ESCI 502 Final Exam

Luke Ghallahorne Fall 2023

This exam is worth 52 points and is worth 25% of the grade for the class. You will turn in a Word/PDF file and your R code. Not all questions will require R work. To complete your exam, you will upload your files to Canvas by the deadline.

This exam is open book. As such, I expect precise language in your answers, which should be concise and to the point.

Do your own work
This exam is to be completed by you and you alone. You may not work with your classmates or with your data analyst partner. You may consult your own notes and labs, books, Canvas, R documentation, and internet help. You may contact me with questions. If you are uncertain about what it means to turn in your own work, please see the WWU Academic Honesty policy.

Tips for full credit
Use clear, concise, and precise language.
Be specific about what you find. Think about what you are testing and what exactly your findings refer to; provide statistical evidence.
Copy plots to your Word document using the *export* function on the plots pane of R Studio so all elements are legible (and make them big enough to read on a laptop). Label plot axes for all final plots (not needed for exploratory plots).
Read the full problem before diving in so you know where you are headed and can get there efficiently, with a logical workflow.

Code
Provide clear code. Use comments and line breaks. Use logical object names that are not too long. Make sure that your code will run from top to bottom without error. To check this, restart R once you are done and run your entire script, checking the console for output errors. *In addition to turning in your R file, please copy/paste code at the end of your document (this helps with a quick scan while grading).* If your code does not run, you will lose points.

Tips if you get stuck
Check your object labels.
Check capitalization and punctuation.
Run ?*function* (e.g. ?lm) to make sure you have specified the arguments correctly.
Save your code, restart R, and re-run your code again. Sometimes an object will be saved in the environment that was a relict of old code.
Check any subsets and intermediate analysis as you go to make sure they are what you expect.

Grading
Point values for each question are listed below next to the questions. Partial credit will be given. However, answers without associated code (where needed to produce an answer) will not be given credit. Your code counts for 10 points on this exam—make sure it runs and is legible.

You've got this!

1. Describe (5 pts):

Define the three components of a generalized linear model and describe how you would choose a proper distributional family.

The three components of a generalized linear model are a random component, a systematic component, and a link component. The random component is the response variable (i.e., Y) and its probability distribution, usually a member of the exponential family (normal, binomial, Poisson, gamma, negative binomial, etc.), described by the mean and the spread. The systematic component is the predictor variables (i.e., X_i), which can be categorical or continuous and may interact. The link component is a function that relates the predictors to the response, determined by the distribution of the response variable.

To choose a proper distribution family, first examine the response variable type. If the response is a continuous variable, a normal, gamma, or other continuous distribution should be considered. If it is discrete, a binomial, negative-binomial (presence-absence), or Poisson (count) is more appropriate. Then, visualize the variable with a histogram or similar tool, and compare the distribution to candidate distribution families.

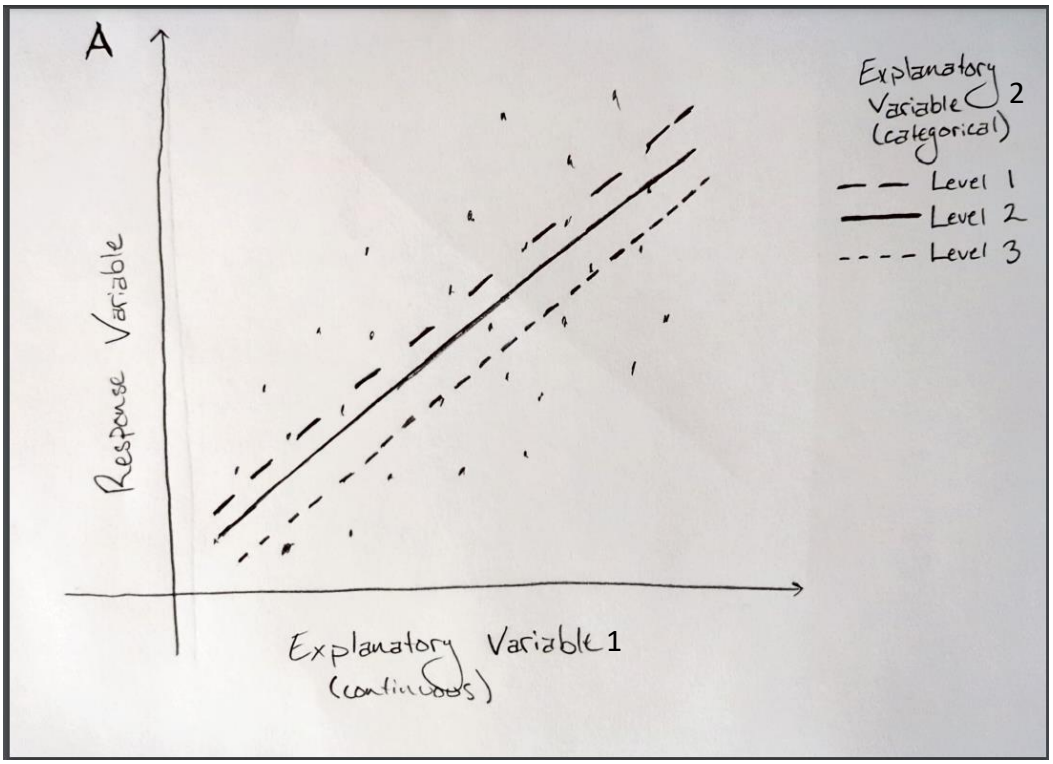
2. Draw (5 pts):

Draw (roughly, by hand is fine) two plots. Each plot will have a continuous response variable and a continuous explanatory variable as well as a categorical explanatory variable with 3 levels. Draw the predicted line for 3 separate groups (levels) where:

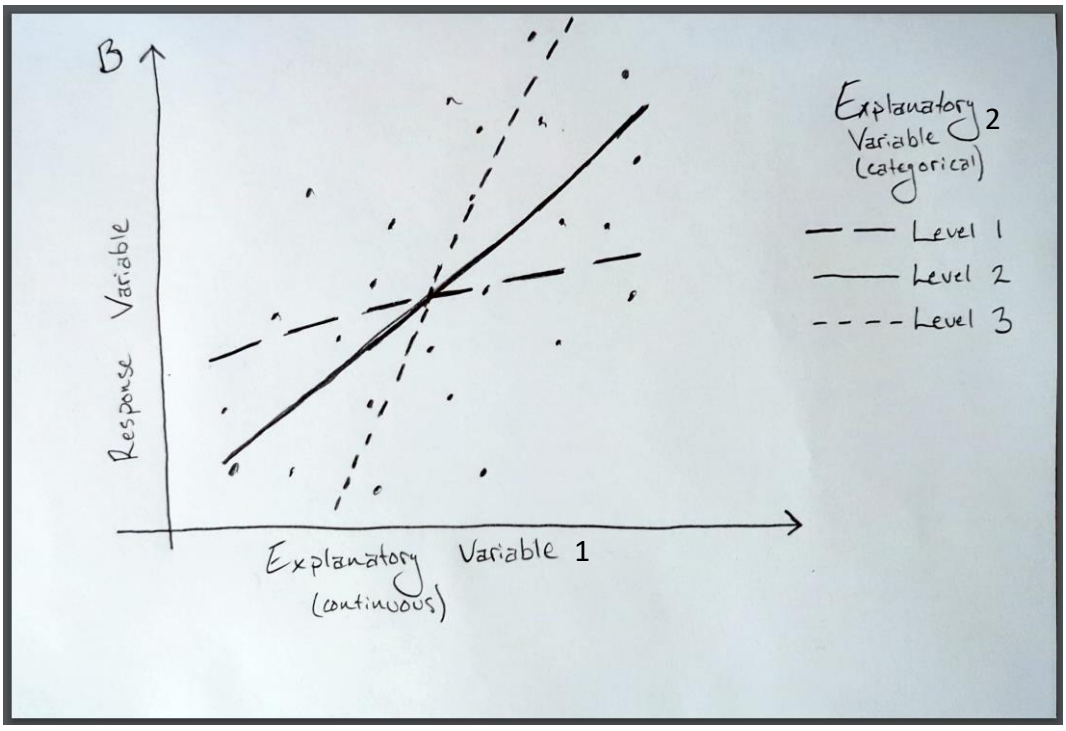
Plot (A) will show the 3 groups with no interactions
Plot (B) will show an interaction effect between the categorical and continuous predictors, with respect to the response variable.

In short, demonstrate what an interaction effect would look like and what a plot without an interaction effect would look like, and describe an interaction effect in words.

An interaction effect occurs when the predicted response of one explanatory variable changes as another explanatory variable changes. As seen in plot B, the level of the categorical variable alters the predicted response for a given value of the continuous explanatory variable. Interaction plots visualize this through the different slopes and intersection of regression lines as the levels of the categorical variable change.



Plot A: Multiple regression of a continuous response variable on one continuous and one categorical explanatory variable. Prediction lines are parallel, indicating no interaction between the explanatory variables.



Plot b: Multiple regression of a continuous response variable on one continuous and one categorical explanatory variable. Prediction lines intersect at some value of EV1, indicating that the variables are interacting with each other as well as the response.

3. Define (3 pts):

Random Effects

Random effects are blocking units that are randomly selected from a whole, representing realized possibilities from an overall set of means and not purely independent samples.

Collinearity

Collinearity occurs when explanatory variables are strongly correlated with one another, allowing one explanatory variable to be predicted from the others more accurately than the response.

Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a method of parameter estimation used in generalized linear models that calculates likelihood of a response vs a prediction with respect to the link function, instead of ordinary least squares. This is done because a GLM does not output predictions in the same space as the response variable and needs the link to translate.

4. Discuss (9 pts)

a. Why using a generalized linear model is preferable to transforming your lognormal data and using a linear model.

While transformations of data to create a linear model is possible and sometimes useful, it fundamentally changes the data and how the variables are related to each other. Log transformed values can be difficult to interpret compared to raw measurement units. If the data include a lot of zeroes, a fudge value must also be added, further altering the data. Moreover, not all data can be normalized with transformations. Generalized linear models keep the input data in their original form and distributions, while using a link function to translate output into the scale of the response variable, allowing for easier interpretation of relationships between variables.

b. Model goodness of fit versus model selection.

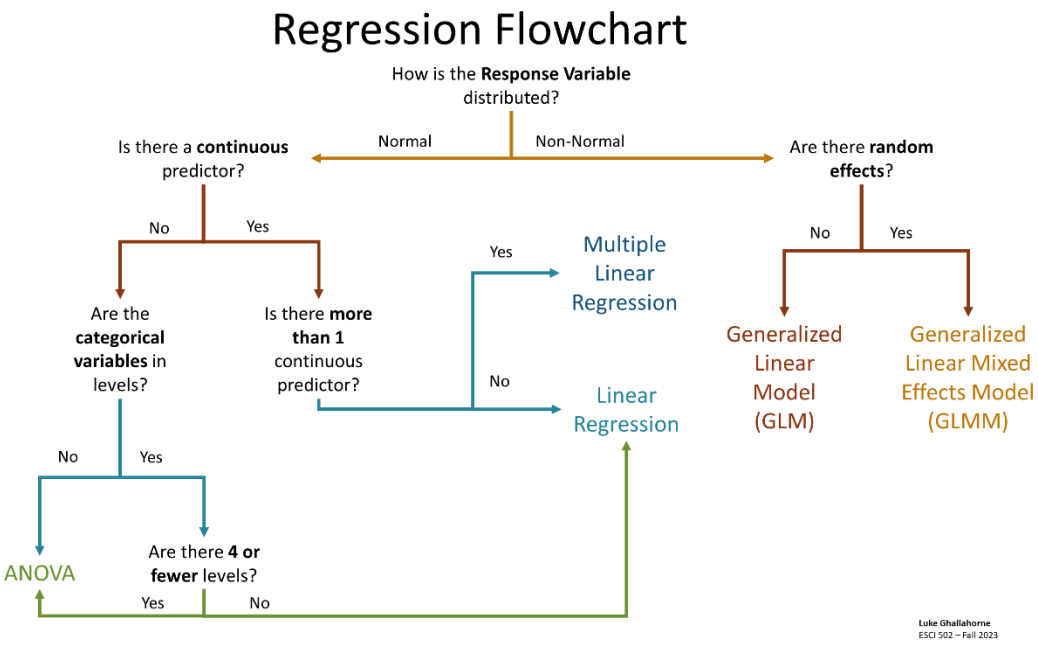
The goodness of fit of a model is a test verifying the accuracy of a model compared to the original data. It typically (R^2, pseudo R^2, etc.) provides the percentage of variance in the data that is explained by the modeled predictors. Model selection uses goodness of fit tests as a tool to determine the best fit model from a suite of candidates. We can compare goodness of fit or likelihood tests to compare models with added or dropped explanatory variables.

c. Why partial regression plots are important to interpreting a multiple regression problem.

With multiple explanatory variables, it is very difficult to visualize the impact of each variable. Partial regression plots depict the relationship between the response variable and one explanatory variable in a multiple linear regression, while holding all other variables constant at a single value (typically the mean). This allows us to see the impact each predictor is having on the response, and the magnitude of that relationship compared to other predictors.

5. Design (10 pts)

This quarter you have learned a number of statistical techniques centered around the general regression framework. Build a flowchart with the techniques we covered. Explain when/why you would use each technique. There are many examples out there, but your task is to design a decision framework with the tools you have learned this quarter (not all of the possible statistical techniques available on the internet). You can do this by hand and upload an image (make sure it is clear) or use a graphical software like PowerPoint or other



I included my flowchart as a separate file on Canvas as well so it will be easier to see.

ANOVA and linear regressions assume the response variable (and explanatory variables) are normally distributed. They can be transformed to be normal if needed, but there are better practices if the response variable is non-normal – generalized linear models (GLMs).

ANOVA compares the means between groups and is applicable to a categorical explanatory variable. If the categories cannot be understood as levels (e.g., low-medium-high versus male-female), ANOVA is a good choice. If the categories can be made into levels, then a linear regression is more powerful than ANOVA when there are more than 4 levels.

If more than one continuous predictor is being considered, multiple linear regression can account for each, including interactions (if justifiable).

Generalized Linear Models can model response variables with many probability distributions by using a link function. GLMs can be expanded to include random effects as Generalized Linear Mixed Effect Models (GLMMs), though they are difficult to fit accurately.

6. Do (20 pts)

Suppose you've been busily working in the laboratory sectioning abalone shells for the purpose of age determination. In addition to age information, you've also noted the sex of each specimen and you have taken several shell measurements (length, diameter, and height) and weight measurements (whole, shucked, viscera, and shell). Ultimately, you would like to develop the best predictive relationship of age as a function of the variables you've measured. Because sectioning abalone is tedious and you'd rather be skiing.

Preliminary analysis has shown that **shell diameter**, **shell weight,** and **shucked weight** are promising indicators. Formulate a candidate set of linear model parameterizations (designed to predict abalone age) involving combinations of the three variables; also include the **categorical sex variable**, as gonad may make a difference in the relationship between weight and age. Do not fit any interactions, even though you might suspect sex and shucked weight have a potential interaction based on the previous info (here, main effects only).

Assess the input data and assumptions of your modeling framework with standard techniques.

Write out the candidate models, fit them to the data, and use model selection to discriminate among the parameterizations.

Report your AICs and associated results in table form to demonstrate how you selected a best fitting model.

Assess your best fitting model for violation of assumptions and report the parameter estimates, standard errors, and relevant statistics associated with the best fitting model(s).

Make a concluding statement about your results.

Data file: *abalone2.csv*

1. Assessing data and assumptions.
   a. Visualizing data:

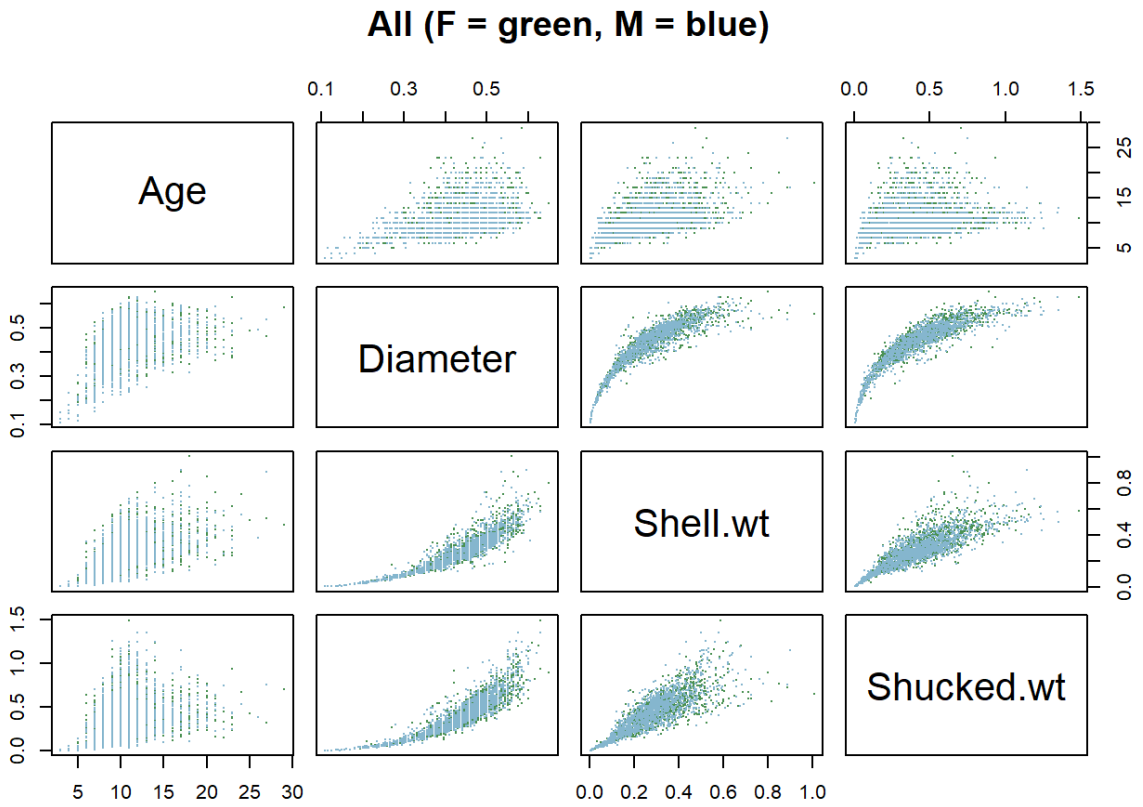


Fig 1: Scatterplots of response and continuous predictor variables. Colors represent sex.

All three predictors have some relationship to age (as expected, given the introductory information); it appears funnel shaped, with tighter data at smaller values and greater values as the Xi's increase. Diameter, shell weight, and shucked weight are also interrelated, with exponential/logarithmic funneling patterns.

These patterns do not change when comparing male and female abalone.

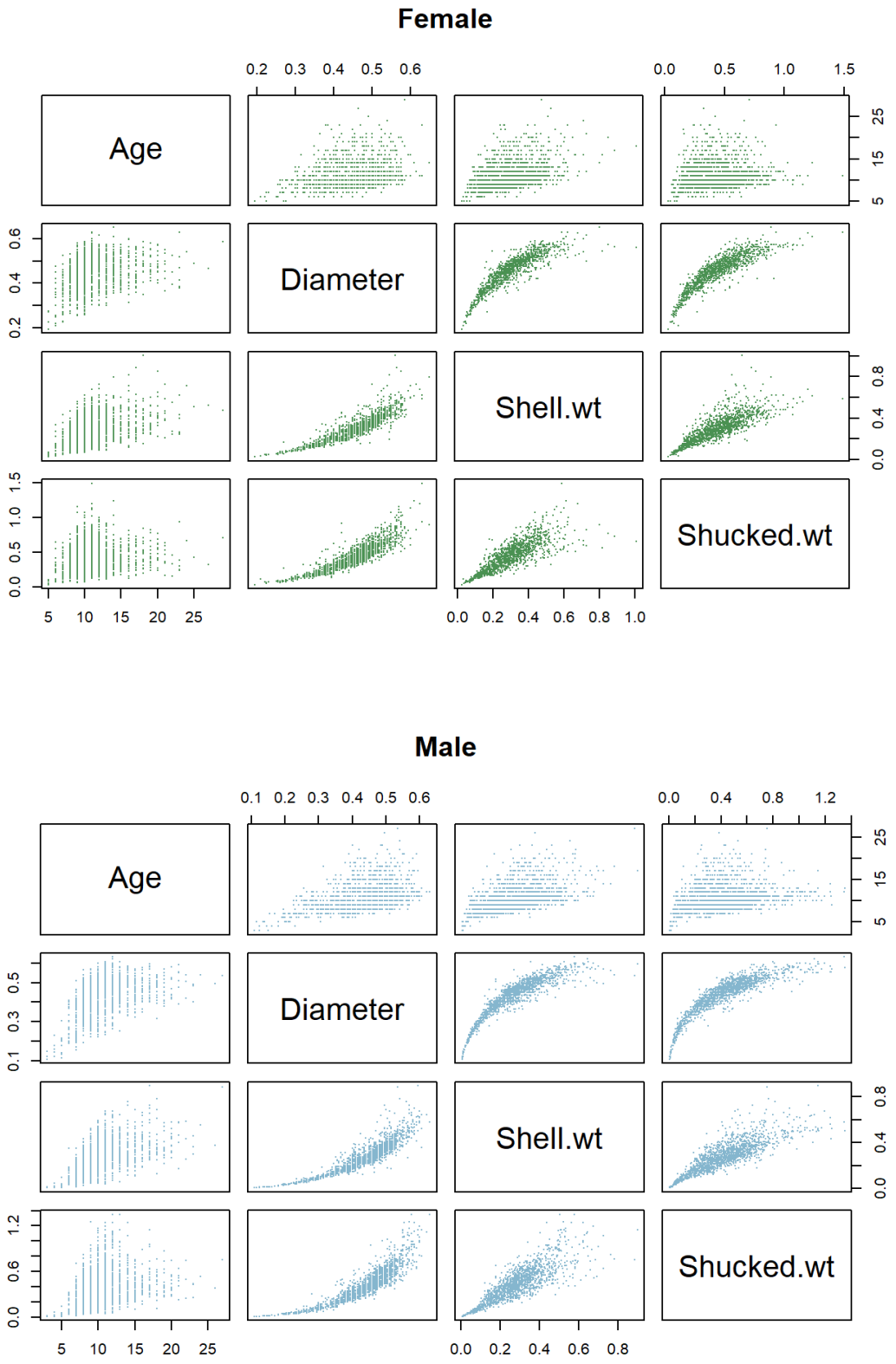


Fig 2: Scatterplots of variables separated by male and female abalone.

b. Assessing distributions:

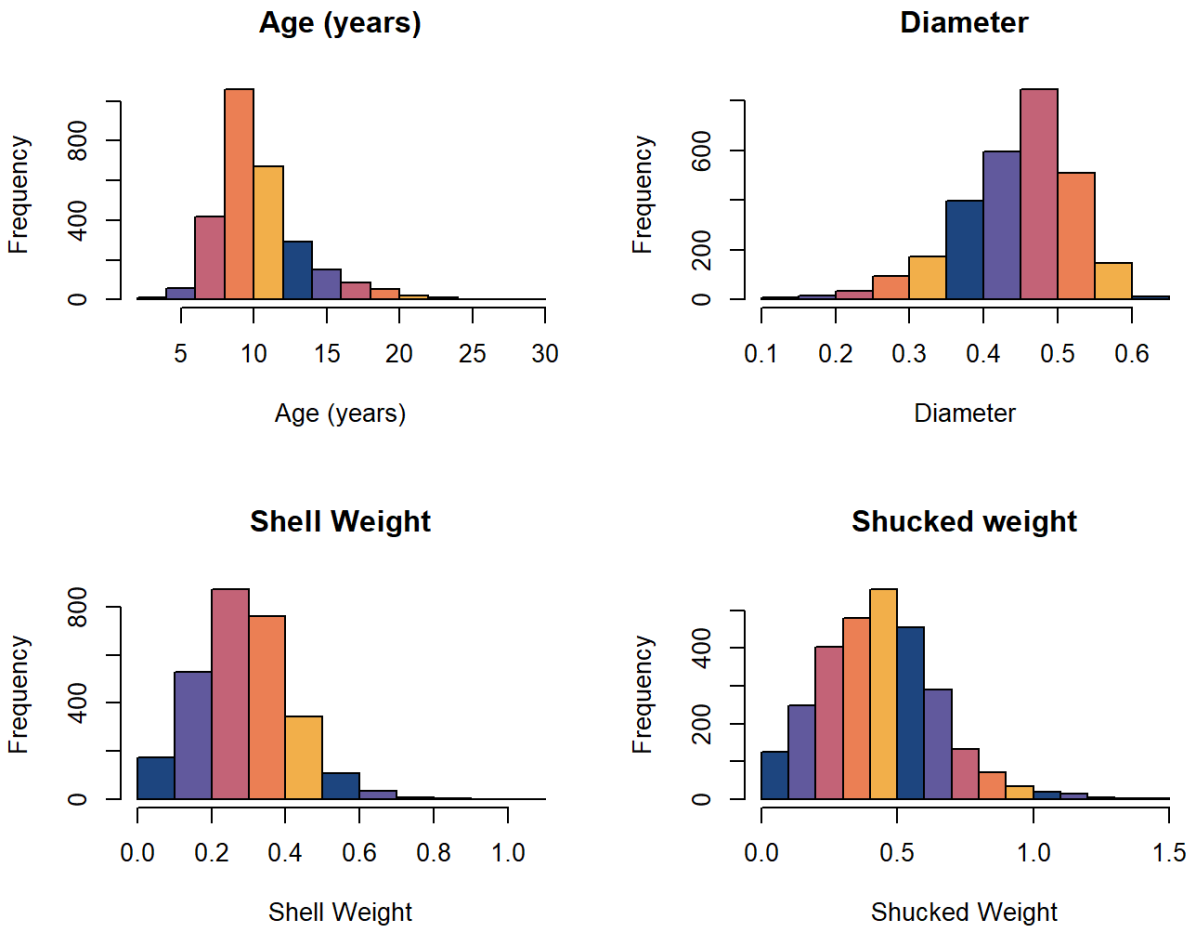


Fig 3: Histograms of continuous variables (age, diameter, shell weight, and shucked weight).

The three numeric predictors - diameter, shell weight, and shucked weight - are all normally distributed, with some skew to the left or right. Age appears somewhat normal, but age is a positive, discrete variable, so a Poisson distribution may be more appropriate.

c. Assessing variance in response variable among categorical groups:

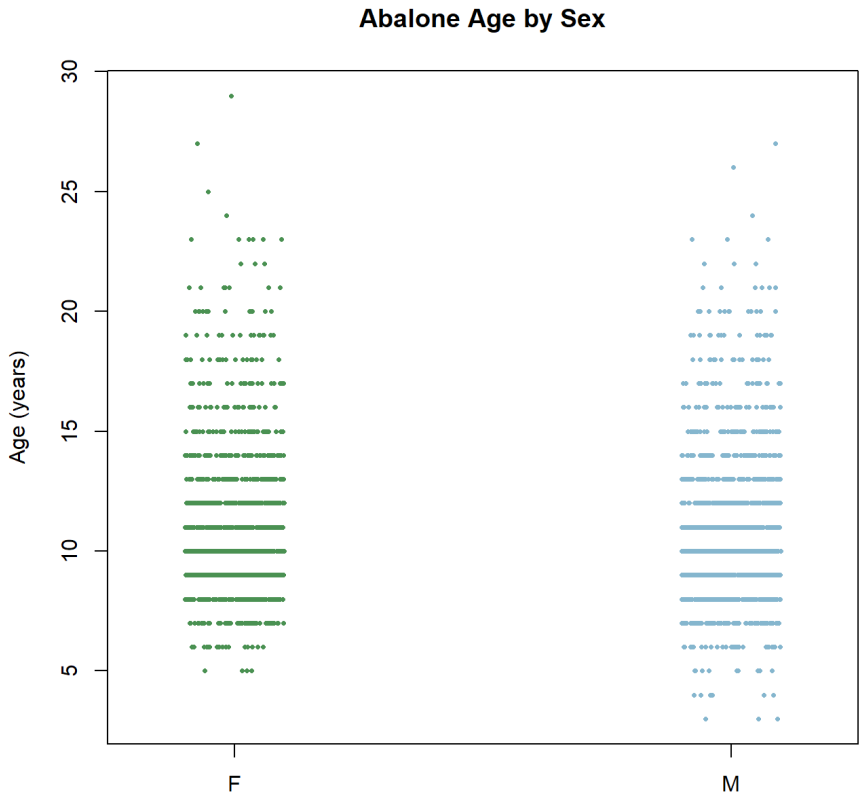


Fig 4: Stripchart of abalone age by sex. Points jittered for easier visualization.

Variance in age is similar between male and female abalone.

d. Testing for correlation between predictor variables:

	Diameter	Shell.wt	Shucked.wt
Diameter	1.0000000	0.8771149	0.873718
Shell.wt	0.8771149	1.0000000	0.822242
Shucked.wt	0.8737180	0.8222420	1.000000

Table 1: Pearson correlation test results comparing diameter, shell weight, and shucked weight.

The predictors are all strongly correlated to each other, as shown by the r values from Pearson correlation tests in the table above. However, since we are exploring main effects only for the purposes of the exams, I will omit any interactions from my candidate models.

e. Assessing log transformations of predictor variables:

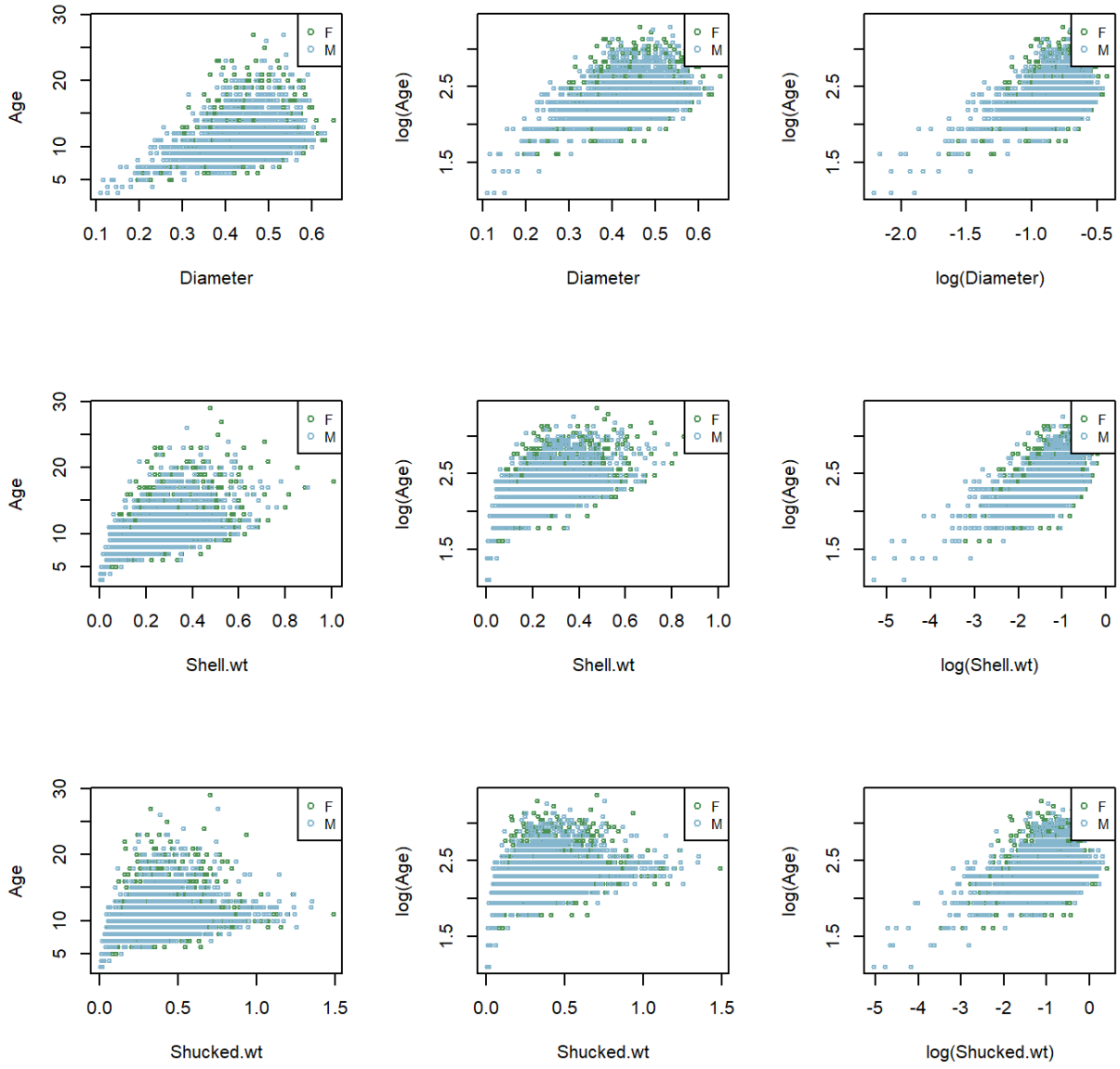


Fig 5: Scatterplots of predictor variables (diameter, shell weight, and shucked weight) and age. The first column shows raw data; the second column shows log-transformed age; the third column shows log-transformed response and predictor.

Relationships between variables and log of age (column 2) are a little less funneled and appear more linear/logarithmic in pattern than the raw data (column 1). Taking the log of both x and y variables (column 3) reveals a more linear pattern between the log of each predictor and the log response. However, this becomes very difficult to interpret, and since the predictors are already normally distributed, there is no need to log-transform them.

2. Model Selection.
    a. Full Model:

$$Age \sim Diameter + Shell.wt + Shucked.wt + Sex$$

```
m1 <- glm(Age ~ Diameter + Shell.wt + Shucked.wt + Sex, data = aba, family =
"poisson")
summary(m1)
```

```
##
## Call:
## glm(formula = Age ~ Diameter + Shell.wt + Shucked.wt + Sex, family = "pois
son",
##     data = aba)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.8288218  0.0572745  31.931  < 2e-16 ***
## Diameter     1.0987078  0.1859380   5.909 3.44e-09 ***
## Shell.wt     1.6569206  0.0834098  19.865  < 2e-16 ***
## Shucked.wt  -0.9688535  0.0576405 -16.809  < 2e-16 ***
## SexM        -0.0009045  0.0115182  -0.079    0.937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2272.1  on 2834  degrees of freedom
## Residual deviance: 1428.6  on 2830  degrees of freedom
## AIC: 13367
##
## Number of Fisher Scoring iterations: 4
```

```
anova(m1)
```

```
## Analysis of Deviance Table
## Model: poisson, link: log
##
## Response: Age
```

```
## Terms added sequentially (first to last)
##             Df Deviance Resid. Df Resid. Dev
## NULL                         2834     2272.2
## Diameter     1  289.779       2833     1982.4
## Shell.wt     1  256.662       2832     1725.7
## Shucked.wt   1  297.076       2831     1428.6
## Sex          1    0.006       2830     1428.6
```

Three variables are contributing significantly to the model: diameter ($z = 5.91, p < 0.001$), shell weight ($z = 19.86, p < 0.001$), and shucked weight ($z = -16.81, p < 0.001$). The categorical variable of sex did not affect the model fit ($z = -0.08, p < 0.94$). However, I will include sex in some model variations as it is a variable of interest (per instructions, gonad type may affect the relationship between age and size, and we were told to include sex in the model options).

	VIF
Diameter	6.119518
Shell.wt	3.927146
Shucked.wt	4.359157
Sex	1.021480

Table 2: VIF test for collinearity between predictor variables.

There is some indication that Diameter is slightly collinear with the other explanatory variables (VIF = 6.12), but with all values lower than 10, we can assume no strong collinearity and continue with the model.

Checking for overdispersion:

```
D <- deviance(m1)
degf <- summary(m1)$df[2]
phi <- D/degf
phi
```

```
## [1] 0.5048138
```

```
# using Chi^2 test
pp <- sum(resid(m1, type = "pearson")^2)
1 - pchisq(pp, m1$df.resid)
```

```
## [1] 1
```

There is no overdispersion in the model ($\phi <\approx 1$).

b. Model variations:

	df	AIC	formula
m1	5	13366.79	$Age \sim Diameter + Shell.wt + Shucked.wt + Sex$
m1.1	4	13364.80	$Age \sim Diameter + Shell.wt + Shucked.wt$
m2	4	13657.65	$Age \sim Diameter + Shell.wt + Sex$
m2.1	3	13659.87	$Age \sim Diameter + Shell.wt$
m3	4	13731.31	$Age \sim Diameter + Shucked.wt + Sex$
m3.1	3	13729.42	$Age \sim Diameter + Shucked.wt$
m4	4	13400.22	$Age \sim Shucked.wt + Shell.wt + Sex$
m4.1	3	13398.79	$Age \sim Shucked.wt + Shell.wt$
m5	3	13913.16	$Age \sim Diameter + Sex$
m5.1	3	13913.16	$Age \sim Diameter$
m6	2	13686.60	$Age \sim Shell.wt + Sex$
m6.1	3	13685.58	$Age \sim Shell.wt$
m7	3	14128.00	$Age \sim Shucked.wt + Sex$
m7.1	2	14136.07	$Age \sim Shucked.wt$

Table 3: Model variations with associated AIC value and degrees of freedom (df).

By AIC comparison, the best fit model includes main effects of diameter, shell weight, and shucked weight, excluding sex. However, its AIC is very slightly lower than the full model (including sex), and there is biological reasoning to include sex despite its low significance. Moreover, with an n of 2835, there are plenty of data to justify including more predictor variables. I ultimately settled on the full model ($Age \sim Diameter + Shell.wt + Shucked.wt + Sex$) as the best model.

c. Compare to Gaussian model:

To be sure that a Poisson distribution was the best for the model, I compared my GLM to a Gaussian model.

```
m1.2 <- glm(Age ~ Diameter + Shell.wt + Shucked.wt + Sex, data = aba, family
= "gaussian")

summary(m1.2)
```

```
##

## Call:

## glm(formula = Age ~ Diameter + Shell.wt + Shucked.wt + Sex, family = "gaus
sian",

##     data = aba)

##

## Coefficients:

##              Estimate Std. Error t value Pr(>|t|)

## (Intercept)   5.87705    0.44079  13.333  < 2e-16 ***

## Diameter      7.90255    1.47111   5.372 8.43e-08 ***

## Shell.wt     21.72073    0.76459  28.408  < 2e-16 ***

## Shucked.wt  -10.98454    0.46119 -23.818  < 2e-16 ***

## SexM         -0.01190    0.09317  -0.128    0.898

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for gaussian family taken to be 5.979576)

##

##     Null deviance: 26697  on 2834  degrees of freedom

## Residual deviance: 16922  on 2830  degrees of freedom

## AIC: 13122

##

## Number of Fisher Scoring iterations: 2
```

The Gaussian-fit model has a lower AIC (13122) than the original Poisson model (13367). However, comparing AIC between Gaussian and Poisson GLMs is controversial given the different fitting of each model, so I used log likelihood as well.

```
logLik(m1)
## 'log Lik.' -6678.396 (df=5)
logLik(m1.2)
## 'log Lik.' -6555.174 (df=6)
```

The Gaussian model has a higher log likelihood than the Poisson, indicating that a normal distribution might be a better fit for the age data. Nonetheless, since age is a positive, discrete numeric variable and the difference in log likelihood is relatively small, I will continue assessing the Poisson model:

$$Age \sim Diameter + Shell.wt + Shucked.wt + Sex$$

3. Assessing model fit and assumptions.
    a. Diagnostic plots/residual assessment:

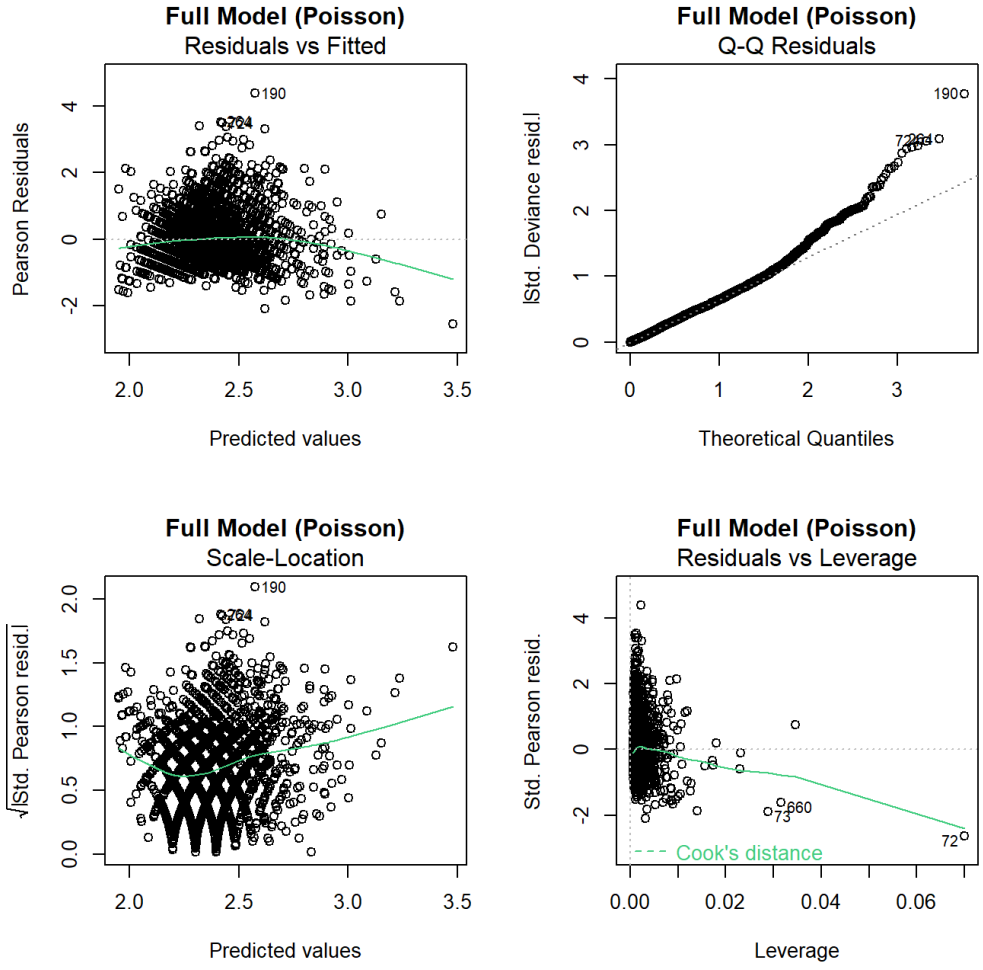


Fig. 6: Diagnostic plots of the GLM.

The residuals from the full model show distinct patterns. There is a strong clumping of points in the Pearson residuals, and the upper tail in the QQ plot strays significantly. These plots show that the model is not a very good fit for the data; if this were a more complete project, significant work would be needed to improve the model.

Just to be sure, I compared the residual plots with the Gaussian model, and found similar patterns.

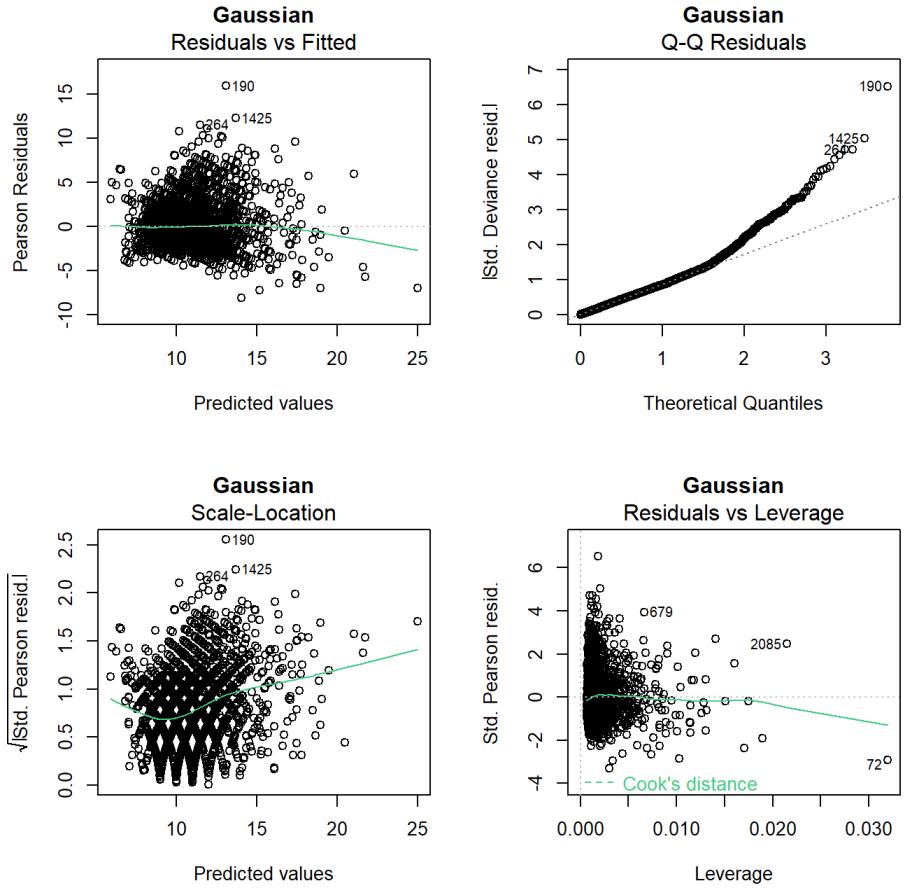


Fig. 6: Diagnostic plots for the Gaussian LM.

b. Goodness of fit – pseudo-R^2:

```
# goodness of fit: pseudo-R^2
pR2 <- 1 - (m1$deviance / m1$null)
pR2
## [1] 0.3712452
```

The model explains approximately 37.1% of the variance in the data ($pseudo\ R^2 = 0.371$).

4. Visualizing model predictions.
    a. Added-Variable Plots:

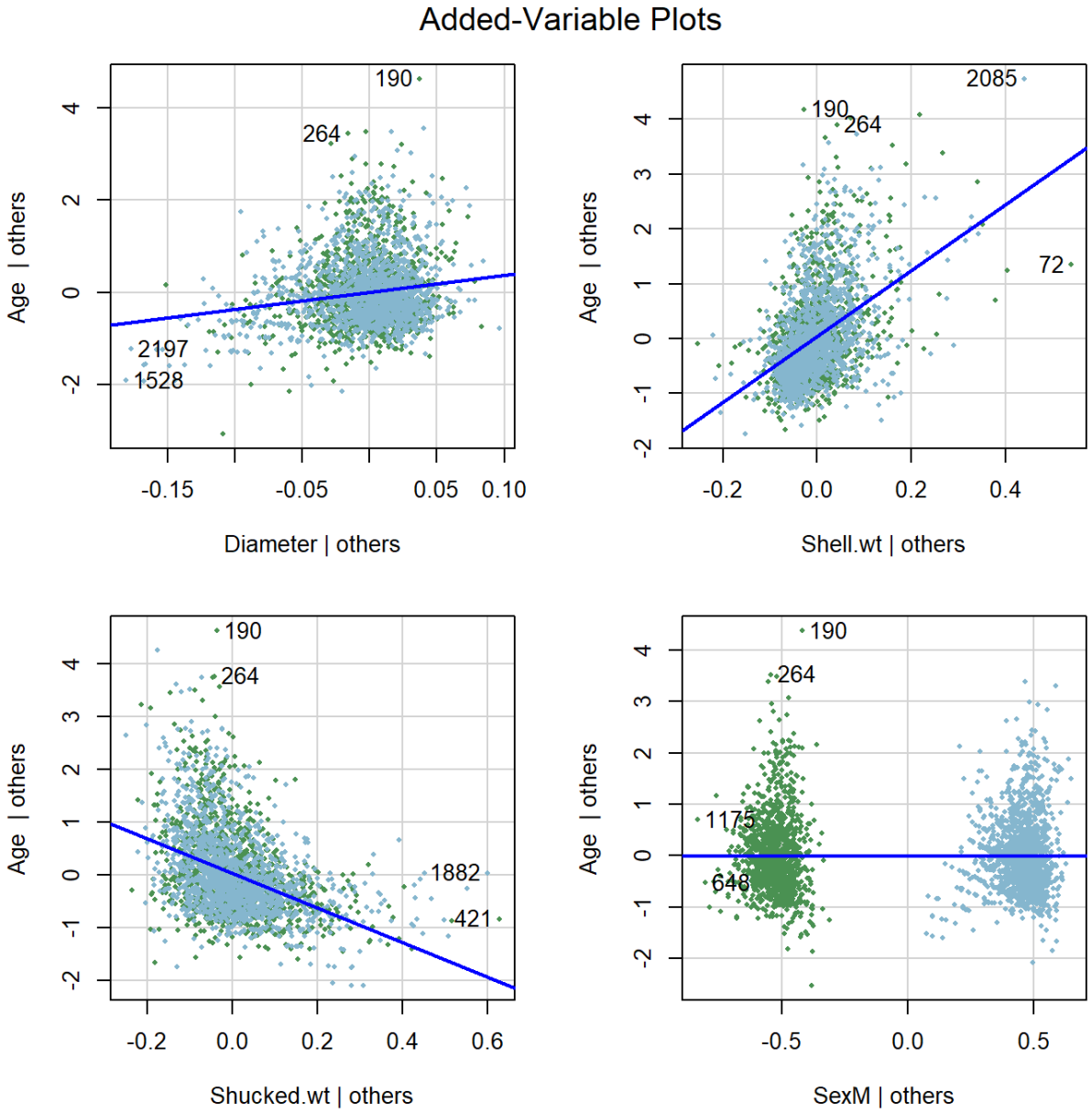


Fig 7: Added-variable plots showing residuals from predictions for each predictor variable, holding all others constant at their means.

Plotting the residuals with Added-Variable plots highlights the direction and strength of each variable's relationship to age, holding all others constant at their means. Diameter has a slight, positive relationship to age; shell weight has the strongest, positive influence; shucked weight has a weaker and negative relationship.

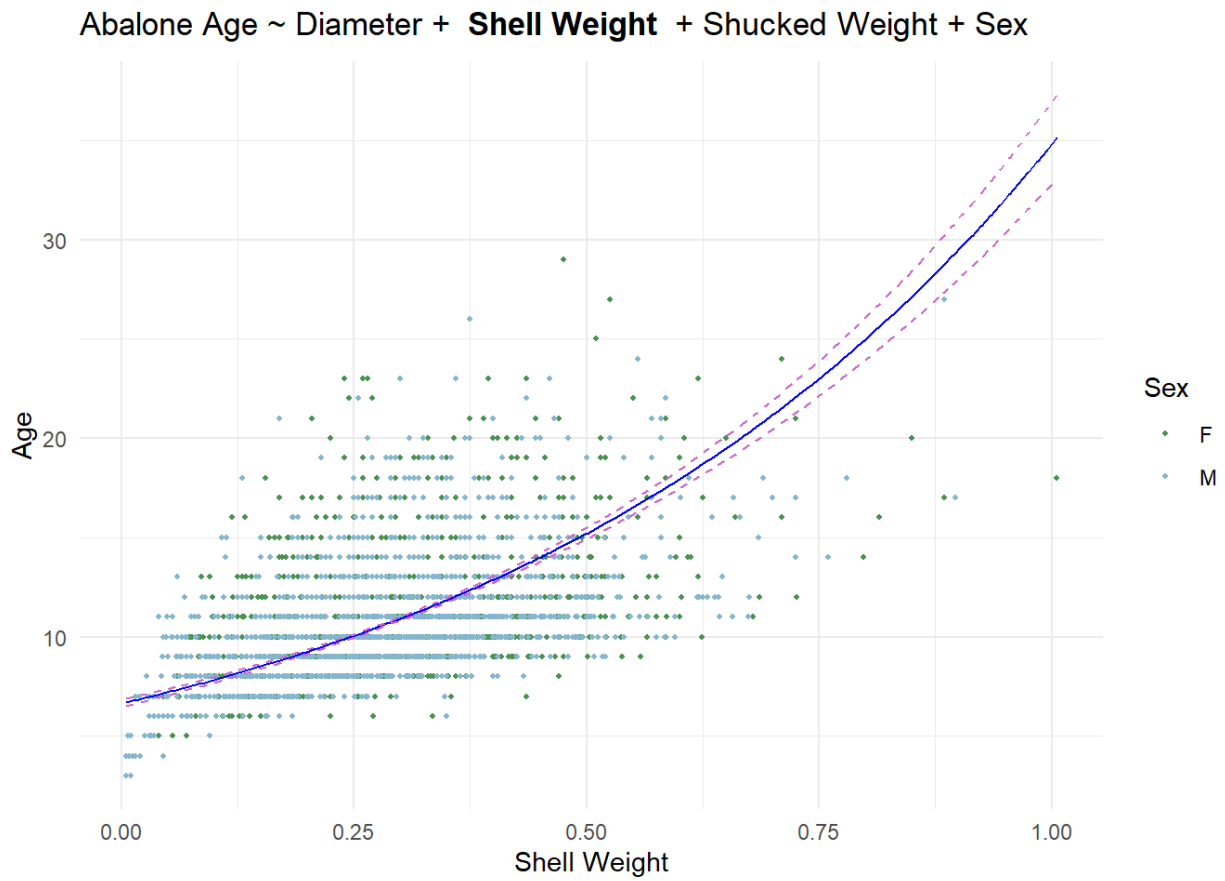


Fig 8: Relationship between abalone age and shell weight, colored by sex.

Figures 8 through 10 show real data for abalone age by each main effect - diameter, shell weight, and shucked weight - as well as sex. Blue lines show exponentially transformed estimated marginal means from the model across the range of predictor values; purple lines show transformed EMM +/- standard error. Values were transformed per the log link in a Poisson generalized linear model.

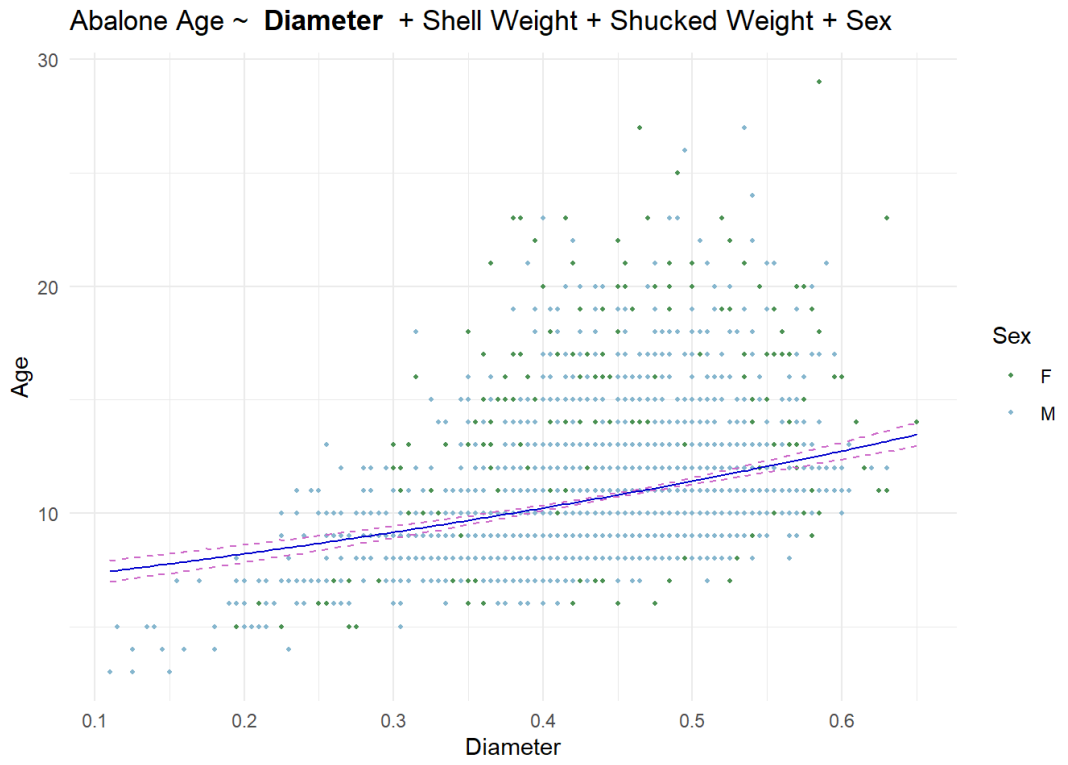
Conclusion:

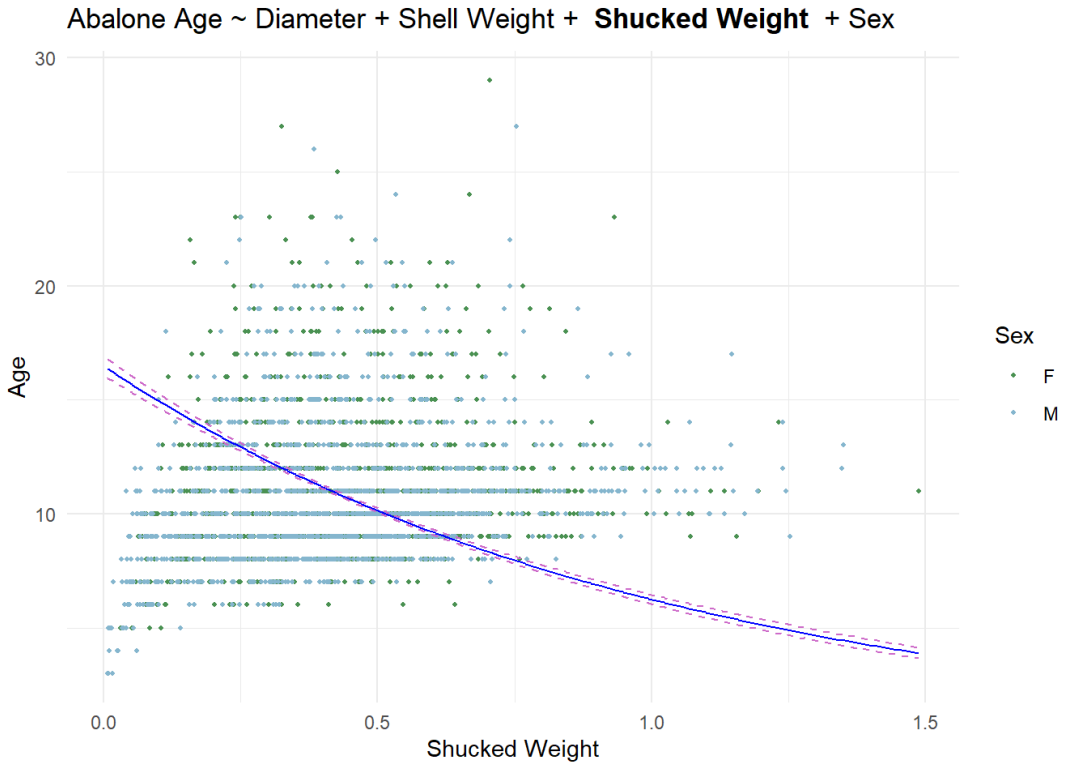
Abalone age can be described by the generalized linear model

$$Age \sim Diameter + Shell\ Weight + Shucked\ Weight + Sex$$

fitted with a Poisson distribution ($intercept = 1.83,\ se = 0.057,\ z = 31.9,\ p < 0.001$). Shell weight had the strongest relationship to age ($slope = 1.657,\ se = 0.083,\ z = 19.87,\ p < 0.001$), followed by shucked weight ($slope = -0.969,\ se = 0.058,\ z = -16.81,\ p < 0.001$), then diameter ($slope = 1.099,\ se = 0.186,\ z = 5.91,\ p < 0.001$). Although I kept it in my final model for biological/assignment reasons, sex did not contribute significantly ($slope = -0.0009,\ z = -0.079,\ p < 0.937$).

The model is not a great fit, only explaining 37.1% of the variance in the data ($pseudoR^2 = 0.371,\ G^2 = 1$). This can also be seen best in the plot of Age vs. Shucked Weight, where the model poorly predicts age at low or high shucked weights (see below).





Figs. 9 & 10: Relationship between abalone age and diameter (9) and shucked weight (10), colored by sex.

+5

Make a plot showing your model predictions with a caption that describes what your plot depicts. Predictions must be estimated using the `predict()` function. Credit will not be given for using `geom_smooth()` in `ggplot()` unless you have modified this function to plot your model fit (e.g., no “method=lm” or the default loess, unless you can clearly justify why this would be a true representation of your model).

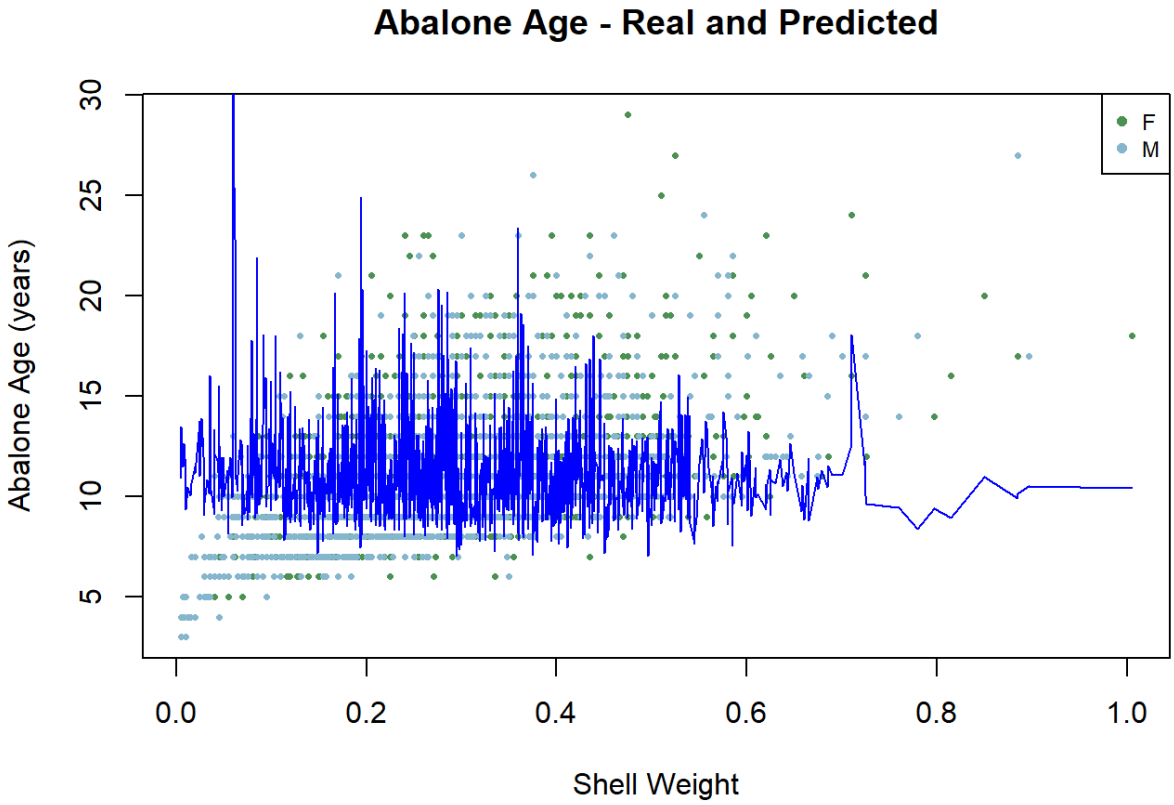


Fig 11: Model predictions for abalone age, shown along range of shell weight. Points show real data, colored by sex. Blue line shows model predictions for each value of shell weight with associated predictor values. The erratic nature of the line and the wide variation from the real data show that this model could be improved.

Assessment (1 pt each)

How long did it take you to do this exam?

About 10 hours total, split over the last few days (and including the 2 hours in class, which were very helpful to get started). At least one hour of that was struggling with palettes and colors after finishing.

How fair do you think this exam was (1-5, 5 being fair)?

5

Looking back over the course, what do you wish we had covered in class better?

Some more time with GLMs and getting more into GLMMs. Multiple comparisons and limits to analyzing data from too many directions.

What is the most important thing you learned this quarter?

To be critical of the statistics I am reading in published studies, as many researchers are not analyzing their data properly; and in the same vein, to be highly cautious of my own approach and to document the whole process.