# Stat 505 Assignment 4

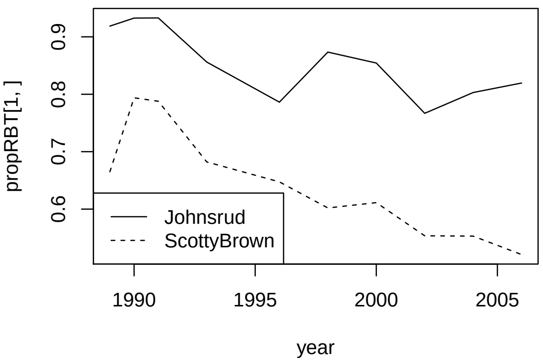
Solutions

We have data on fish caught in the Blackfoot River by Fish, Wildlife, & Parks personnel over a number of years. They used electrofishing equipment to attract the fish to the boat, then dipped them out of the water with nets, measured length in mm and weight in grams. They are often working in cold conditions in late autumn or early spring, so some measurement error is expected. This is real data, so some cleaning is necessary!

These data are not from a random sample. The goal is to catch all fish within a reach or section of the Blackfoot River every few years to assess the health of the population. Changes over years are important to the biologists.

1. Remove Bull trout and WCT (whitefish) and any fish with missing weight.

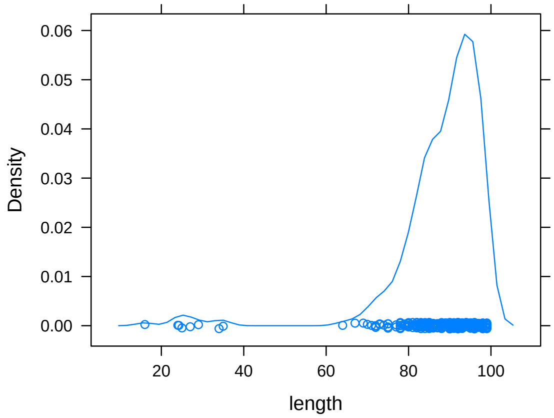
2. Use a plot to show how the proportion of RBT changes with year and section. Does one section always have a higher proportion RBT than the other?

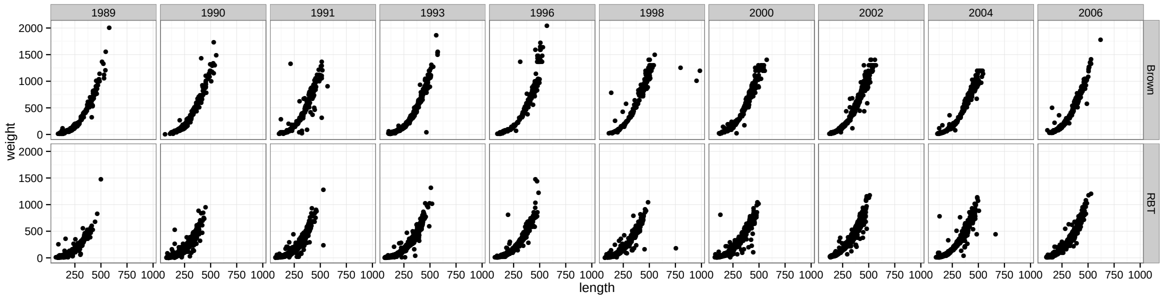


*The proportion of Rainbows is generally decreasing over time, especially at Scotty Brown. Johnsrud always has a higher proportion RBT that Scotty Brown.*

3. Plot weight as a function of length, separating into panels by species and year.

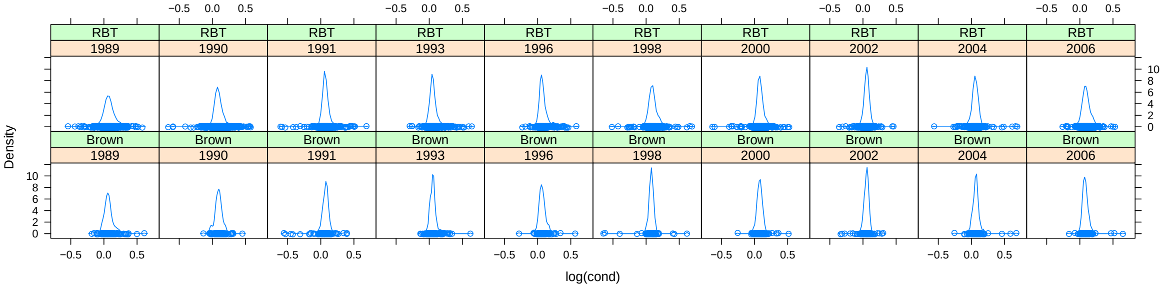
   What seems to be the lower limit of length that is catchable with this equipment? Filter out fish you deem too short to have actually been caught.



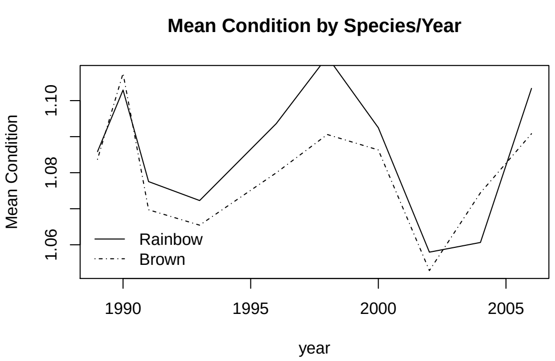


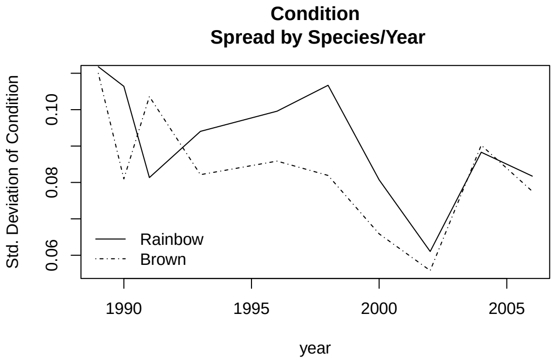
*I think that lengths less than 60 mm are probably errors because there is a large gap between 35 and 62 mm in the density plot. Very few fish are less than 100 mm (about 4 inches). I am removing fish less than 60 mm in length.*

4. Which pairs of (weight, length) combinations seem difficult to believe? Filter these out as well.
   One way to look for unusual pairs is to use what fisheries biologists call a "condition index" $w^{1/3}/l * 50$ where $w$ = weight and $l$ = length. If fish are highly unusual in this scale, it would be best to remove them, but you might need to compare only within species and within year. (If years are different in this regard, we should make a note for discussion with the biologists.)



*I filtered out 17 fish with condition over 2, because these are generally short fish with an unusually high weight. Based on the plot of weight by length, I still see some long fish with unusually low weights, so I removed 12 fish with condition less than .5. Only 3 fish had lengths over 750 mm (all were Browns caught at Johnsrud in 1998). I suspect that these could be errors, since their weights are also low, but perhaps they had just* *spawned and lost a lot of weight. In any case, they will change the relationship between length and weight, so I am going to remove them as well.*





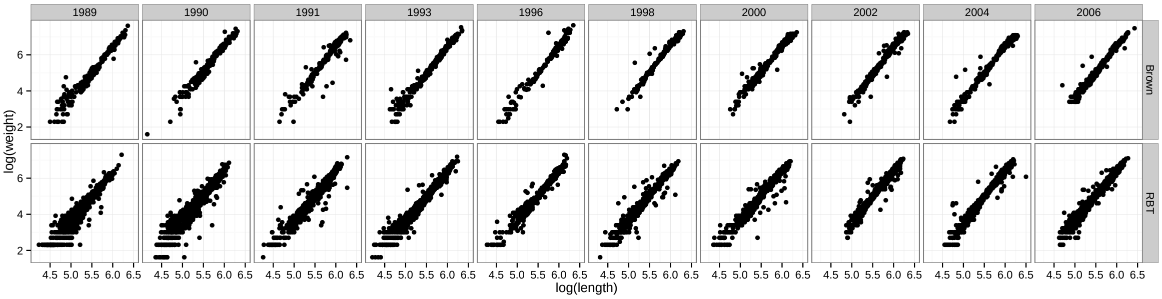
*To look for changes in condition index across years and species, I found mean and standard deviation for each species/year combo and plotted them. We can see that means for the two species move mostly in parallel in the two plots. Low values are seen in 2002 and 2004, and high values in 1990, 1998, and 2006. The pattern is a bit different for spreads, but 2002 was a year with low spread in conditions of both species.*

*I did not see a need to apply different filtering (based on conditions) for different years or species.*

5. Build a model for weight as a function of length.

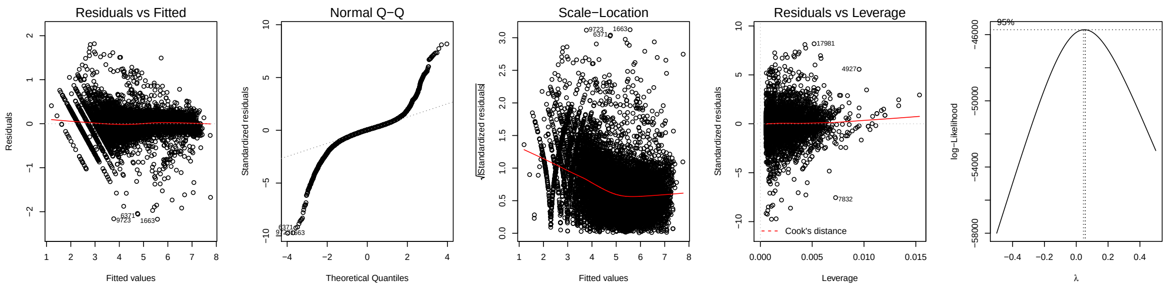
   We used a geometric argument to link tree volumes to height and girth. If fish bodies are of constant density (mass per unit volume), come up with a similar argument for what scale to use to model weight as a function of length.

   Include species and year in an appropriate way.



*I began with a model for log(weight) based on log(length), species, section, year (as a factor) and two-way interactions between log(length) and the other main effects. The anova command partitions sums of squares, giving the lion's share to log(length). Each term has a very small p-value (less than $5 \times 10^{(-8)}$), but we have a huge dataset, so some of these terms are showing statistical significance with no practical significance. I'll discuss the output in the report below.*

6. Write up your model including a discussion of the diagnostic plots.



*I won't reproduce the comments I made above, but will summarize the important points here.*

*The data are from electrofishing runs on the Blackfoot river in Montana at two sites (Johnsrud and Scotty Brown) from years 1989, 1990, 1991, 1993, 1996, 1998, 2000, 2002, 2004, and 2006. The researchers attempted a census of the fish at each time point.*

*We were asked to look at weight as a function of length, and to see if the relationship is the same for Rainbow (RBT) and Brown trout, and if it changes over time and space. To accomplish that goal, we first cleaned the data, removing unusually large or small fish, as well as Bull trout, whitefish, and fish which had no weight measurement. Because volume, and therefore weight has a multiplicative relationship to fish length, thickness, and height, and assuming each dimension of a fish's body grows in proportion to the other dimensions, we took logs to make the relationship additive. The linear model was fit to log(weight) using log(length) (centered at the average log length), species, section, and year (as a factor) plus two-way interactions between log(length) and the other predictors.*

*The slope for Browns from Johnsrud in 1989 is estimated as 2.75 (0.018), so a small fish with log(length) of 4.2 is estimated to have 6.44 less log(weight) than the largest fish (with log(length) = 6.5).*

*Adjustments to intercept:*
*RBT of mean log(length) are estimated to have slightly smaller log(weight) than Browns (of mean log(length) at Johnsrud in 1989), but only by 0.022( SE = 0.0055), so this seems to be a difference which is not of practical importance. Similarly, the estimate for ScottyBrown is only 0.034 (0.0044), which is not large compared to the range of log(weights) from 1.6 to 7.6. Year effects range from -0.99 (0.12) in 2004 to 0.107 (0.11) in 1998 (all are relative to 1989 Browns). These indicate to me that the relationship does vary across years.*

*Adjustments to slope:*
*The slope for RBT's differs by -0.105 (0.012) from that of Browns, which might be of interest. However, across the range of log(length)s, that changes log(weight) by only 0.246. The difference between slope at Johnsrud and at ScottyBrown is smaller, estimated as 0.058 (0.016), so I think not large enough to worry about. Similarly, year effects range from -0.003 (0.02) in 2000 to 0.18 (.02) in 2004.*

*Diagnostics:*
*The plots of log(weight) over log(length) show a curious fan shape – broader at the bottom left. I'm guessing that this is because a small error in measuring length is amplified when we convert to logs. That shows up in the first and third diagnostic plots as weird bands*

*of residual points on the left side of the plots. It appears that variance is decreasing with fitted values. I did look at the Box-Cox plot for the raw data, and found that log transform ($\lambda = 0$) is not optimal, being a bit too small, but is not far from optimal, so no other transformation seems to help more. The plot of ordered residuals against normal quantiles shows very long tails. There is no way to fix this, but we can appeal to the CLT because we have lots of data, and can conclude that our p-values, though not exact, are generally trustworthy.*

*Scope:*
*The analysis presented is simply a summary of the observations. It does not apply to a broader population of fish (no random sampling) and does not provide causal evidence (no treatments were applied).*

## R Code Appendix

```
propRBT <- prop.table(with(fish, table(species, section, year)), 2:3)[2,
    , ]
year <- as.numeric(colnames(propRBT))
plot(propRBT[1, ] ~ year, type = "l", ylim = range(propRBT))
lines(propRBT[2, ] ~ year, type = "l", lty = 2)
legend("bottomleft", lty = 1:2, c("Johnsrud", "ScottyBrown"))
```

```
fish <- droplevels(subset(blackft, !is.na(weight) & weight > 0 & (species ==
    "RBT" | species == "Brown")))
```

```
require(lattice)
densityplot(~length, fish, subset = length < 100)
fish <- subset(fish, length > 60)
```

```
require(ggplot2)
qplot(x = length, y = weight, data = fish, facets = species ~ year) + theme_bw()
```

```
fish$cond <- with(fish, weight^0.333/length * 50)
densityplot(~log(cond) | factor(year) * species, fish, subset = cond < 2 &
    cond > 0.5)
fish <- subset(fish, cond < 2 & cond > 0.5 & length < 750)
```

```
summary(tt.sd <- as.numeric(with(fish, tapply(cond, list(species, year),
    sd))))
par(mfrow = c(1, 2))
summary(tt.mn <- as.numeric(with(fish, tapply(cond, list(species, year),
    mean))))
plot(year, tt.mn[seq(1, 20, 2)], type = "l", lty = 4, main = "Mean Condition by Species/Year",
    ylab = "Mean Condition")
lines(year, tt.mn[seq(2, 20, 2)])
legend("bottomleft", c("Rainbow", "Brown"), lty = c(1, 4), bty = "n")
plot(year, tt.sd[seq(1, 20, 2)], type = "l", lty = 4, main = "Condition\n Spread by Species/Year",
    ylab = "Std. Deviation of Condition")
lines(year, tt.sd[seq(2, 20, 2)])
legend("bottomleft", c("Rainbow", "Brown"), lty = c(1, 4), bty = "n")
```

```r
qplot(x = log(length), y = log(weight), data = fish, facets = species ~ year) +
    theme_bw()
fish$Llen <- scale(log(fish$length), T, F)
fish.fit1 <- lm(log(weight) ~ Llen * (species + factor(year) + section),
    fish)
xtable(anova(fish.fit1))
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Llen | 1 | 17101.92 | 17101.92 | 344788.03 | 0.0000 |
| species | 1 | 4.32 | 4.32 | 87.16 | 0.0000 |
| factor(year) | 9 | 15.09 | 1.68 | 33.80 | 0.0000 |
| section | 1 | 3.03 | 3.03 | 61.09 | 0.0000 |
| Llen:species | 1 | 5.92 | 5.92 | 119.43 | 0.0000 |
| Llen:factor(year) | 9 | 6.18 | 0.69 | 13.85 | 0.0000 |
| Llen:section | 1 | 1.48 | 1.48 | 29.81 | 0.0000 |
| Residuals | 13996 | 694.22 | 0.05 | | |

```r
xtable(summary(fish.fit1))
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.8112 | 0.0081 | 591.23 | 0.0000 |
| Llen | 2.7571 | 0.0183 | 150.65 | 0.0000 |
| speciesRBT | -0.0216 | 0.0055 | -3.94 | 0.0001 |
| factor(year)1990 | 0.0970 | 0.0082 | 11.82 | 0.0000 |
| factor(year)1991 | 0.0583 | 0.0082 | 7.10 | 0.0000 |
| factor(year)1993 | 0.0373 | 0.0085 | 4.41 | 0.0000 |
| factor(year)1996 | 0.0943 | 0.0107 | 8.84 | 0.0000 |
| factor(year)1998 | 0.1034 | 0.0098 | 10.60 | 0.0000 |
| factor(year)2000 | 0.0869 | 0.0084 | 10.36 | 0.0000 |
| factor(year)2002 | 0.0261 | 0.0103 | 2.54 | 0.0112 |
| factor(year)2004 | 0.0133 | 0.0093 | 1.43 | 0.1539 |
| factor(year)2006 | 0.1007 | 0.0087 | 11.52 | 0.0000 |
| sectionScottyBrown | 0.0345 | 0.0044 | 7.83 | 0.0000 |
| Llen:speciesRBT | -0.1054 | 0.0124 | -8.50 | 0.0000 |
| Llen:factor(year)1990 | 0.0652 | 0.0204 | 3.20 | 0.0014 |
| Llen:factor(year)1991 | 0.0477 | 0.0208 | 2.30 | 0.0215 |
| Llen:factor(year)1993 | 0.0281 | 0.0195 | 1.44 | 0.1501 |
| Llen:factor(year)1996 | 0.0558 | 0.0237 | 2.35 | 0.0187 |
| Llen:factor(year)1998 | 0.0733 | 0.0218 | 3.36 | 0.0008 |
| Llen:factor(year)2000 | -0.0038 | 0.0210 | -0.18 | 0.8579 |
| Llen:factor(year)2002 | 0.1533 | 0.0243 | 6.31 | 0.0000 |
| Llen:factor(year)2004 | 0.1842 | 0.0220 | 8.39 | 0.0000 |
| Llen:factor(year)2006 | 0.0693 | 0.0216 | 3.21 | 0.0013 |
| Llen:sectionScottyBrown | 0.0580 | 0.0106 | 5.46 | 0.0000 |

```r
par(mfrow = c(1, 5))
plot(fish.fit1)
MASS::boxcox(lm(weight ~ Llen * (species + factor(year) + section), fish),
    lambda = seq(-0.5, 0.5, 0.1))
```