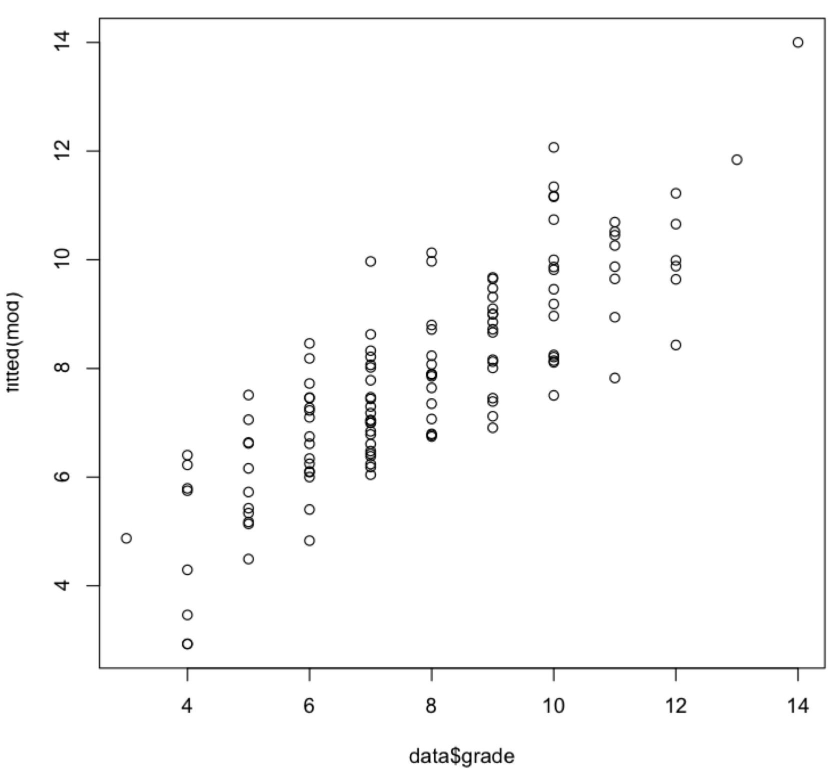


There is nothing to concerning with this graph, as displayed by the lack of an obvious pattern. It should be noted that the trace line increases slightly, which may imply that we are in violation of our constant variance assumption. But when viewing the non-standardized residuals verse fitted values graph that is shown on the next page, there is not an apparent funnel shape to our data. Therefore, we conclude that our homoscedasticity assumption holds.

Linearity and Constant Variance of Residuals Continued

Next we asses our linearity assumption.





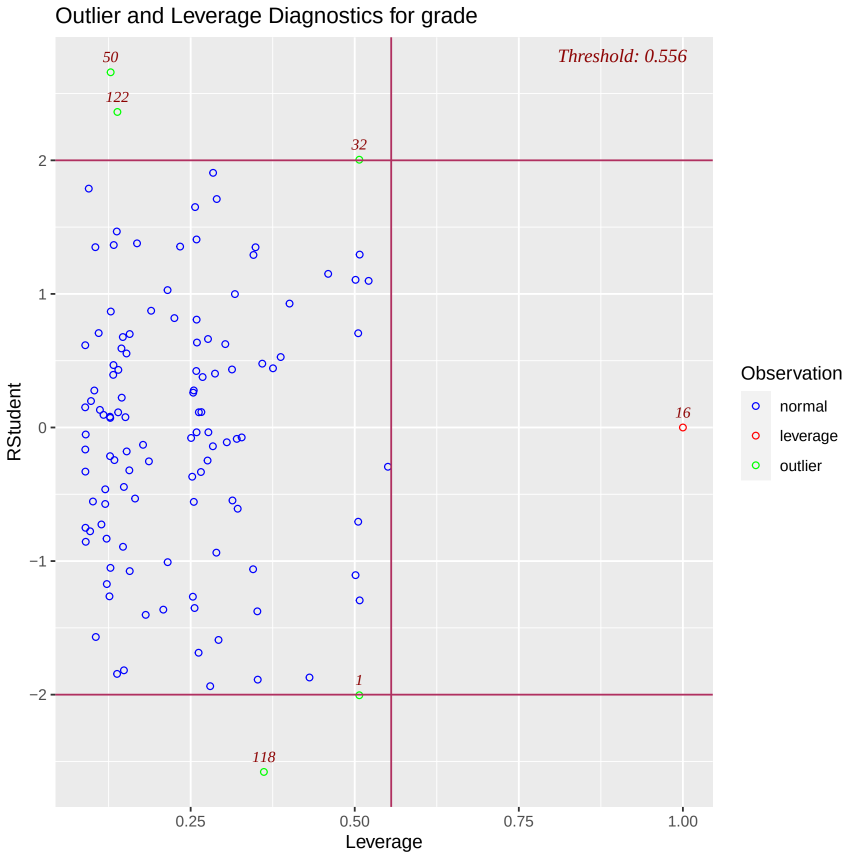
First, when looking at the residuals verse years climbing, the median observation for the residuals over the levels corresponding to climbers with less than 4 years of experience tend to be negative. This may be an indication that the relationship between hardest grade and years climbing do not form a linear relationship. I make this conclusion not only based upon the graph, but also because it makes sense when taking into account a basic learning theory model which would suggest that this factor be best modeled logarithmically.

Next it appears there may be an issue with the max pull predictor. It may be overestimating the climbing grade for those who can manage to do 40 or more pull ups. As numerically supported by the curvature test, which reports a p-value of 0.012, which suggests nonlinearity. Next when looking at the weekly hangboard frequency as a predictor, there is an apparent outlier in this category that is influencing the quadratic line of fit. The curvature test also supports a p-value of 0.05521, which may imply that there is more to be I would suspect this individual may be over training and it may be why he is climbing at a level far under the models mean predicted grade.

Lastly, the residuals verse fitted values do not indicate an apparent violation of our linearity assumption. Nor does the plot of fitted values verse the observed values. However, our model may be improved by addressing some of the beforehand items. So my conclusion is that our linear assumption is indeed suspect of violation. Further discussion of how to deal with a potential linearity violation in our use of hangboard frequency occurs below.

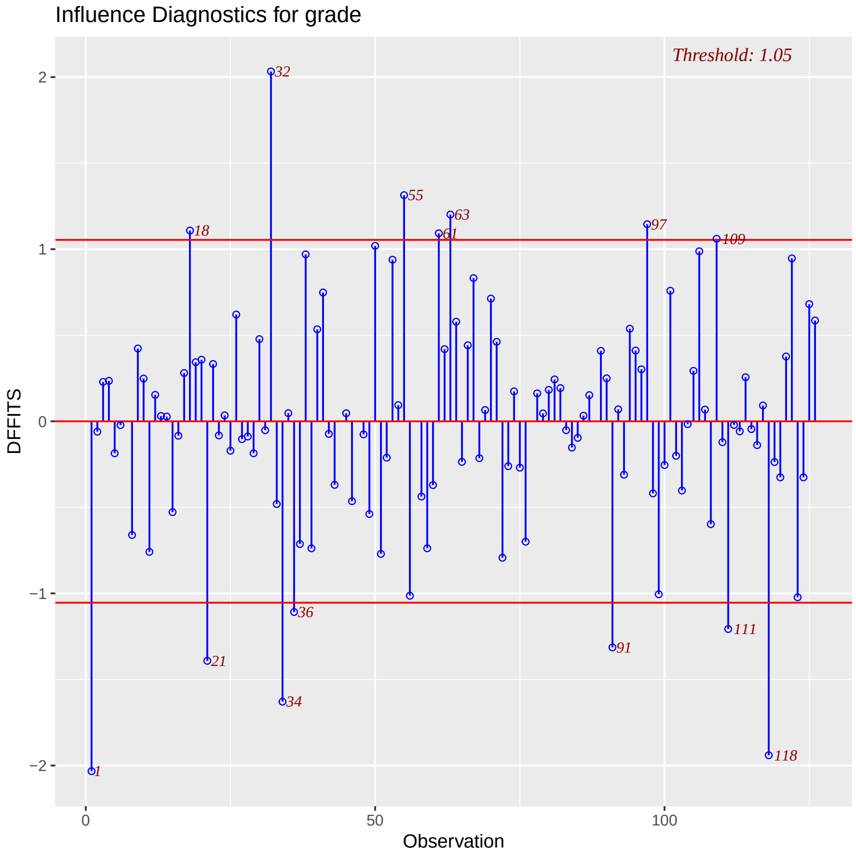
Influential Observations

When examining out residuals verse leverage plot, there are no obvious influential observations. Certainly some outliers, but there are no outliers with high leverage.



On the next page we look at the DFFITS plot. We will use the rule that any DFFITS$_i > 2 \cdot \sqrt{((p+1)/(n-p-1)} = 1.271997$ should be labeled as influential.

Influential Observations Continued



There are a few influential points under this rule. Specifically observations $1, 21, 34, 91, 118$ and 32.
Due to all the categories in years climbing the DFBETA index plots show an enormous amount of information and not much of it is meaningful. There was one observation that was deemed influential by the threshold line created from the dfbetaPlots in the car package. Specifically, it was observation 97 that influenced the weekly hangboard frequency coefficient drastically and this was the same outlier that was identified in our linearity discussion. However, the observation also happened to be a very elite climber and there where very few of those so consequently it impacted the hangboard frequency coefficient.

This is a symptom in of my underlying model. In my opinion there are few possible solutions:

(1) Blame the messy data. Ideally, we could collect a sample that would contain more elite climbers.
(2) Model weekly hangboard frequency as a categorical predictor instead of a numerical variable.
(3) Remove the observation.

The first option is the easy way out. Then for the second, when replacing it as a categorical factor in our model and performing an F-test yields a p-value of 0.13 (adjusted R^2 decreases to 0.597). These results imply that our current treatment of hang board frequency should remain as is and further investigation into a suitable transformation should be performed (either on the numeric or categorical data). As for removal, there are no obvious inaccuracy's in observation 97. One could even argue that due to observation 97's level of climbing, his responses are likely the most accurate. If the goal of the model is to be used for predicting and training purposes then the best solution is to remove the observation. However, if we are trying to make a statistical inference on the population it is likely best to leave as is. Though that opens an entirely new discussion as to what our population is, if it is of all climbers then the chances observing a climber of this level in our sample size is quite small. Hence, an argument still may be made for removal.

The model will remain as is.

Interpretation of Interaction Terms

The only interaction term in my model is between the numeric bmi and the categorical sex. However, we also include sex and bmi independently so it is a ”different slopes, different intercepts” model. The variable bmi was calculated from the reported height and weight values. R decided to have Female as the control for the sex dummy term in the model. The coefficient on bmi refers to the slope for female climbers. Then the interaction term between bmi and sex effectively reduces the bmi coefficient (by a very small amount) when a male climber is being inputed into the model But the presence of the male climber changes the intercept of the model. Apart from this, there really isn't a meaningful way to describe the coefficient between the interaction of male and female.

Linear Hypothesis Test of Years Climbing

Given so many levels to the categorical variable of years climbing, our model may benefit from collapsing some of the categories. This may also eliminate all of the NA coefficients that appeared when I tried to make this variable interact with others in my model. Which may allow us to have years climbing interact with other terms and lead to a much better model. This will be investigated in the next submission. Since the majority of the NA's appeared in the levels of experience greater than 5.5 years, I did a crude test to see if collapsing everyone who has been climbing for 5.5-15+ years into one category would suffice. The p-value with this test is 0.2686, so we will keep the original model for now.