Amber Swain

1) Let the shucked weight be a response variable and all other variables as explanatory variables. Use multiple linear regression to obtain an equation for the shucked weight. Report w*.

w* = [-0.015580, 0.003618, 0.005630, 0.144493, 0.045869, 0.018446, 0.700702, -0.429031, -0.697786, -0.006225]

<u>Equation:</u>

Shucked Weight = -0.015580 + 0.003618*SexI + 0.005630*SexM + 0.144493*Length + 0.045869*Diameter + 0.018446*Height + 0.700702*WholeWeight + (-0.429031)*VisceraWeight + (-0.697786)*ShellWeight + (-0.006225)*Rings

For the Sexi and SexM columns, if Sex is I, then SexI would be 1 and SexM would be 0. If Sex is M, then SexI would be 0 and SexM would be 1. If Sex is F, both SexI and SexM would be 0. Since the Sex is a categorical variable, R treats categorical variables as factors, and when you include a categorical variable with multiple levels in a linear regression model, R automatically creates dummy variables for each level.

Output from lm() command:

Coefficients:

- Intercept: -0.015580
- SexI: 0.003618
- SexM: 0.005630
- Length: 0.144493
- Diameter: 0.045869
- Height: 0.018446
- Whole Weight: 0.700702
- Viscera Weight: -0.429031
- Shell Weight: -0.697786
- Rings: -0.006225

2) Use adjusted R^2 to remove unnecessary variables (if any). Report w*_1 - a final vector of coefficient. Discuss.

Adjusted R squared: 0.969218777222254

What We Removed	New Adjusted R Squared	Should It Be Removed(Y/N)
Sex	0.969126024963765	No
Length	0.969075893301744	No
Diameter	0.969216274443189	No
Height	0.969222822438314	Yes
Whole Weight	0.902108650741001	No
Viscera Weight	0.966460677925198	No
Shell Weight	0.956716543608706	No
Rings	0.964906824102078	No

New R^2 With Removing Unnecessary Variables: 0.969222822438314

New w1* without Height = [-0.015062, 0.003543, 0.005607, 0.145200, 0.048474, 0.700748, -0.428203, -0.696632, -0.006205]

3) Use multiple linear regression with Lasso regularization to obtain another equation for the shucked weight (use any available tool you want, for example glmnet() and cv.glmnet() for cross-validation part) Report w*_2 in this case. Discuss.

For this problem I used this article/website as guidance:
https://www.statology.org/lasso-regression-in-r/

w2* = [-0.013953132, 0.002748447, 0.125744177, 0.061478795, 0.006623031, 0.679744775, -0.373466288, -0.658569684, -0.006322613]

(Intercept) -0.013953132
Sex 0.002748447
Length 0.125744177
Diameter 0.061478795

Height 0.006623031

Whole weight 0.679744775

Viscera weight -0.373466288

Shell weight -0.658569684

Rings -0.006322613

4) Compare two vectors w*_1 and w*_2. Discuss.

Variable	**w1***	**w2***
Intercept	-0.015062	-0.013953132
Sex	0.003543 = SexI, 0.005607 = SexM	0.002748447
Length	0.145200	0.125744177
Height	NA/0	0.061478795
Diameter	0.048474	0.006623031
Whole Weight	0.700748	0.679744775
Viscera Weight	-0.428203	-0.373466288
Shell Weight	-0.696632	-0.658569684
Rings	-0.006205	-0.006322613

All of the coefficients in w2* either increased or decreased closer to zero. They all kept the same sign in both vectors, so the ones that were negative in w1*, stayed negative in w2* and therefore increased in value by getting closer to zero, and the ones that were positive in w1*, stayed positive in w2* and therefore decreased in value by getting closer to zero.

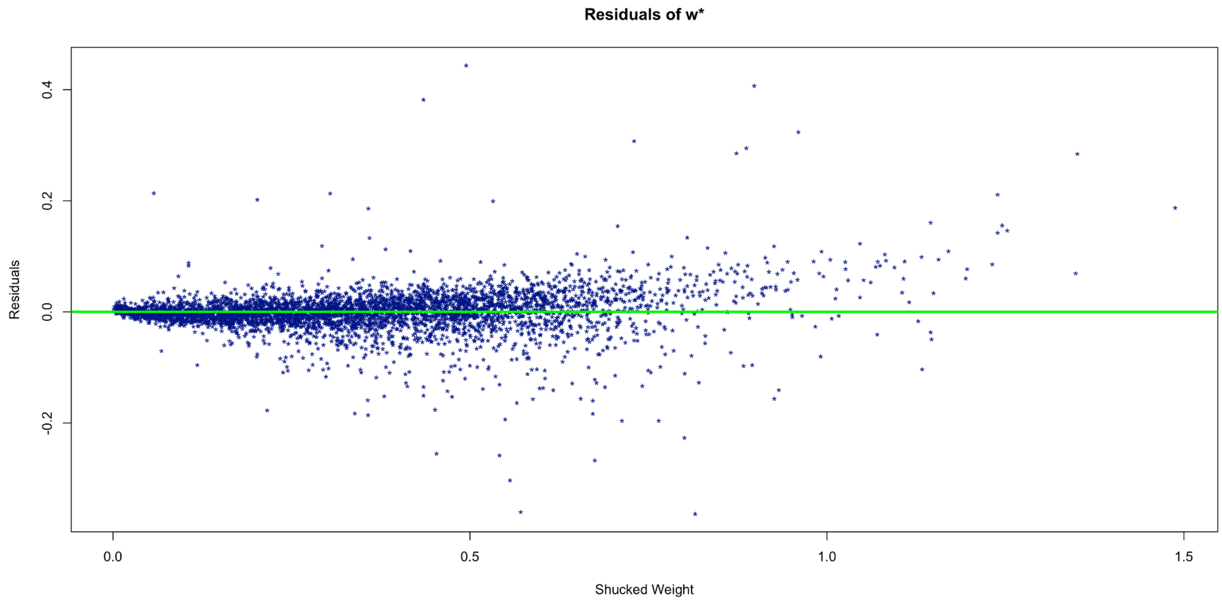
5) Repost MSE on the whole training data for both equations. Compare and discuss.

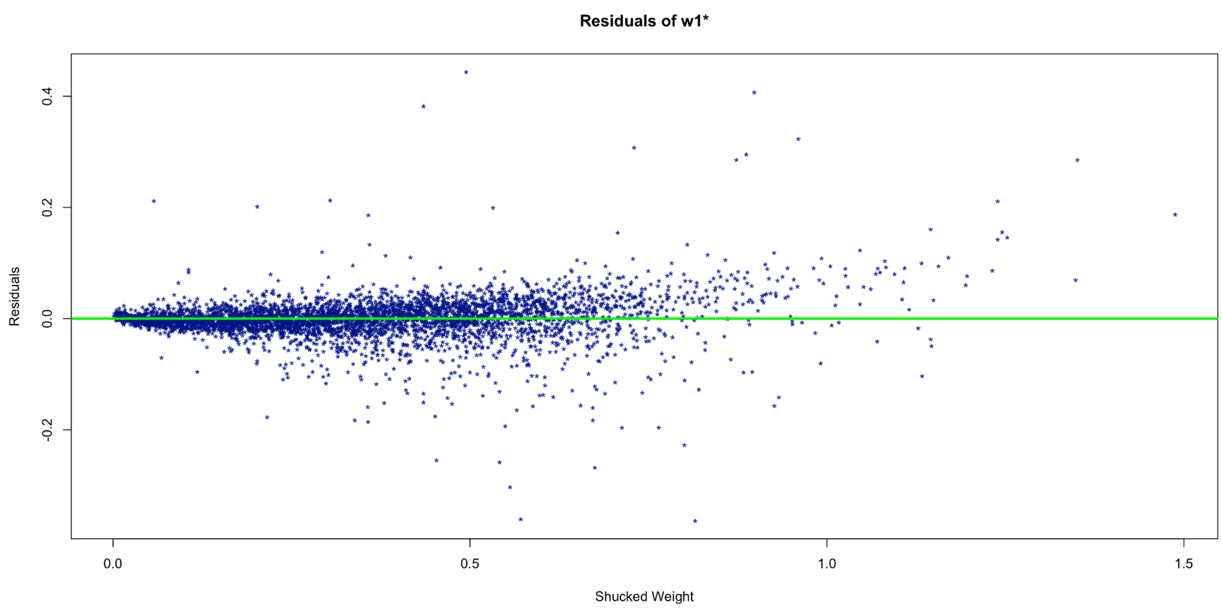
MSE for w1* = 0.001513049

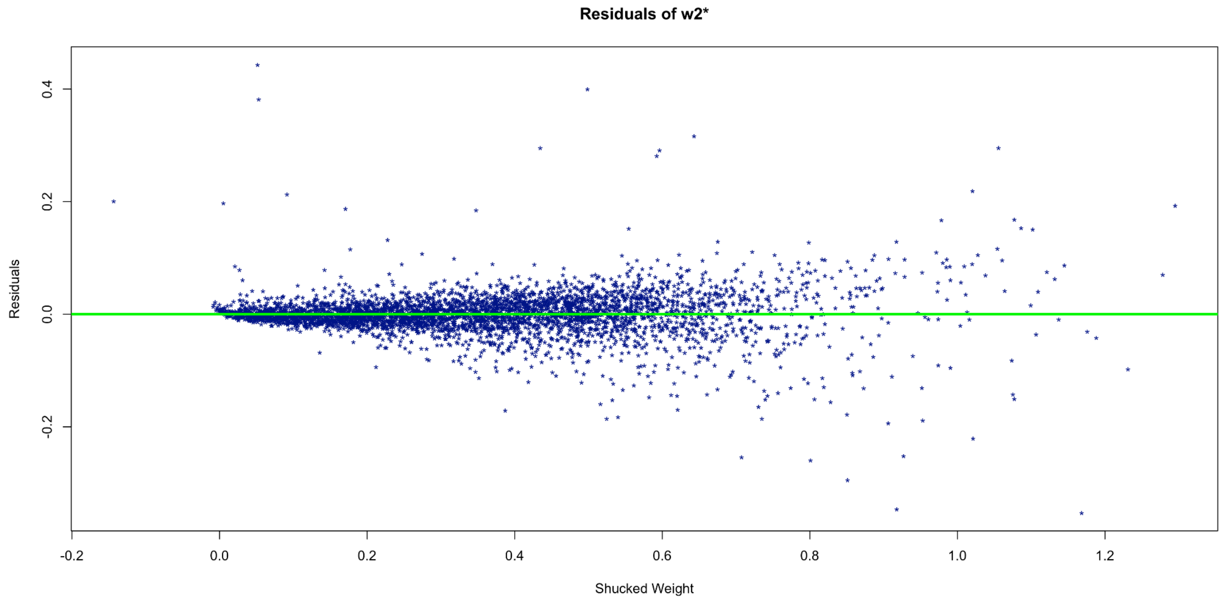
MSE for w2* = 0.001516864

They are almost the same, with w2* being 0.000003815 greater than w1* MSE.

6) Output the residual plots for all w: w*, w*_1 and w*_2. Discuss.






Residuals of w2*
Residuals
0.4
0.2
0.0
-0.2
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Shucked Weight