Horse Mussel Measurements

Arden Miller

Executive Summary

This report investigates using regression models to relate the edible mass M of horse mussels to four measurements made on their shells. The shell measurements were length (L), width (W), height (H), and shell mass (S). It was found that only two of these measurements are required in such a model although several different pairs produce models that work approximately as well as each other. One such model uses height and shell mass:

$$\widehat{\mathrm{E}(\mathsf{M})} = \exp(-3.742 + .09306 \times \mathsf{H} - .000332 \times \mathsf{H}^2 + .00337 \times \mathsf{S}).$$

The precision of predictions made using this model is evaluated.

Horse Mussel Data

The purpose of this investigation was to model the mass of the edible part of horse mussels as a function of other measurements related to the size of the mussel. The data used to create the model consists of measurements made on horse mussels sampled from the Marlborough Sounds. A summary of the measured variables is given in the following table:

Variable	Mean	Range
M: mass of the mussel's muscle (g)	20.88	1–52
L: length of mussel's shell in (mm)	242.6	132–331
W: width of mussel's shell in (mm)	41.73	20–68
H: height of mussel's shell in (mm)	114.3	65–158
S: mass of the mussel's shell (g)	124	10–359

Figure 1 contains pairwise scatter-plots of these measurements.

1 Modelling the Mussel's Mass

The purpose of this investigation was to explore how the mass of the edible part of a horse mussel (M) is related to measurements made on the mussel's shell. The pairs plot in Figure 1

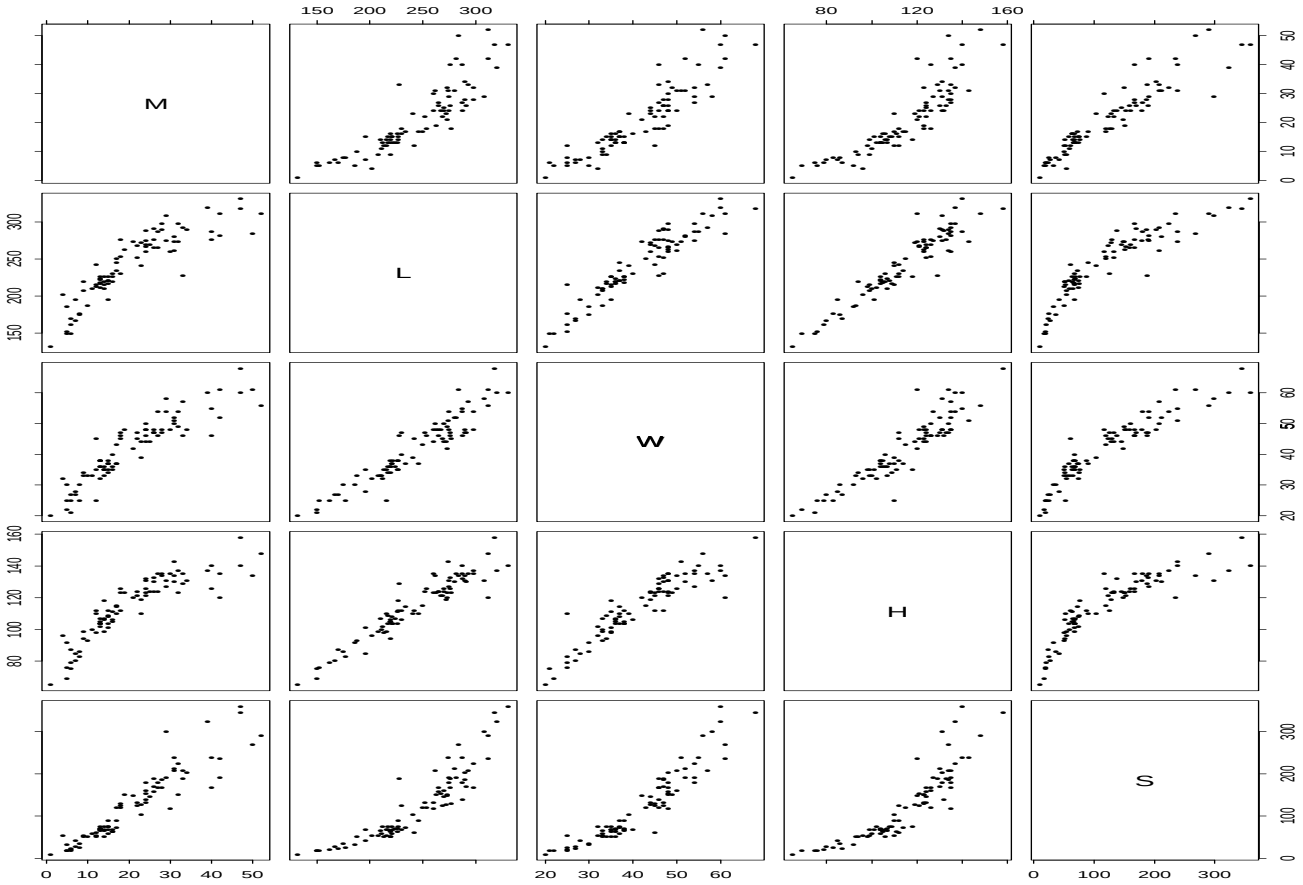

Figure 1: Pairwise Scatterplots of Horse Mussel Measurements.

indicates that for each of the four explanatory variables (L, W, H, and S) considered individually, there is a strong positive relationship between M and each variable. Figure 1 also shows that the four explanatory variables are strongly related to each other. This suggests that there may be a substantial overlap in the information that these variables provide about M.

Multiple linear regression models were used to explore which combinations of the explanatory variables could provide the best explanation of M. It was found that a number of different combinations produced models that worked approximately as well as each other. Most of these models only required two of the explanatory variables.

I identified the following model as working quite well:

$$\log \widehat{\mathrm{E}(\mathsf{M})} = -3.742 + .09306 \times \mathsf{H} - .000332 \times \mathsf{H}^2 + .00337 \times \mathsf{S}.$$

$\widehat{\mathrm{E}(\mathsf{M})}$ represents the predicted average edible mass for all mussels with the specified values of H and S. This model can also be written as:

$$\widehat{\mathrm{E}(\mathsf{M})} = \exp(-3.742 + .09306 \times \mathsf{H} - .000332 \times \mathsf{H}^2 + .00337 \times \mathsf{S}).$$

Figure 2 indicates how this model relates the average edible mass to shell height and shell mass. In the first plot shell mass was fixed at its mean value and the predicted mussel mass is plotted over the range of shell height. This plot indicates that $\hat{\mathsf{M}}$ increases with H until H reaches about 140mm and then decreases. For the second plot shell height was fixed at its mean value

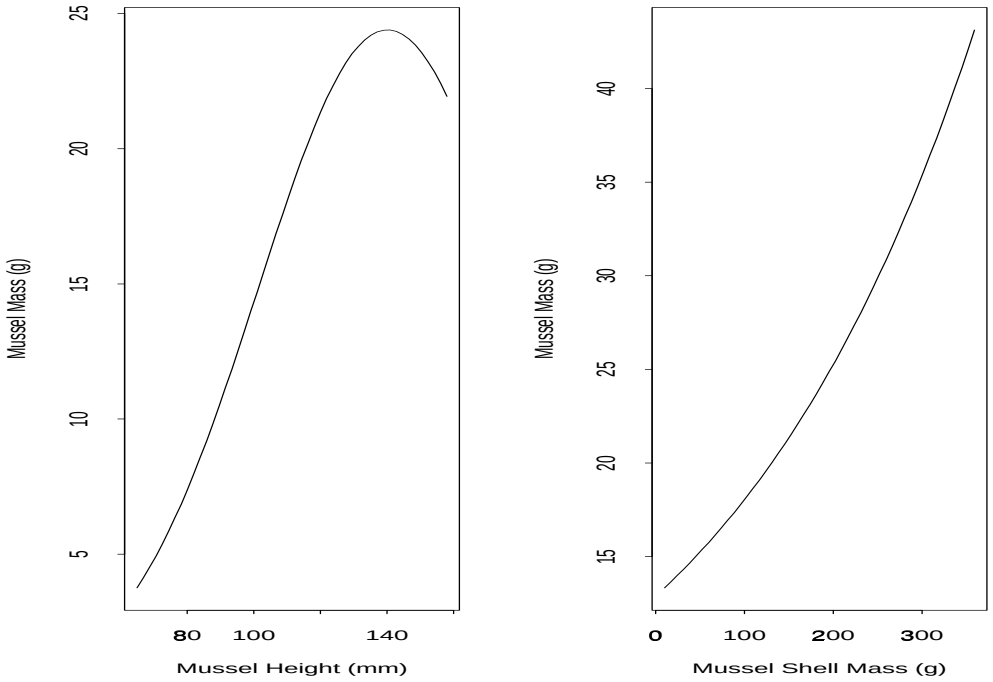

Figure 2: Pairwise Scatterplots of Horse Mussel Measurements.

and the predicted mussel mass is plotted over the range of shell mass. This plot shows that $\hat{\mathsf{M}}$ increases as S increases. The rate of increase gets larger as S increases.

To evaluate the precision of predictions made using this model, I considered predicting the average amount of edible mass for three scenarios: a small mussel, an average sized mussel, and a large mussel. The small mussel scenario uses the minimum values of H and S from the data and the average and large scenarios use the mean values and maximum values respectively. The following table contains the values of H and S used for each scenario, the predicted average edible mussel mass ($\widehat{\mathrm{E}(\mathsf{M})}$), and a 95% confidence interval for the true average edible mussel mass. The confidence interval can be interpreted as having a 95% probability of containing the true average edible mussel mass for all mussels with the specified combination of H and S.

Scenario	H	S	$\widehat{\mathrm{E}(\mathsf{M})}$	interval
small	65	10	2.55	2.02 to 3.22
average	114.3	124.0	19.56	18.2 to 21.0
large	158	359	48.4	37.3 to 62.1

Statistical Appendix

The pairs plot for this data (see Figure 1) indicates that the 4 explanatory variables are clearly related to each other. This was confirmed by finding the VIF's which were > 10 for L, W, H and

approximately 8 for S. These values indicates strong multicollinearity among the regressors. As a result it is unlikely that all 4 of the explanatory variables will be needed in the model for M.

I started by fitting the basic multiple regression model that used M as the response and just linear terms for the four regressors (L, W, H, S). The plot of residuals versus fitted values for this model indicated a clear funnel effect which was confirmed by the plots from the funnel command. I decided to try transforming the response. The output from funnel indicated that a Box-Cox transformation with $p \approx 0.3$. I tried several transformations using p from .5 to 0 (log). All of these seemed to alleviate the funnel problem. I decided to use the log transformation on the basis that this makes the model easier to explain.

For log(M) as the response, I fitted the full model that contained linear terms for all four regressors. The T-tests indicated that one or more of L, W, and S could be dropped from the model (only H is clearly significant). I used all.poss.regs to identify the best subset models. The output indicates that either 2 (H and L) or 3 regressors (H, L and W) should be used. Fitting the three variable indicates that W can be dropped (P-value from t-test is .16). I also looked at partial plots and ace plots for the full model which indicated that a polynomial model may be required for one or more of the regressors. I investigated adding a quadratic model terms for the explanatory variables one at a time. I found that a number of models that consisted of a quadratic polynomial in one of the regressors plus a linear term for one other regressor performed approximately as well as each other. I selected the model that used H, H^2, and S since it had a slightly higher R^2 that the other models of this type. The output for summary for this model is:

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   2.4330   0.1034    23.5381   0.0000
poly(H, 2)1   3.4741   0.5764     6.0273   0.0000
poly(H, 2)2  -1.3818   0.2832    -4.8790   0.0000
          S   0.0034   0.0008     4.1684   0.0001

Residual standard error: 0.2337 on 78 degrees of freedom
Multiple R-Squared: 0.8875
F-statistic: 205.2 on 3 and 78 degrees of freedom, the p-value is 0
```

Diagnostic plots for this model are given in Figure ?. In these plots observations 48 and 8 stand out as having large residuals and observation 1 has a high leverage. The output from influence.measures for these point is:

```
   .Intercept. poly.H..2.1 poly.H..2.2      S dffits cov.ratio cooks.d  hats
 1       0.015       0.023       0.049 -0.013  0.060     1.498   0.001 0.297
 8      -0.011       0.301       0.237 -0.118 -0.732     0.428   0.108 0.026
48      -0.339       0.557      -2.136  0.173 -3.117     0.388   1.784 0.250
```

I explored the effect of deleting observation 48 and then observations 48 and 8. Observation 48 has a very large Cook's Distance indicating it is having a large overall effect on the fitted model. The output for summary for this model is:

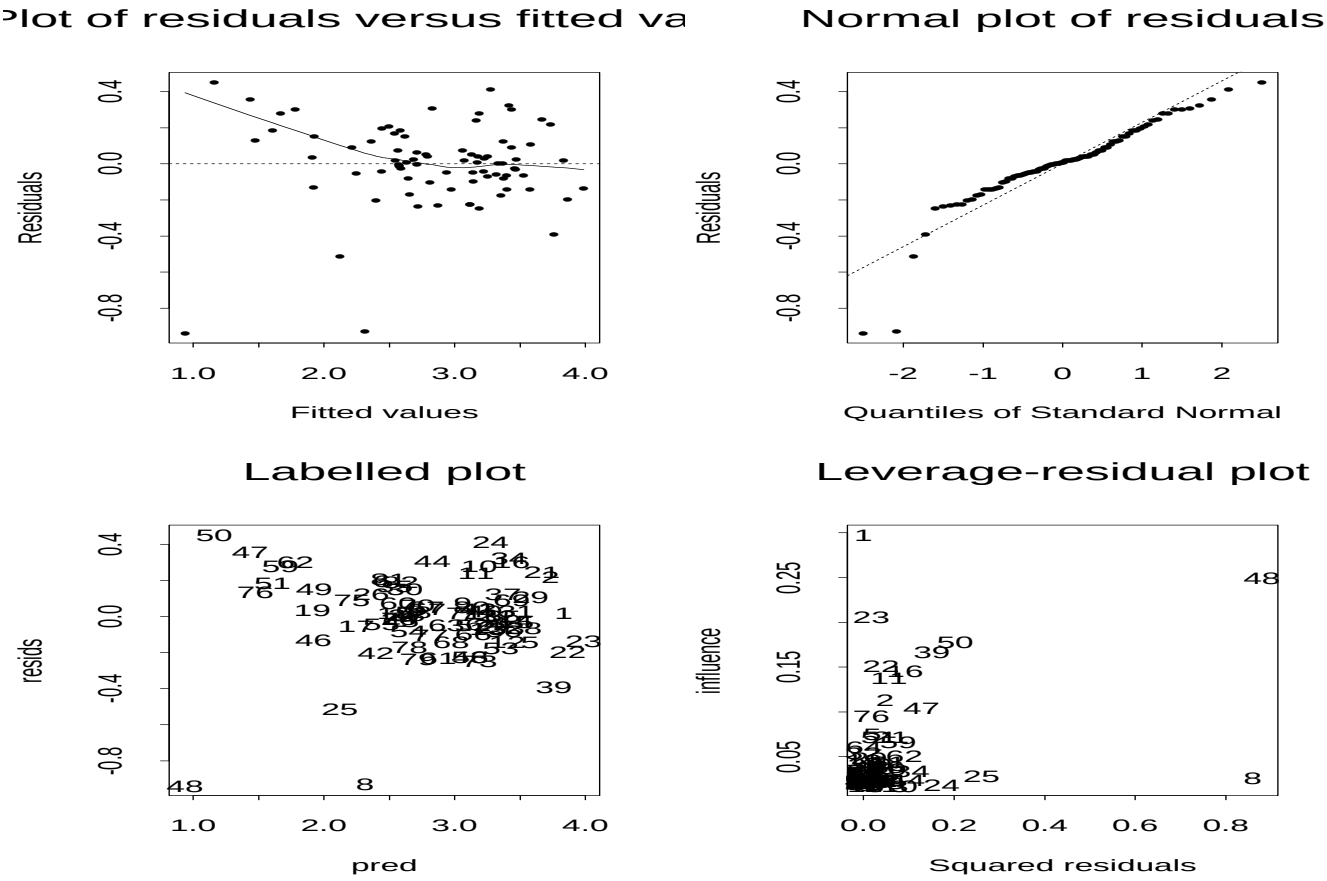

Figure 3: Diagnostic Plots

```
Coefficients:
                Value Std. Error  t value Pr(>|t|)
(Intercept)    2.4835    0.0758    32.7642   0.0000
poly(H, 2)1    2.7834    0.4098     6.7914   0.0000
poly(H, 2)2   -0.8300    0.1994    -4.1624   0.0001
          S    0.0033    0.0006     5.7282   0.0000

Residual standard error: 0.1682 on 76 degrees of freedom
Multiple R-Squared: 0.9215
F-statistic: 297.3 on 3 and 76 degrees of freedom, the p-value is 0
```

The main impact of dropping these two points is on the estimated coefficients for H and H^2 and there is a big decrease in the "Residual standard error". Dropping these points also improves the diagnostic plots obtained from diag.plots – there are no obvious outliers and the Normal plot of residuals is straighter.

It is interesting that observations 48 and 8 represent the two observations with the smallest values of M (1 and 4). This suggests that we could use the model fitted for the reduced data set as long as we are not concerned about really small mussels ($M \leq 4$). However, for my report I have used the model fitted for the full data set since we do not know that this would be the case.