Stat 505 Assignment 4

Name here?

Excellent. 17.5 of 18

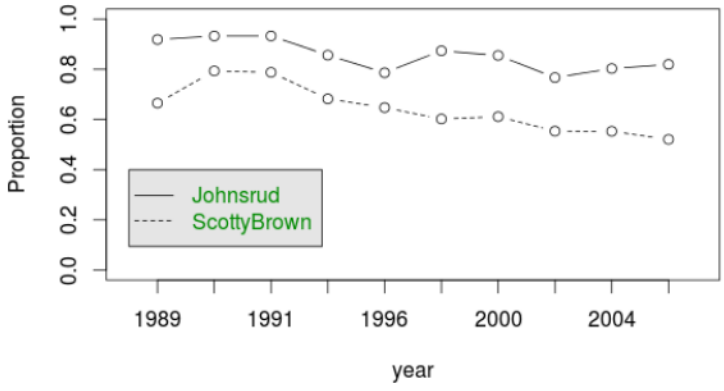
Due: September 26

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.

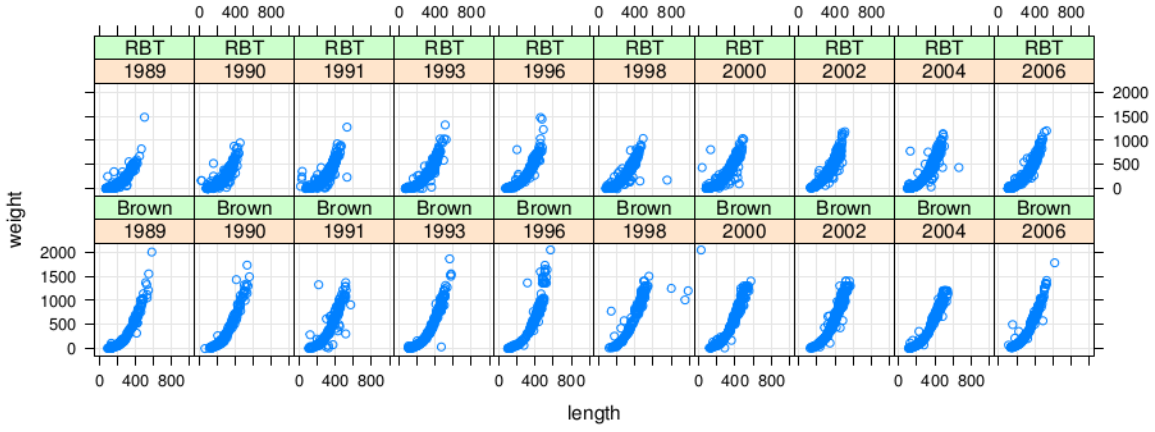
2. Use a plot to show how the proportion of RBT changes with year and section. Does one section always have a higher proportion RBT than the other?



Is this a decline across years?
-.5

Section Johnsrud has a higher proportion of RBT across all years.

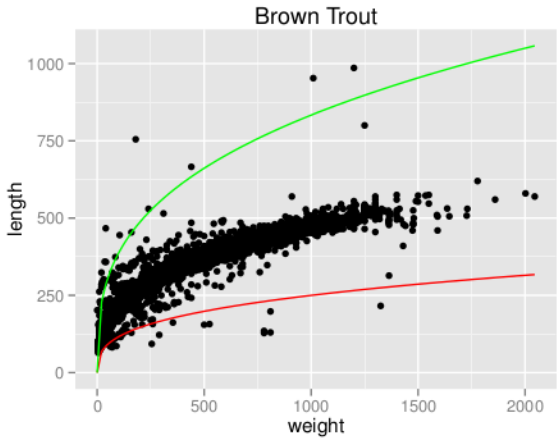
3. Plot weight as a function of length, separating into panels by species and year.

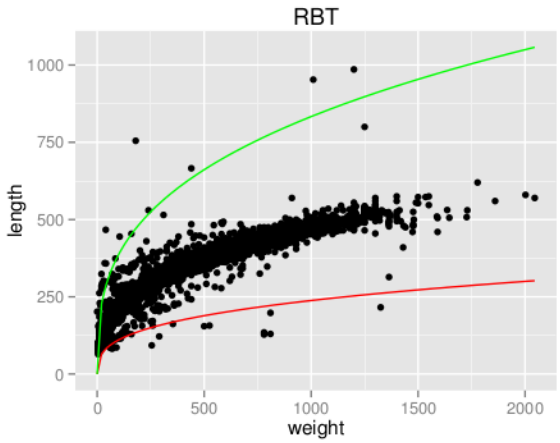


What seems to be the lower limit of length that is catchable with this equipment? Filter out fish you deem too short to have actually been caught.

The lower limit of fish that is catchable with this equipment seems to be about 64 *mm which is about* 2.5 *inches. After the above adjustments are made to the dataset, there are only* 5 *fish smaller than* 64 *mm, and all five are less than* 36 *mm.* 36 *mm is about a* 1.4 *inch fish, and no matter how fine meshed of a net the researchers are using, I think it would be very difficult to prevent that size of a fish from swimming through the holes in the net.*

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.)





So these lines show the boundaries you filter by, below?
Conditions of .6 and 2.1?

Log scale would give a better view of the bottom left.

Based on the plots above, for brown trout I will delete all fish with condition index less than 0.6 *or greater than* 2. *For RBT, I will delete all fish with condition index less than* 0.6 *and greater than* 2.1.

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.

If fish bodies are of constant density, the weight of the fish is a function of the length cubed. Because we know that there is a multiplicative relationship between length and weight, it makes sense to log transform weight so that our interpretation will be in terms of multiplicative changes. This also assumes thickness & height increase prop to length, right?

6. Fit a linear model for log(weight) on log(length). Does the intercept depend on site and or species? Does the slope? Fit a model with main effects and appropriate interactions (let's leave out the 3-way interaction). Interpret each coefficient estimate. Explain exactly what effect each is measuring.

   The answers to these questions are in my discussion below.

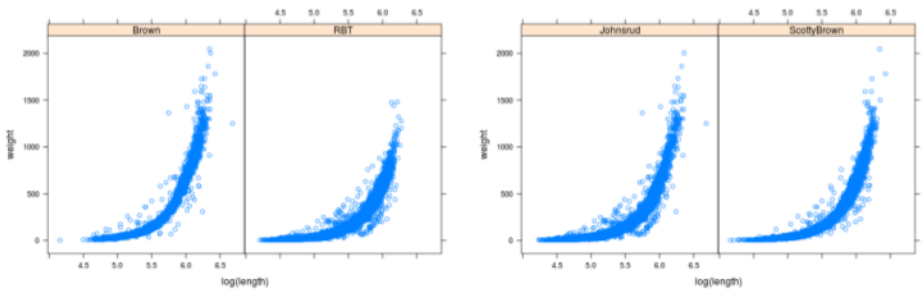
7. Write up your model including a discussion of the diagnostic plots.

   Introduction

   Information was collected about the lengths, weights, and species of fish in the Blackfoot river from 1989 *to* 2006. *The section and year in which they were caught was also recorded.*

   Statistical Prodedures

   I first looked at some plots to decide which interaction terms to include in the model. The log(length) effect appears to depend on species. This we would expect biologically. Due to differences in habitat, I also expected that the relationship between log(length) and log(weight) would depend on section. Even though the scatterplot does not suggest a log(length) by section interaction, I still included it in the model because of the biological argument. I didn't think any major changes could happen in the time span of this study, so I chose not to include a log(length) by year interaction. The paneled plot above supports this decision, showing that, within a species, the relationship between log(length) and log(weight) doesn't change much across years. Note that I considered treating year as a continuous variable, but in the end I chose to treat it as a factor because I didn't see a linear trend in log(weight) across year (and our sample size is so large I wasn't worried about the loss of df).

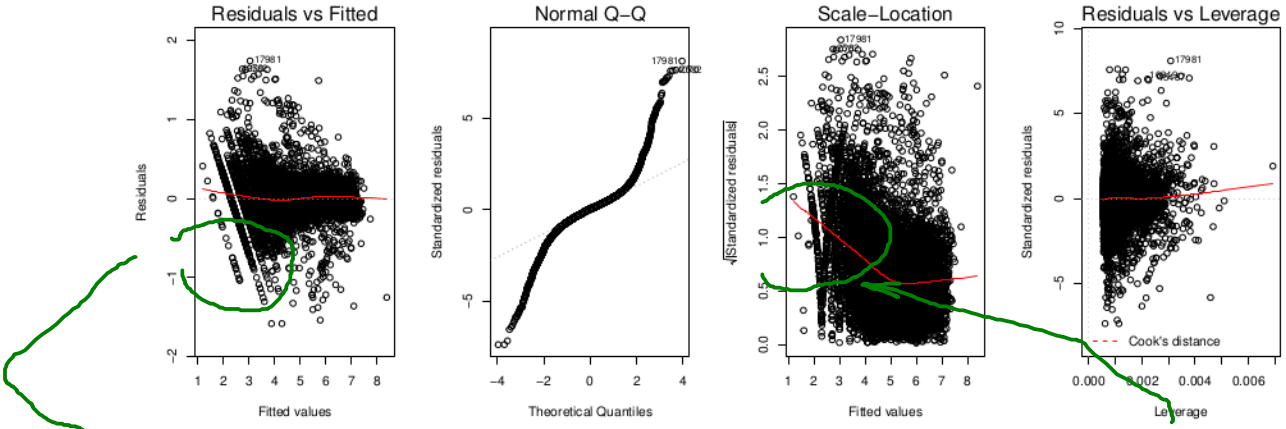


I think temporal changes are possible w/in this time span. Whirling disease came in around 2000, but has not had a big impact (yet?). That might change the RBT proportion.

I fit a model of log(weight) on log(length), species, year, and section as well as the above mentioned two factor interactions. The diagnostic plots are below. There are some concerns. It seems that the spread of the residuals decreases with increasing fitted

I didn't specify an audience for the report, but would avoid abbrev's

values. As a result, we see a downward trend in the scale-location plot. Additionally, the normal Q-Q plot reveals a long-tailed distribution. But with this large of a sample size the central limit theorem kicks in and our sampling distribution of sample means is normally distributed even when the normality and HOV assumptions are not met. The assumption of independence, however, is still important and could cause concern if researchers have reason to be believe it is not met.



These look like length measurement errors and may cause the trend here

Summary of Findings

There is strong evidence that log(weight) depends on log(length), species, year, and section (p-values< 0.0001). There is also strong evidence that the relationship between log(weight) and log(length) depends on species and section (p-values< 0.0001). For every doubling of length, the mean weight of brown trout in section Johnsrud is estimated to increase by a factor of 7.1 with a 95% confidence interval from 6.96 to 7.18. Holding all other variables constant, the intercept is estimated to increase by 0.589 when going from brown trout to rainbow trout, with a 95% confidence interval from 0.435 to 0.696. The intercept is estimated to decrease by 0.255 when going from section Johnsrud to section Scotty Brown, with a 95% confidence interval from a 0.491 to a 0.257 unit decrease. The intercept is estimated to change by the amounts shown in the below table for each year. The slope is estimated to decrease by 0.1117 when going from brown trout to rainbow trout in section Johnsrud, with a 95% confidence interval from a 0.13 to a 0.09 decrease. The slope for rainbow trout is estimated to increase by 0.0527 when going from section Johnsrud to section ScottyBrown with a 95% confidence interval from 0.0454 to 0.0861.

log scale?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
log(length)	1	17122.61	17122.61	368792.87	0.0000
species	1	4.00	4.00	86.05	0.0000
year	9	15.30	1.70	36.62	0.0000
section	1	2.77	2.77	59.69	0.0000
log(length):species	1	5.26	5.26	113.21	0.0000
log(length):section	1	1.24	1.24	26.71	0.0000
Residuals	13991	649.59	0.05		

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-10.6112	0.0651	-163.08	0.0000
log(length)	2.8282	0.0115	245.92	0.0000
speciesRBT	0.5890	0.0666	8.84	0.0000
year1990	0.0858	0.0073	11.68	0.0000
year1991	0.0483	0.0075	6.44	0.0000
year1993	0.0219	0.0078	2.83	0.0047
year1996	0.0796	0.0099	8.05	0.0000
year1998	0.0873	0.0090	9.74	0.0000
year2000	0.0743	0.0077	9.68	0.0000
year2002	0.0358	0.0084	4.26	0.0000
year2004	0.0046	0.0086	0.53	0.5948
year2006	0.0862	0.0080	10.80	0.0000
sectionScottyBrown	-0.2555	0.0567	-4.50	0.0000
log(length):speciesRBT	-0.1117	0.0118	-9.46	0.0000
log(length):sectionScottyBrown	0.0527	0.0102	5.17	0.0000

Scope of Inference

There is no indication that fish were randomly selected from the fish in the Blackfoot river, so inference does not extend beyond the fish sampled. In fact, it is probably safe to assume that this was a convenience sample because those fish caught were probably the ones that were easiest to target. They aim to get a census, but electrofishing has some biases.

Fish were not randomly assigned lengths, so statistically we cannot establish a causal relationship between length and weight, but we can acknowledge the physical law that increasing length must be accompanied by increasing weight.

R Code Appendix

```
lst <- tapply(blackft$species, blackft[,c(5,7,6)], length)
lst1 <- as.data.frame(lst)
prop.john <- lst1[,2]/(lst1[,1]+lst1[,2])
prop.scotty <- lst1[,4]/(lst1[,3]+lst1[,4])
plot(c(rep(-1,10))~unique(blackft$year), ylim=c(0,1), ylab="Proportion", xlab="year")
points(prop.john~unique(blackft$year), type="b")
points(prop.scotty~unique(blackft$year), type="b", lty=2)
legend(.5, .4, c("Johnsrud", "ScottyBrown"), lty = c(1,2),
       text.col = "green4",
       merge = TRUE, bg = "gray90")
```

```
require(lattice)
blackft$year <- factor(blackft$year)
#with(blackft, xyplot(weight~length|year, grid=TRUE, group=species, auto.key=TRUE))
with(blackft, xyplot(weight~length|year+species, grid=TRUE, auto.key=TRUE))
```

```
blackft <- subset(blackft, blackft$length>36)
```

```
blackft$condind <- 50*blackft$weight^(1/3)/blackft$length
require(ggplot2)
condindhigh <- function(x)(50*x^(1/3)/2)
condindlow <-  function(x)(50*x^(1/3)/.6)
qplot(x=weight, y=length, data=subset(blackft,species="Brown"), main="Brown Trout")+stat_function(fun = condindhigh, colour = "r
condind1 <- function(x)(50*x^(1/3)/2.1)
condind2 <-  function(x)(50*x^(1/3)/.6)
qplot(x=weight, y=length, data=subset(blackft,species="RBT"), main="RBT")+stat_function(fun = condind1, colour = "red")+stat_fun
```

```
blackft1 <- subset(blackft, species=="RBT" & condind<2 & condind>0.6)
blackft2 <- subset(blackft, species=="Brown" & condind<2.1 & condind>0.6)
blackft <- rbind(blackft1,blackft2)
```

```
#yr <- as.numeric(blackft$year)
lm.add <- lm(log(weight)~log(length)+species+year, data=blackft)
```

```
require(lattice)
xyplot(weight~log(length)|species, data=blackft)
xyplot(weight~log(length)|section, data=blackft)
```

```
yr <- as.numeric(blackft$year)
lm.fit <- lm(log(weight)~log(length)+species+year+section+log(length)*species+log(length)*section, data=blackft)
```

```
par(mfrow=c(1,4))
plot(lm.fit)
```

```
require(xtable)
xtable(anova(lm.fit))
xtable(summary(lm.fit))
```