STAT 230 Homework 2

Owen Forman

1. (Manatees)

Load the following dataset:

```
manatees <- read.csv("https://www.math.carleton.edu/ckelling/data/manatees.csv", header = TRUE)
```

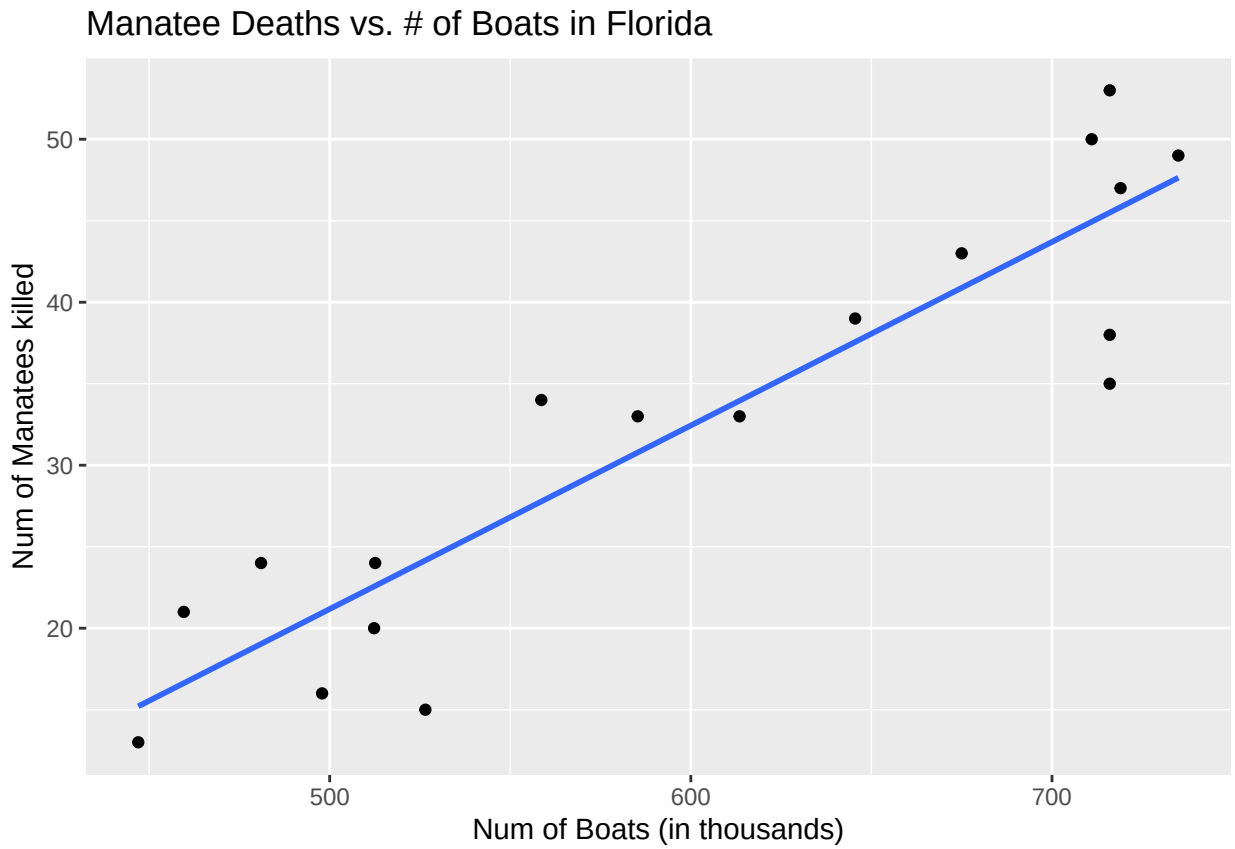
The data records the number of manatees killed by powerboats and the number of powerboats registered in Florida (in thousands) from 1982 to 2000. ;-(

(a) Make a plot of the number manatee deaths vs. number of powerboats and add the least squares line to the plot. Comment on the apparent relationship (shape, strength, and direction of association).

Answer: Visually, the relationship appears relatively linear, with a decently strong positive correlation.

```
#code with comments here

#loads libraries used
library(ggplot2)


#creates plot of killed vs boats, adds points, and least squares line
ggplot(data = manatees) + aes(x = boats, y = killed) + geom_point() + geom_smooth(method = "lm", se = F
  labs(title = "Manatee Deaths vs. # of Boats in Florida", x = "Num of Boats (in thousands)", y = "Num
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



(b) The year 1991 is missing from this dataset. An estimate of the number of powerboats in that year, by averaging the values from 1990 and 1992, is 599,400 boats. Assuming that to be the number of boats in 1991, give a 90% prediction interval for the number of manatees killed in that year and interpret the interval. **Answer:** We are 90% confident that in 1991 the number of manatees killed in Florida by powerboats was between 22.666 and 42.073 (rounded to the nearest thousandth)

```
#code with comments here

#creates fitted regression line
manatee_lm = lm(killed ~ boats, data = manatees)

#forms a prediction for # of killed manatees when there were 599.4 (thousand) boats
prediction <- predict(manatee_lm, newdata = list(boats = 599.4), interval = "prediction", se = TRUE, le

#prints prediction and interval
prediction
```

```
## $fit
##        fit      lwr      upr
## 1 32.36963 22.66632 42.07294
##
## $se.fit
## [1] 1.275324
##
## $df
## [1] 16
##
```

```
## $residual.scale
## [1] 5.40952
```

(c) Since we don't actually know the number of boats registered in 1991, do you think that accounting for this additional source of uncertainty will increase or decrease the width of the prediction interval you calculated in part (b)?

Answer: I think it will increase the width of the prediction interval. As one adds uncertainty the width of the prediction interval will increase since it must now cover a larger span to maintain the 90% accuracy of the prediction.

(d) Do the assumptions of linearity, constant variance, and normality appear to be appropriate for these data? Justify your opinion. (Note: I am not asking you to check the independence assumption.)

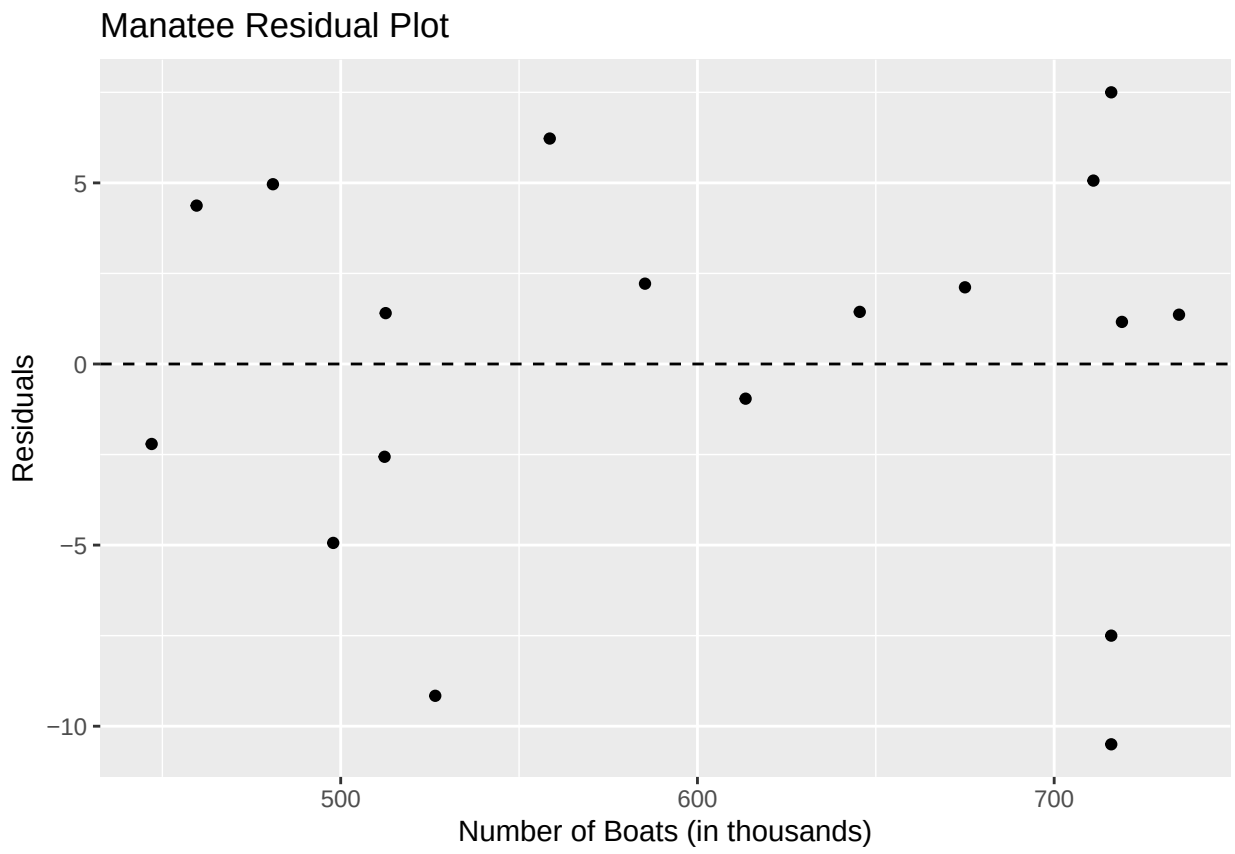
Answer: We can check our assumptions of linearity and constant variance by observing the residual plot shown below. Although we cannot be 100% sure, there does not appear to be a discernible pattern within the plot, which instead appears to have random scattering above and below the y = 0 line. This agrees with our assumption of linearity and constant variance.

Next, to check the assumption of normality, we can observe the Q-Q plot also shown below. From observation we can see that our sample quantiles do not follow the expected normal quantiles perfectly, however this is likely due to our small sample size. We haven't gone over how to simulate from a random of size n distribution so that analysis is not included here. Since, using the information we have available to us, the answer is unclear, we cannot make an accurate assessment of whether or not our assumption of normality holds.

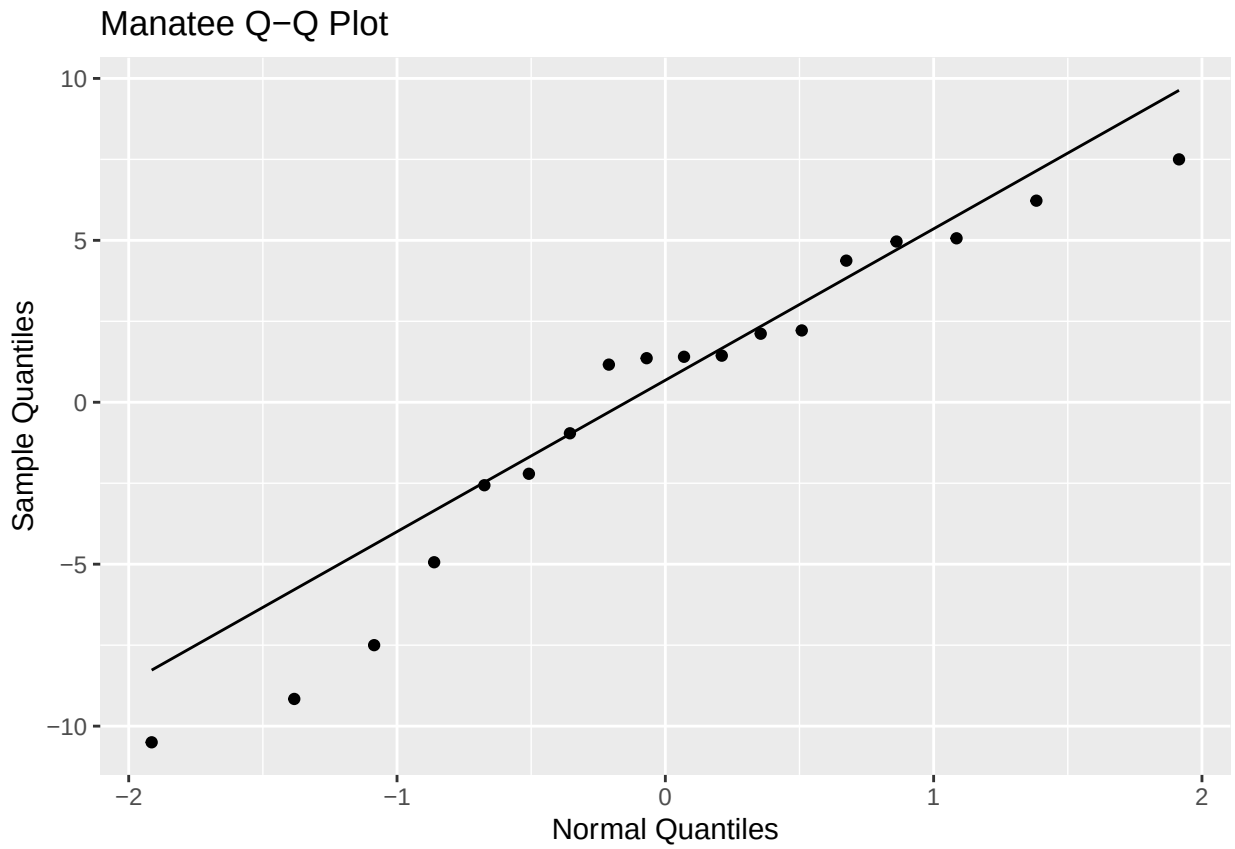
Finally, we must consider the assumption of independence. It is extremely hard to say, especially because we don't have knowledge about how the data was collected. The only possible reason they wouldn't be independent that I can think of is that the number of manatees killed in the first year will slightly affect the manatee population in the second year, which could have an affect on the number of manatees killed in year 2. This would cause the data points collected to be slightly (although probably marginal to the point of it not mattering) affected by previous points, which would make them not truly independent.

To illustrate the point consider the example: imagine there were 100 manatees in year 1 (the first year of recordered data) and 10 of them get killed by boats. There'd then be 90 manatees that make it to year 2 plus however many manatees are born, for the sake of the example say they perfectly couple up and all have kids, so there would then be 135 manatees in year 2. However if no manatees were killed though we'd have 100 that survive and 50 newborn manatees for a total of 150 manatees. The number of boats in year 2 is the same across both examples, but the number of manatees are different. So, it is possible that the number of manatees killed in year 1 slightly affects the number of manatees killed in year 2. While there is no (given) proof that there is a correlation between more manatees and more manatees killed, it logically follows that it is, at the very, least a possibility. Thus we cannot claim with certainty that the assumption of independence holds when there's a chance it doesn't.

```r
#code with comments here
#loads necessary library
library(broom)

#augment dataset to include resid values
manatee_aug <- augment(manatee_lm)

#resid plot
ggplot(manatee_aug, aes(x = boats, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Manatee Residual Plot", y= "Residuals", x = "Number of Boats (in thousands)")
```




```
#Q-Q plot
ggplot(manatee_aug, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Manatee Q-Q Plot", y = "Sample Quantiles", x = "Normal Quantiles")
```

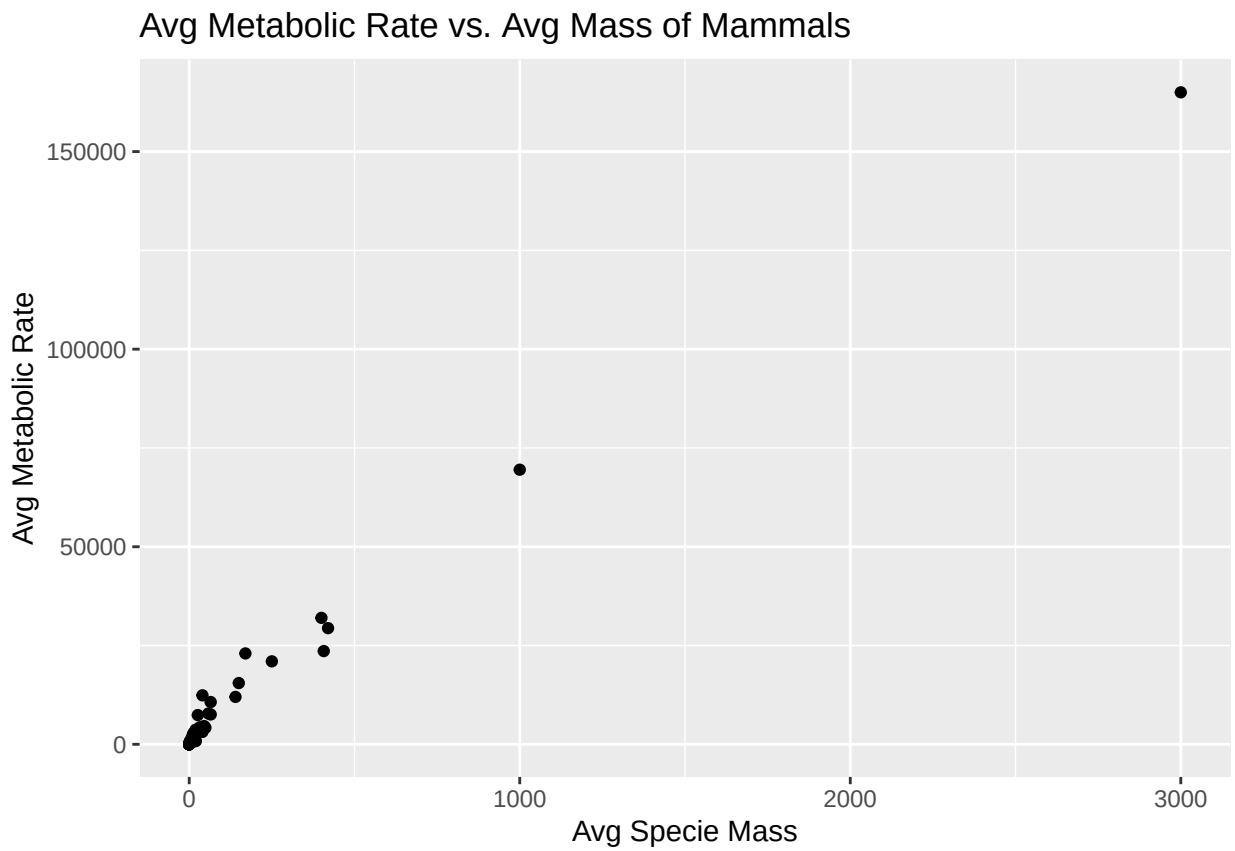


2. *Note*: the following problem involves log transformations, the topic of Friday's class. You should, however, be able to complete the question before Friday's class since the question only asks you to consider interpreting the regression diagnostic plots and R^2 value (from Wednesday) rather than interpreting the model itself. You can think about it as preparation for Friday's class if you choose to do it before class.

This is Adaptation of Exercise 8.26 (the data is `ex0826` in the `Sleuth3` R package). First read the problem statement for Exercise 8.26 to understand what the data is.

(a) Make a plot of average metabolic rate vs. mass for the 95 animals in this dataset. Make sure the axes are appropriately labeled. Comment on the relationship (shape, strength, direction of association). **Answer:** It is hard to comment on the graph because there are lots of data points clumped around 0, and a few extreme outliers. That being said, from what I can visually tell there does appear to be a strong positive correlation between Mass and Avg Metabolic Rate. The relationship shape is extremely hard to tell because of given reasons, but if I had to make an assessment based on the visual it appears to be at least a little curved (aka not linear).

```
#code with comments here

#loads library for problem
library(Sleuth3)

#simple scatterplot of Avg Metabolic Rate vs Avg Mass of Mammals
ggplot(data = ex0826, aes(x = Mass, y = Metab)) +
  geom_point() +
  labs(title = "Avg Metabolic Rate vs. Avg Mass of Mammals", x = "Avg Specie Mass", y = "Avg Metabolic
```

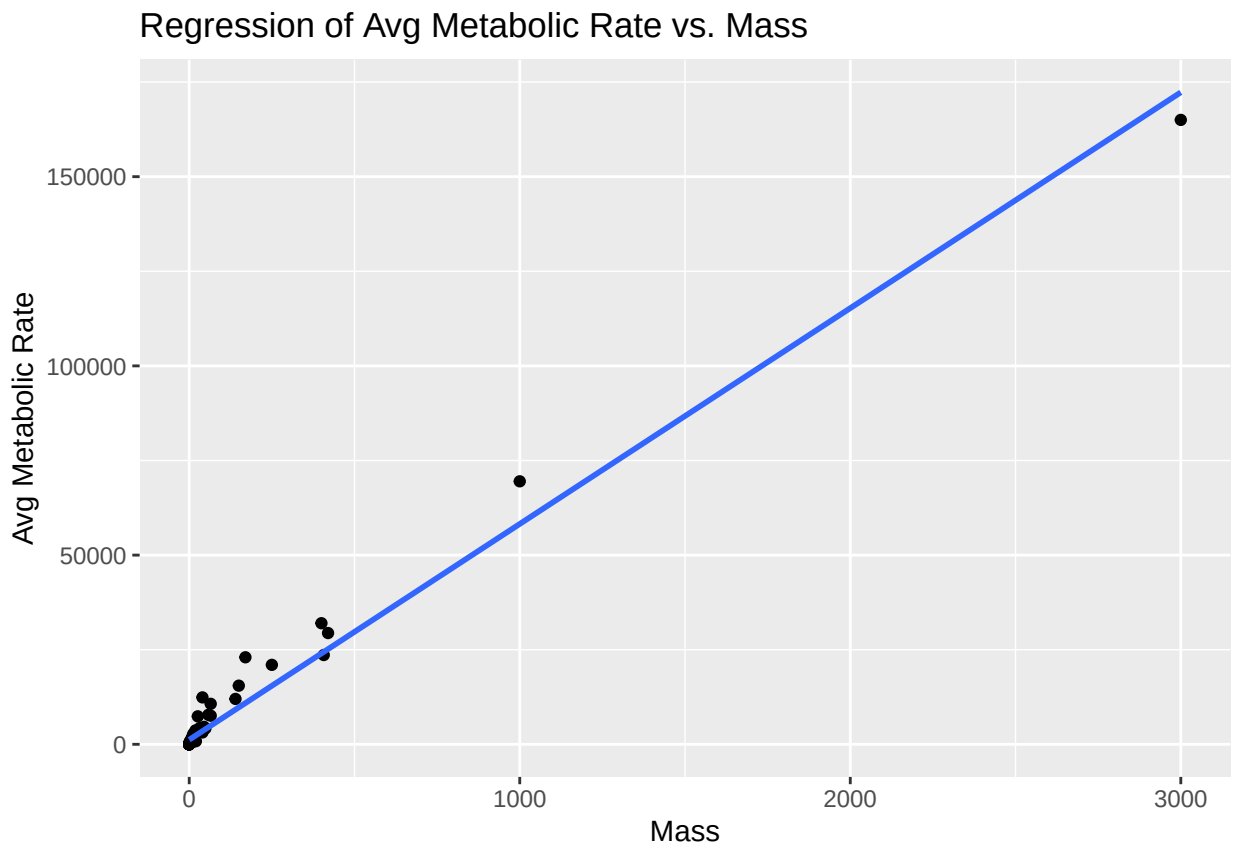


(b) Fit the simple linear model of metabolic rate against mass. (You do not need to report or interpret the parameter estimates.) Give the R^2 value for this model and interpret this number. On the basis of the R^2 value, is this model "good"?

Answer: The R^2 value is .9788. The interpretation of this that approx 97.88% of the variation in a mammals observed avg metabolic rate can be explained by its mass. On the basis of the given R^2 value, this model is exceptionally "good" HOWEVER, because of what was mentioned in part a) that there are massive outliers and the graph doesn't appear to be linear, so we cannot interpret or use this R^2 statistic.

```
#code with comments here

#prints plot with linear regression line
ggplot(data = ex0826, aes(x = Mass, y = Metab)) +
  geom_point() +
  labs(title = "Regression of Avg Metabolic Rate vs. Mass", x = "Mass", y = "Avg Metabolic Rate") +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



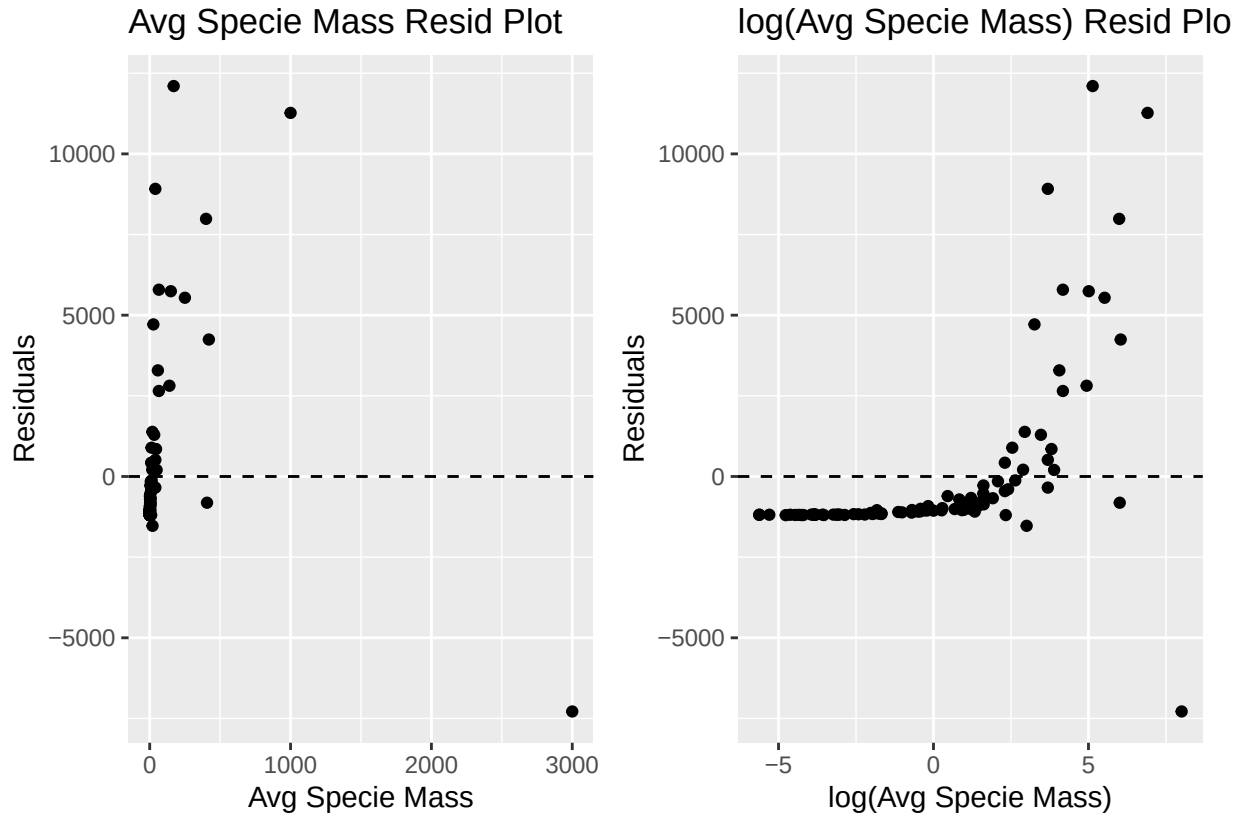

```
#fits linear model onto data
ex0826_lm <- lm(Metab ~ Mass, data = ex0826)

#prints R^2 statistic
summary(ex0826_lm)$r.squared
```

```
## [1] 0.9788373
```

(c) Plot the residuals from the model in part (b) vs. the mass. To see what is going on more easily, also plot the residuals vs. the log(mass). Are the standard regression assumptions satisfied? **Answer:** No, the standard regression assumptions, namely linearity and constant variance are NOT satisfied. In the unaugmented (non log-ged) residual plot the points are clumped up right next to eachother with no apparent random variance and one or two major outliers in the far corners of the plot, and the augmented (log-ed) residual plot the points follow a clear pattern (looks like they're following some kind of exponential curve).

```
#code with comments here

#loads patchwork library
library(patchwork)

#augments linear model
ex0826_aug <- augment(ex0826_lm)

#plots resid plot
residvmass <- ggplot(ex0826_aug, aes(x = Mass, y = .resid)) +
  geom_point() +
```

```
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Avg Specie Mass Resid Plot", y= "Residuals", x = "Avg Specie Mass")

#plots augmented resid plot (log(Mass))
residvmass_log <- ggplot(ex0826_aug, aes(x = log(Mass), y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "log(Avg Specie Mass) Resid Plot ", y= "Residuals", x = "log(Avg Specie Mass)")

residvmass + residvmass_log
```



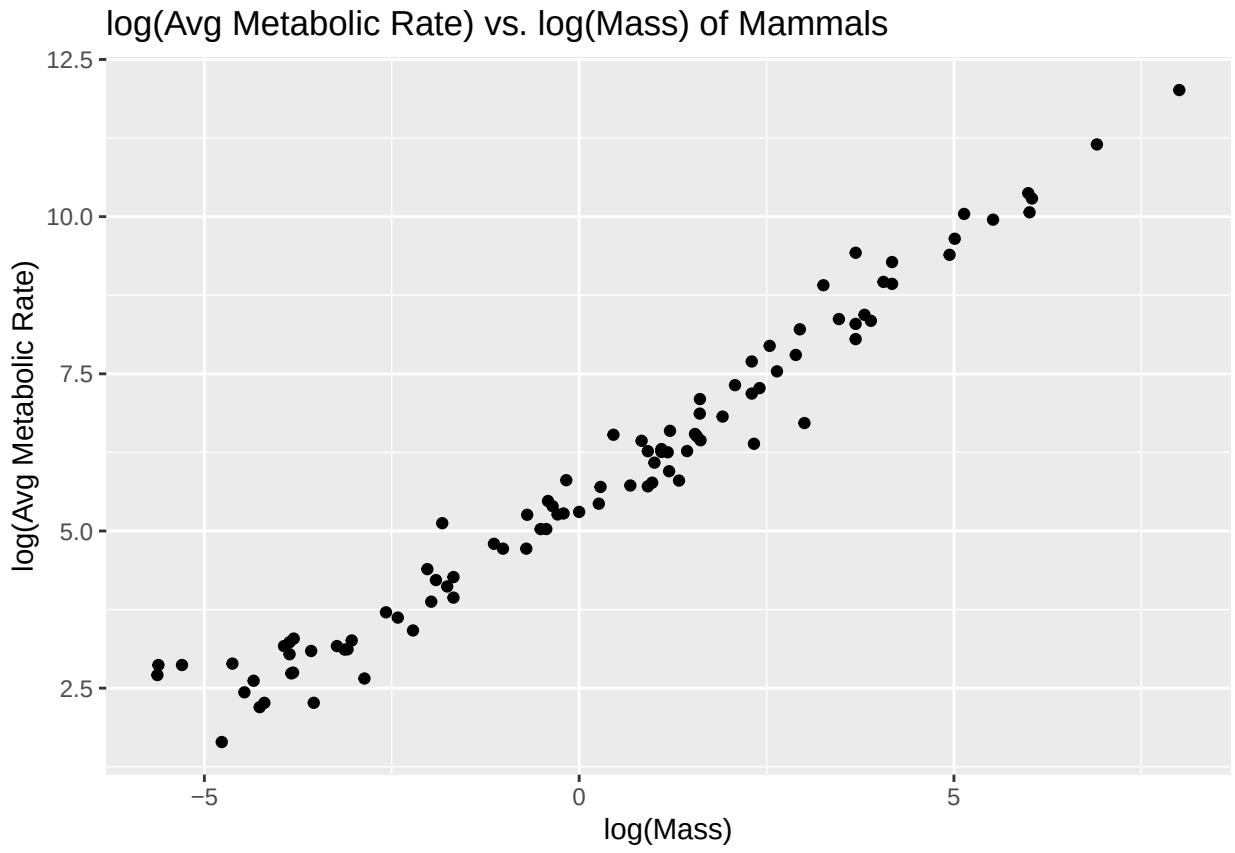
Note: to show two plots side-by-side, you can use the `patchwork` library to combine plots made using `ggplot`. You can use this code to produce the plots:

```
#note that this code doesn't run because of "eval = FALSE" above

# fit model from part (b)
SLR_model <- lm(Metab ~ Mass, data = ex0826)
SLR_aug <- augment(SLR_model)

# make plots for part (c)
library(patchwork)
residVsMass <- ggplot(SLR_aug, aes(x = Mass, y = .resid)) + geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Mass (kg)", y = "residuals")
```

```
residVsLogMass <- ggplot(SLR_aug, aes(x = log(Mass), y = .resid)) + geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "log(Mass)", y = "residuals")

residVsMass + residVsLogMass
```

(d) Make a plot of the log(metablic rate) vs. the log(mass). Compare with the plot in (a). **Answer:** Compared to the plot of in (a) this plot not only looks more linear, but is significantly easier to read and comprehend because the axis's are several orders of magnitude smaller which reins in the outlier cases to be much closer (visually) to the rest of the data.

```
#code with comments here

ggplot(data = ex0826, aes(x = log(Mass), y = log(Metab))) +
  geom_point() +
  labs(title = "log(Avg Metabolic Rate) vs. log(Mass) of Mammals", x = "log(Mass)", y = "log(Avg Metabol
```



(e) Fit the simple linear model of log(metablic rate) against log(mass). Give the R^2 value for this model. Which R^2 value is bigger: the original model, or the log-log model? **Answer:** The R value of this model is .9649 which is slightly less thanthe original model. However it is still quite a large R^2 value.

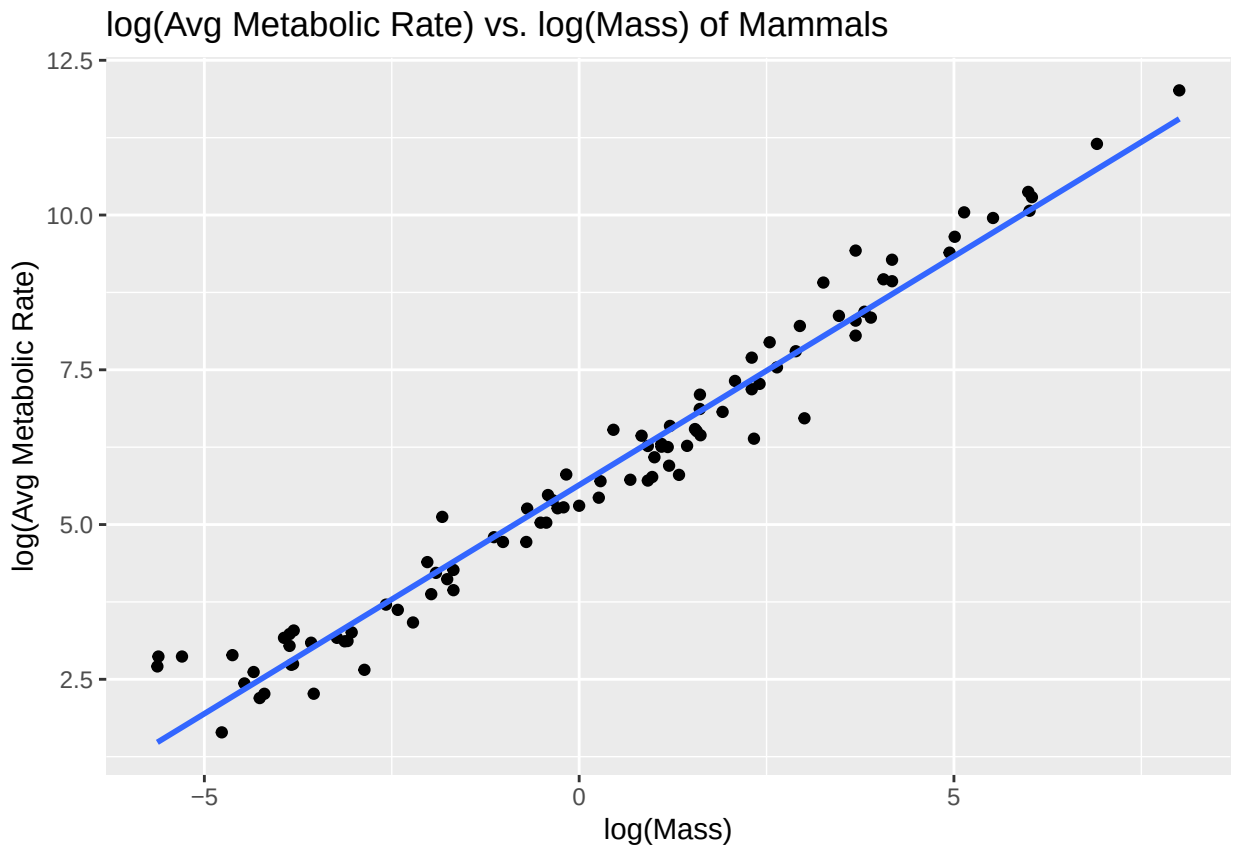
```
#code with comments here

#fits transformed model
ex0826log_lm <- lm(data = ex0826, log(Metab) ~ log(Mass))

#augments fitted model
```

```r
ex0826log_aug <- augment(ex0826log_lm, newdata = ex0826)

#prints R^2 value
summary(ex0826log_lm)$r.squared
```

```
## [1] 0.9648579
```

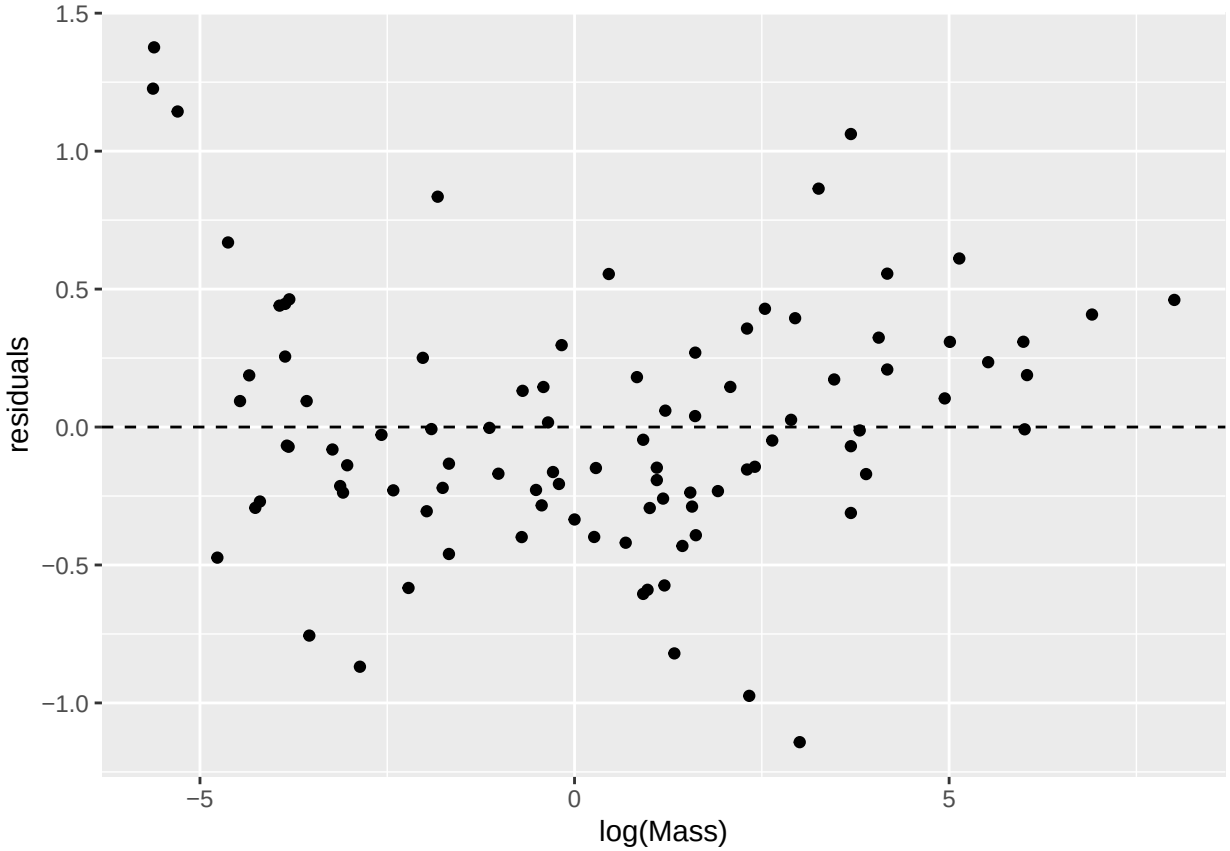
```r
#prints model with fitted regression line
ggplot(data = ex0826, aes(x = log(Mass), y = log(Metab))) +
  geom_point() +
  labs(title = "log(Avg Metabolic Rate) vs. log(Mass) of Mammals", x = "log(Mass)", y = "log(Avg Metabol
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



(f) Plot the residuals from the model in part (e) vs. the log(mass). Are the standard regression assumptions satisfied? Remember that the `augment` function needs to be given the original data if there are transformations in your model. You can use the following code to make the residual plot: **Answer:** The standard regression assumptions do appear to be satisfied given this model. There is no visual pattern to the residuals, and an equal amount of points appears to fall above and below the y=0 line

```r
#code with comments here

#prints resid model
ggplot(ex0826log_aug, aes(x = log(Mass), y = .resid)) + geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "log(Mass)", y = "residuals")
```




```
#note that this code doesn't run because of "eval = FALSE" above

loglog_model <- lm(log(Metab) ~ log(Mass), data=ex0826)
loglog_aug <- augment(loglog_model, newdata = ex0826)

ggplot(loglog_aug, aes(x = log(Mass), y = .resid)) + geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "log(Mass)", y = "residuals")
```

(g) Which model is better: the original model or the log-log model?

Answer: I believe that the log-log model is better, as the regression assumptions are met so we are confidentaly able to fit a regression line onto the data. However, it is important to note that one con of the log-log model is that interpretation's of the model are significantly less easy to come up with than the original. We haven't covered interpretation of log-log models in class at this point (I think we will Friday), but it will definitely be harder to derive than with an un-transformed model. However since the original model wasn't linear, no fitting of a regression line and interpretation of it could happen in the first place, so even if it's a little more confusing to interpret it's still "better" than no model at all.

Recommended exercises (do not turn in, some answers in book):

Chapter 8 exercises 1, 5, 10, 17, and 20