Applying alternative assessments to approximate abalone age

[redacted]

1 Introduction and Hypotheses

Abalone (*Haliotis rubra*) is a highly sought after food resource which grows off the coast of Australia (Duncan 2003). Since these shellfish grow very slowly, many countries regulate harvesting them. To maintain the populations, it is important to know the age of abalone so that younger abalone are not harvested. In fact, several countries regulate the size of abalone harvested to varying degrees of success.

The current method used to approximate the age of abalone is counting the number of dark rings in their shells and adding 1.5. Previous work has found that abalones form three minor rings in the first 16 months of life, a major ring in their second year, and another major ring in each of the following years (Prince et al. 1988). In related species of bivalves, these rings form following an annual cycle. For example, Atlantic surf clams form a dark ring in late summer to early fall and form a light ring during the rest of the year (Jones 1980). To count the number of dark rings an abalone has, researchers have to cut open the shell, stain it, and then use a microscope to count the rings (Dua and Graff 2017). This process is tedious and researchers are interested in determining if there is an easier to obtain measurement that can also be used as a proxy for age. Since abalone follow an annual cycle to form rings, it is important to either measure the sampled abalone at the same time of year for each sampling or track the time of year that each abalone is measured.

Since abalones grow with age, we hypothesize that a measurement of size or weight will be able to predict the number of dark rings in the shell and, thus, predict age. Further, Prince et al. (1988) developed an age-length key to predict the number of dark rings an abalone has from its measured length which suggests length could be a good predictor of number of rings. It is not clear whether the length in that paper is the same measurement as the length in our data set since Prince et al. (1988) do not describe their measurement technique. We expect that a model describing size or weight of abalone will not explain all of the variation in the data since there are other contributing factors like resource availability which were not measured in this study.

2 Data Description

The data we are studying for this project was originally published in a Tasmanian Fisheries report (Nash et al. 1994). We accessed this data via the UCI Machine Learning Repository (Dua and Graff 2017). The data set includes measurements on 4177 samples of nine variables which are described in Table 1. There are no missing values in the data since the researchers removed those data points prior to publication. The researchers note that the majority of the missing data was missing response variable, rings. The continuous data is scaled by $\frac{1}{200}$. The researchers say this scaling was applied so that they can use it with an artificial neural network.

Table 1: Description of measurements collected on abalone.

Variable Name	Data Type	Units	Description
Sex	Nominal		female, male, or infant
Length	Continuous	mm	longest shell measurement
Diameter	Continuous	mm	perpendicular to length
Height	Continuous	mm	with meant in shell
Whole Weight	Continuous	g	weight of whole abalone
Shucked Weight	Continuous	g	weight of meat
Viscera Weight	Continuous	g	weight of gut after bleeding
Shell Weight	Continuous	g	weight of dried shell
Rings	Ordinal	count	number of rings in shell

3 Regression Analysis

3.1 Initial Analysis and Predictor Transformations

To begin our analysis, we scale each of the continuous predictors by 200 to undo the scaling applied by the previous researchers. This rescaling is not necessary and does not change the fit of the model, but it does allow for easier interpretation of the resulting model. From out initial analysis of the data, we find two points with zero height (1258 and 3997) and two points with heights at least two times larger than the heights of the other abalones (1418 and 2052). These points are shown in Figure 1. Additionally, we find that the data points with absurdly large heights are leverage points when we fit a Poisson distributed with all of the predictor variables. We decide that these four data points are outliers and remove them from the data set.

Next, we plot the marginal distributions of each predictor variable versus the response variable, rings, to determine if there is a non-linear relationship and if predictor transformations would be useful. These plots are shown in Figure 2. The box plots for female and male abalone versus rings overlap and show similar behaviour, while the box plot for infants shows that infant abalone tend to have fewer rings. Since infants show different behaviour from adults in this marginal distribution, we factor sex to create dummy variables for the predictor variable. Even though the female and male abalone tend to have similar numbers of rings, we leave

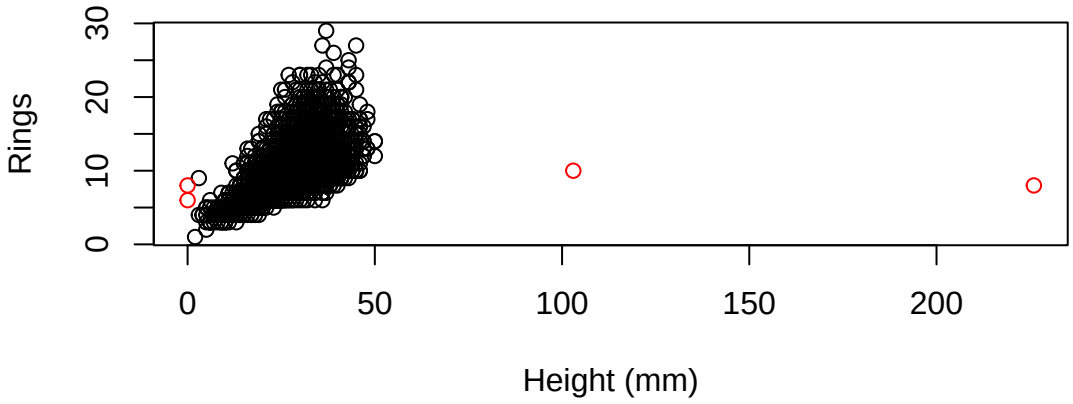

Figure 1: Plot of the data collected on heights of abalone. Outliers are shown in red.

these factors separate and do not make a combined adult group since there are sex differences in abalone. We note that each of the continuous predictors has increasing spread in the data. This heteroskedastic behaviour is expected from Poisson data since $\mathbb{E}[y] = \mathbb{V}ar[y] = \lambda$. By observation, we find that the size measurements (length, diameter, and height) follow a linear trend and the weight measurements (whole, shucked, viscera, and shell) show concave behaviour.

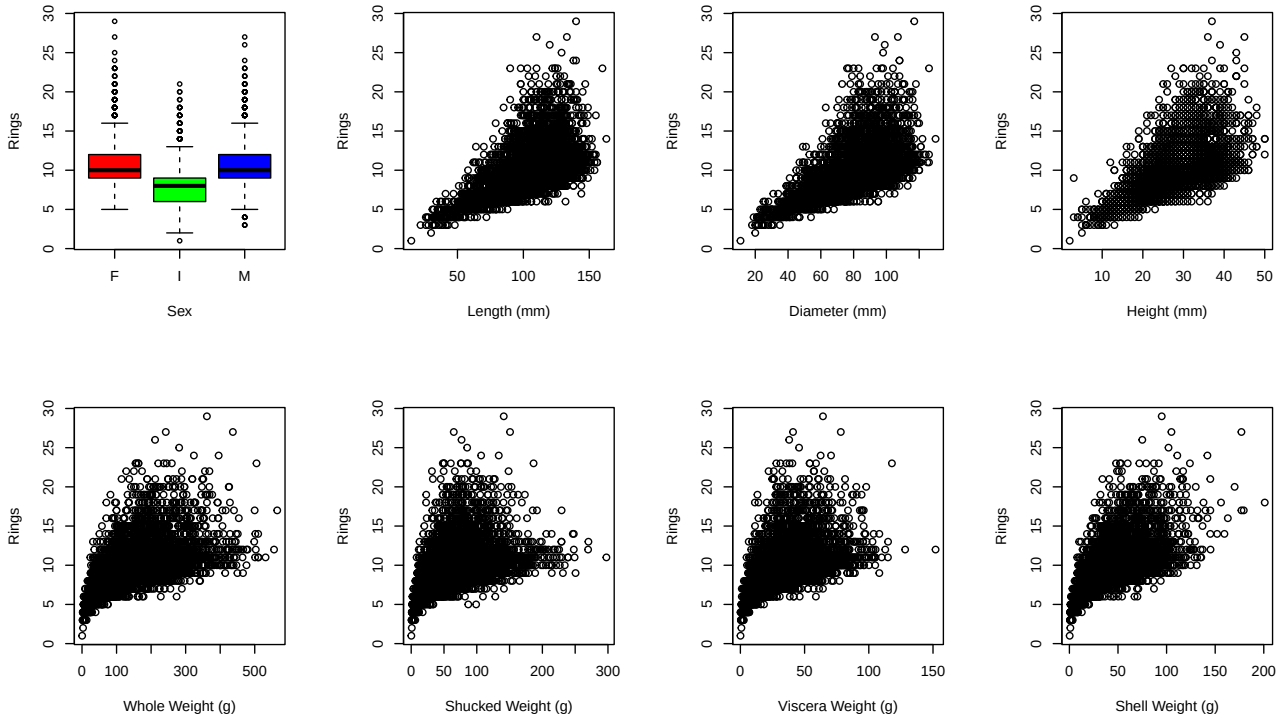

Figure 2: Marginal distributions of each predictor versus the response. In the top left plot, F stands for female, I stands for infant, and M stands for male.

It is not surprising that rings have a concave relationship to weight since weight is related to the volume of the abalone. Abalones are ellipsoidal in shape, but we assume they are perfect spheres for interpretation of this concave relationship. Under this assumption, the volume

of abalone is $\frac{4}{3}\pi r^3$. Since we observe rings are linearly related to the size measurements and weight is proportional to volume, we get rings$^3 \propto$ weight. This relationship suggests cube rooting the weight predictors will result in data that can be fit by a linear model. Performing a log transformation can also account for non-linear relationships. We perform these predictor transformations, fit a linear regression to the resulting data, and compare the resulting fits to determine which transformation is best.

The original model with no transformations applied has AIC = 18938.527. The first transformation we apply is cube rooting the weight predictors which corresponds with our assumed biological relationship between weight and rings. This model has AIC = 18826.977. The second transformation we apply is log-transforming the weight predictors and the corresponding model has AIC = 18750.022. The last transformation we try is log-transforming all of the predictors in case there is a non-linear relationship between size and rings that we did not observe in the marginal distributions. We find AIC = 18758.837 for this model. The most simple and best transformation is the log-transformed weight predictors. We apply a log transformation to the weight predictors and proceed with this transformed data.

3.2 Initial Variable Selection

Since the collected measurements mostly measure abalone size and weight, we expect our biggest obstacle to be substantial multicollinearity in the predictor variables. Hence, we will investigate several different variable selection methods to find a model that has limited multicollinearity and gives a reasonable fit to the data. We choose rings to be the response variable which means we want a distribution that accounts for a discrete response such as Poisson or Negative Binomial. We start by fitting a Poisson model with all of the predictors and check for overdispersion.

From our initial model fit, all of the predictors except diameter and the dummy variable for male abalone are significantly different from zero. However, there are some issues with the model fit. The variance inflation factors (VIFs) show strong multicollinearity in the predictor variables. In fact, max(VIF)=186.810. The data is also underdispersed since the square root of the dispersion parameter is $\sqrt{\phi} = 0.640$. We find that setting the dispersion parameter in our model does not change the parameter values.

We fit a quasi-poisson model to investigate the underdispersion further. The quasi-poisson model calculates a dispersion parameter of 0.409 which is the same as the dispersion parameter for the original poisson model since $\sqrt{0.409} = 0.640$. The parameters of the poisson and quasi-poisson models are very similar. Underdispersion is common in multicollinear models, but there is not a lot of research in the area of dealing with underdispersed data. Since multicollinearity is a potential reason for underdispersed data, we expect underdispersion to become less of an issue as we deal with the multicollinearity in the data.

Forward selection and backward elimination have the potential to bias the resulting model. Additionally, we have nine parameters (including dummy variables) which means we can apply best subset regression without too much concern for overfitting to the data. We compare the resulting models from all three variable selection methods and find that forward selection, backward elimination, and best subset regression all choose the same model. This model

includes all of the predictors except diameter. The dispersion parameter for this model is $\sqrt{\phi} = 0.640$ and the model is still underdispersed. Additionally, while this is the best model so far, we have not accounted for multicollinearity and max(VIF)=186.547.

Since the typical variable selection methods have not helped to reduce the multicollinearity in the model, we try a different method. Here, we remove the predictor with the largest VIF until the model does not show signs of significant multicollinearity. The values of the VIFs for each model throughout the reduction process are given in Table 2. There is at least one predictor with VIF > 10 in every model except the final model. The multicolinearity is at a reasonable level in the final model since max(VIF) < 5. The dispersion parameter has also improved to $\sqrt{0.409} = 0.790$ and underdispersion is less of an issue.

However, the resulting model has a worse fit to the data. In particular, the predictor shucked weight is not significantly different from zero and the model has AIC=19535.222 which is higher than the original multicollinear models with more parameters. Since the model is underdispersed, it is possible that the model is incorrectly rejecting a significant predictor. However, the parameter for shucked weight is -0.016 and it does not make sense biologically that rings decrease when weight increases. We investigate the model fit after removing shucked weight from the model and find a similar AIC value but worse residual deviance. We call this model with predictors sex and height the VIF reduced model. We provide some summary statistics for the VIF reduced model in Table 3 and plots of the fit of this model to data in Figure 3. The model gives a good fit to the data for intermediate heights and a slightly worse fit for abalone with heights further away from the mean.

Table 2: VIFs throughout model selection of removing variable with the largest VIF.

Variable	Full Model	Reduced 1	Reduced 2	Reduced 3	Reduced 4	Final Model
Sex	1.519	1.501	1.491	1.457	1.451	1.446
Length	35.664	35.778				
Diameter	34.426	34.571	14.295	14.103		
Height	5.479	5.468	5.473	5.420	5.066	3.980
Whole Weight	186.810					
Shucked Weight	49.229	17.024	16.283	11.749	8.834	3.901
Viscera Weight	24.783	18.168	17.855			
Shell Weight	38.727	15.492	15.529	13.904	12.168	

Table 3: Summary statistics for our VIF reduced model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.683	0.024	71.108	0.000
SexI	-0.121	0.013	-9.180	0.000
SexM	-0.014	0.010	-1.331	0.183
Height	0.023	0.001	32.687	0.000

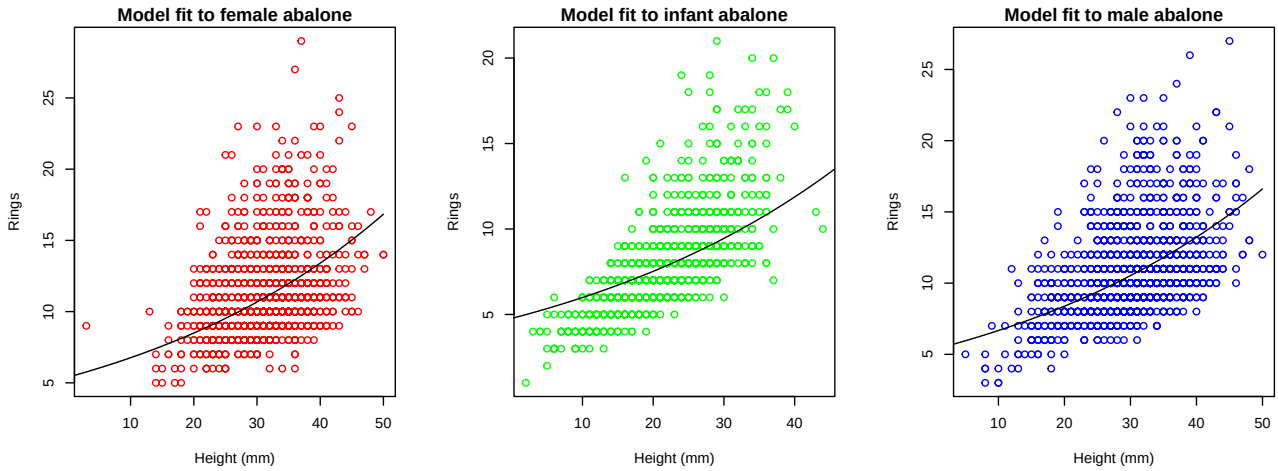

Figure 3: Plots of the fit of the VIF reduced model to data for each abalone sex.

We investigate the rest of the models with sex and one other predictor. We find the model with shucked weight produces a model with the best dispersion ($\sqrt{\phi} = 0.850$) and shell weight produces a model with the best AIC (19282.897) for this subset of models. There appears to be a negative relationship between AIC and dispersion for these models.

The forward selection, backward elimination, and best subset selection methods left us with models that are highly multicollinear which results in poor parameter estimation. The VIF reduced model provides a reasonable fit to the data, but is not backed up by statistical theory. Now, we turn to variable selection using shrinkage estimators to analyze the data further and use proven methods to obtain models with less multicollinearity.

3.3 Variable Selection using Shrinkage Estimators

The two shrinkage estimators we consider here are ridge and Least Absolute Shrinkage and Selection Operator (LASSO). These estimators are biased to decrease the variance of the parameters and attempt to decrease the mean squared error (MSE). This bias amounts to solving a constrained minimization problem where the constraint is $\sum_{i=0}^{k} \beta_i^2 \leq t$ for ridge regression and $\sum_{i=0}^{k} |\beta_i| \leq t$ for LASSO regression. We first apply cross validation to find the optimal value of the penalizing parameter λ. For both ridge and LASSO regression, we find that the MSE is minimized when $\lambda = 0$. The one standard deviation LASSO model is the only one that does not include all of the predictors. This model removes sex, diameter, and viscera weight. The rest of the models (minimized LASSO, and both minimized and one standard deviation ridge) include all of the predictor variables. The minimized ridge and LASSO models both have very similar parameters to the original model we fit without accounting for multicollinearity. This result means using ridge or LASSO regression with non-zero λ increases the MSE and provides a worse fit to the data. It is especially concerning that $\lambda = 0$ minimizes the MSE since it suggests ridge and LASSO regression should not be used and we are still stuck with a highly multicollinear model.

3.4 Principle Component Analysis

All of the models so far are either highly multicollinear or are not backed up by statistical theory. We investigate the data using another biased estimator, principal component analysis (PCA), to combine the collinear variables into new orthogonal predictors. The goal of PCA is to reduce the number of variables in a data set while maintaining most of the information from the original data. PCA is particularly useful for ill-conditioned data, including multicollinear variables (Montgomery et al. 2012).

We remove rings and sex from the data set for now since we only want to create principal components for the predictor variables and PCA does not apply to categorical variables in general. There exist other dimension reducing methods similar to PCA for categorical variables but we will not consider them here because we only have one categorical variable and it is not highyl correlated with the other predictors. The first step of PCA is standardizing the continuous predictor variables so that they are of the same magnitude and one variable does not bias the results. Then we compute the covariance matrix to determine how correlated the predictors are with each other. We find that all of the continuous predictors are highly positively correlated with each other. In fact, the smallest correlation is between height and shucked weight with r = 0.874. An additional test to confirm the multicollinearity in the data is finding the eigenvalues of this covariance matrix. Eigenvalues that are exactly $\lambda = 1$ mean the predictors are orthogonal and eigenvalues close to zero mean the data is highly multicollinear. Not surprisingly, most of the eigenvalues for our data set are $\lambda_i \approx 0$ for $i \in \{1, \ldots, 7\}$.

Next, we compute the principal components by taking linear combinations of the original predictors to form new independent predictors. The algorithm maximizes the amount of information in each vector in descending order, i.e., the first vector contains the most information from the entire data set, the second vector contains the most possible information left over after the first vector is computed, and so on. Some summary statistics for this new data set are given in Table 4. We find that the first principal component accounts for 0.953 of the variation in the data. This results suggests that the rest of the principal components do not add much to the new description of the data and are not necessary for model fitting.

Table 4: Summary statistics for PCA data set.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.583	0.396	0.284	0.201	0.183	0.112	0.062
Proportion of Variance	0.953	0.022	0.012	0.006	0.005	0.002	0.001
Cumulative Proportion	0.953	0.976	0.987	0.993	0.998	0.999	1.000

We plot the predictors in terms of the first and second principal component to determine which components best represent the predictors in the top left of Figure 4. We observe that all of the arrows representing the predictors are close together and are positively correlated with each other. The arrows are also all roughly the same length which means the variables are represented equally well by the first two principal components. In particular, all of the continuous predictors except for height are almost entirely described by the first principal component. The plot on the top right of Figure 4 shows the quality of representation of each

predictor in the first principal component using a squared cosine function. We find the first principal component provides a good representation of all of the predictors. The predictor that has the worst representation in the first principal component is height with about 87% representation. A plot of the data against the first two principal components is given in the bottom left of Figure 4. We note that there appears to be clusters in the data based on the age group of the abalone. In particular, abalone infants tend to have a smaller first principal component value.

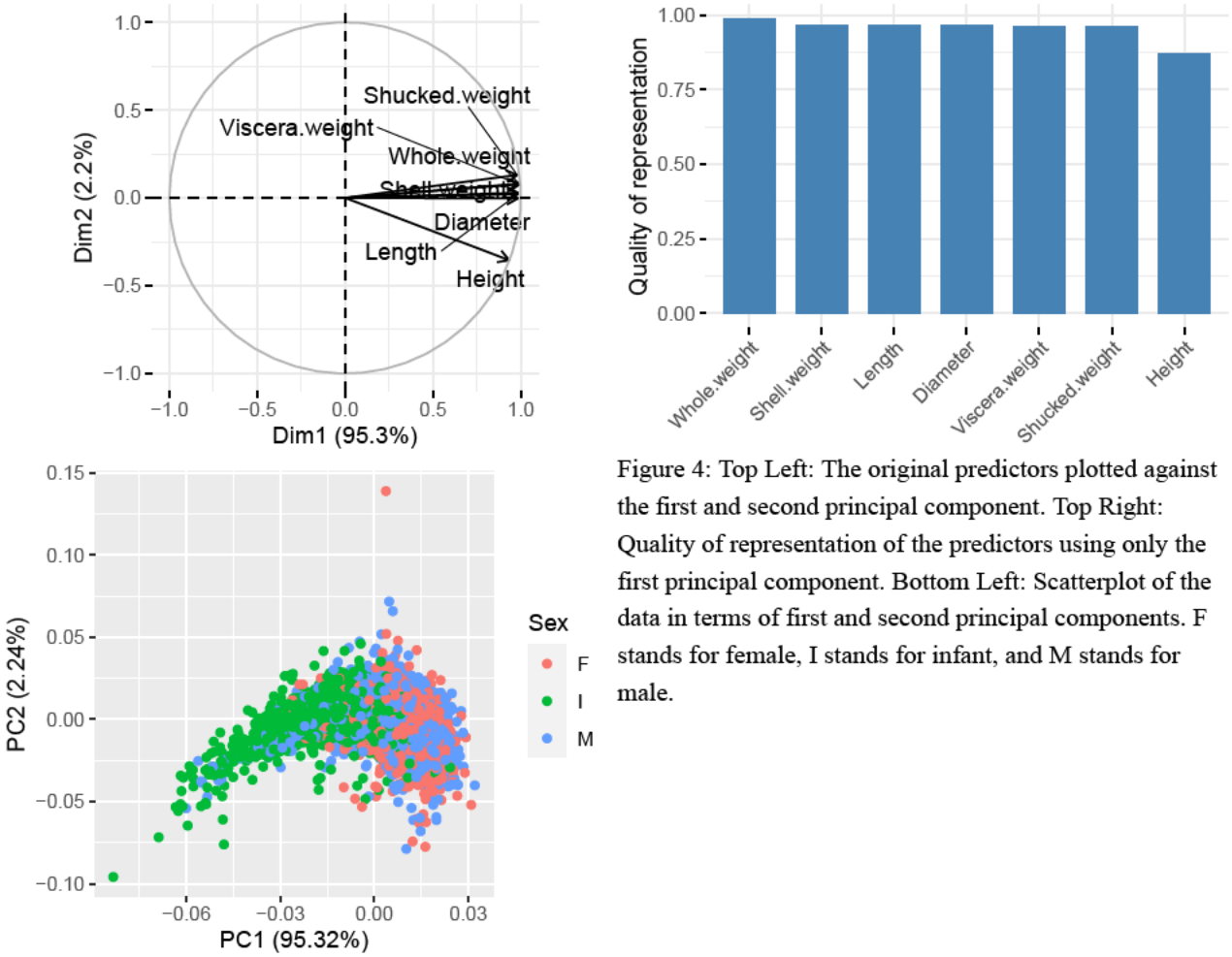

Figure 4: Top Left: The original predictors plotted against the first and second principal component. Top Right: Quality of representation of the predictors using only the first principal component. Bottom Left: Scatterplot of the data in terms of first and second principal components. F stands for female, I stands for infant, and M stands for male.

Before fitting a model to our principal components, we check the eigenvalues of the covariance matrix to ensure the new predictors are orthogonal. We find that the eigenvalues are all $\lambda_i = 1$ for $i \in \{1, \ldots, 7\}$ which means we now have orthogonal predictors. We concatenate sex and rings back onto the principal component data set and fit a Poisson model. We only include sex and the first principal component since the rest of the principal components do not add much more information to the model. Some summary statistics from this model are given in Table 5. We compute the dispersion parameter and find $\sqrt{\phi} = 0.801$ which is an improvement on the underdispersion found in our previous models. The null deviance is 4136.674 and the residual deviance is 2442.189. The residual deviance is much smaller than the null deviance and the current model is giving a much better fit than an intercept only model. This model has AIC =19569.170, which is worse than the original models but this is an acceptable trade-off for the

reduction of multicollinearity in the model. The VIFs given in Table 6 are significantly better than the original model and show that our predictors are no longer multicollinear. Plots of the PCA model fit to the data for each sex are provided in Figure 5. By observation, the PCA model fits the data reasonably well for the female and male abalone. It looks like the infant model would benefit from a larger slope parameter.

Table 5: Summary statistics for our model fit with predictors sex and the first principal component.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.316	0.008	286.423	0.000
data.SexI	-0.113	0.013	-8.355	0.000
data.SexM	-0.016	0.010	-1.526	0.127
PC1	0.071	0.002	31.168	0.000

Table 6: VIFs for our final model.

	GVIF	Df	GVIF^(1/(2*Df))
data.Sex	1.461	2	1.099
PC1	1.461	1	1.209

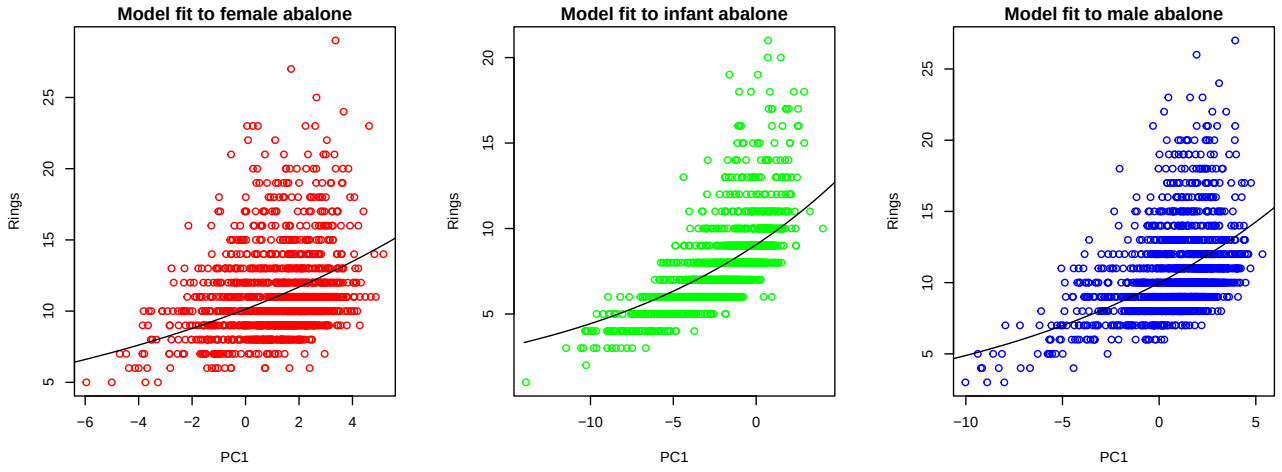

Figure 5: Plots of the fit of the PCA model to data for each abalone sex.

We consider adding more principal components to the model. We find the AIC is minimized when the first five principal components are included in the model. In this case, AIC = 18823.050. However, the dispersion of this model has decreased and $\sqrt{\phi} = 0.654$. Possible reasons for underdispersion include multicollinearity, clusters in the data that are not accounted for in the model, or overfitting to the data. We choose the model that minimizes the violation of the Poisson assumption $\mathbb{E}[y] = \mathbb{V}ar[y] = \lambda$. That is, we choose the model with only sex and the first principal component as predictors. Even though the fit appears to be worse, this

model is at less risk of overfitting to the data. We note that this PCA model is very similar to the VIF reduced model.

4 Discussion and Conclusions

We have several concerns about the data collection methods. In particular, it is concerning that the researchers removed the missing data prior to publication and did not publish the raw data. The authors also note that weather patterns and collection location of the abalone could be necessary to fully model the system, but these measurements were not collected. Location and weather patterns are particularly important variables to include since they relate to resource availability and, therefore, abalone growth. The date of sample collection was also not recorded which presents issues since abalones form rings on an annual basis (Prince et al. 1988). Additionally, based on the descriptions of the weight data, we expect whole weight to be greater than or equal to the sum of shucked weight and shell weight. The whole weight could be larger than the sum of its parts since the shells are dried before being weighed and could lose mass to water evaporation. However, this inequality does not hold in general and the whole weight is less than the sum of the weights of the meat and shell for all but 16 data points. The equality does not hold for any data points. The researchers provide no explanation of this discrepancy and it is not clear why these measurements do not line up.

The lack of variety in the variables also presents major issues. Of the eight predictor variables, four predictors measure the weight of the different parts of the abalone and three predictors measure the size of the abalone. Unsurprisingly, these predictors are highly correlated and cause multicollinearity issues in all of our models except the PCA and the VIF reduced models (i.e., the models where we accounted for the multicollinearity). Another concern is the authors wanted to find a model to estimate number of rings to then estimate age of abalone, but they do not collect any data on the age of the abalone measured in their study. If the end goal of a study is to predict the age of abalone, it only makes sense that the age of the sample abalone should be recorded. Additionally, we found a resource that uses the length of abalone to predict rings as a proxy for age and these authors also formulated an age-length key (Prince et al. 1988). In other words, there already exists a method to predict abalone age from their length.

The best models we found through our regression analysis are the PCA model that only includes the first principal component and the reduced VIF model. The rest of the models we studied have high multicollinearity and are underdispersed. We thought the approach of removing one predictor at a time until the VIFs improve was naive and would not result in a reasonable model. To our surprise, these models are very similar and, since the PCA model includes about 95% of the information from the original predictors, the VIF reduced model also describes a similar proportion of the variation in the original predictors. The high positive correlation in the size and weight predictors suggests that all of the size and weight measurements increase simultaneously. Further, this correlation in the predictors and the similarity between our two best models allows us to more easily interpret the PCA model. The first principal component represents all of the continuous predictors and as these predictors increase together, the number of rings increases. Infants tend to have less rings than female abalone, and male and female abalone do not show a significant difference in the number of rings.

There is not one most important predictor variable. All of the size and weight predictors provide roughly the same amount of explanation for the variation in the number of rings. In particular, we note that the models with sex and one of the size or weight predictors are all fairly similar. The models also all showed the same general behaviour of increases in the size or weight predictor leads to increases in rings. We found that, of this subset of models, the model with the best AIC is the most underdispersed (shell weight) and the least underdispersed model has the worst AIC (Shucked weight). There appears to be an inverse relationship between AIC and dispersion for this data set. Underdispersion causes our model predictions to be more conservative and can lead to us rejecting significant predictors. In the case of our best models, all of the variables are significantly different from zero except for the dummy variable for male abalone. We are not concerned by this variable not showing a significant difference since our initial box plots in Figure 2 do not show a difference in rings between female and male abalone. Hence, we are not concerned that the model is too conservative with its estimates since all the parameters we expect to be significant are, indeed, significant.

Altogether, the data collection methods performed by the previous researchers were poor and do not include enough variety in types of measurements. Most of the measured predictors are highly positively correlated and increase together over the lifespan of an abalone. The authors did not publish their raw data and the measured weights are not consistent. Our best models included sex and one other predictor (either height or the first principal component). Lastly, there appears to be at least one missing variable and the lack of such a variable decreases the fit of our models to data. In future work, we recommend more clarity and consistency in the data collection methods and a larger variety of variables measured such as time of collection, age of the abalone, or location and resource availability.

Sources

Dua D, Graff C (2017) UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences

Duncan PF (2003) SHELLFISH | commercially important molluscs. In: Caballero B (ed) Encyclopedia of food sciences and nutrition (second edition), Second Edition. Academic Press, Oxford, pp 5222–5228

Jones DS (1980) Annual cycle of shell growth increment formation in two continental shelf bivalves and its paleoecologic significance. Paleobiology 6:331–340

Montgomery DC, Peck EA, Vining GG (2012) Introduction to linear regression analysis, 5th edn. Wiley-Blackwell, Hoboken, NJ

Nash WJ, Sellers TL, Talbot SR, et al (1994) The population biology of abalone (Haliotis species) in Tasmania. I. Blacklip abalone (H. rubra) from the North Coast and Islands of Bass Strait. Sea Fisheries Division, Dept. of Primary Industry; Fisheries, Tasmania

Prince J, Sellers T, Ford W, Talbot S (1988) A method for ageing the abalone haliotis rubra (mollusca : gastropoda). Marine and Freshwater Research 39:167. https://doi.org/10.1071/mf9880167