Regression and ANOVA

Overview

The goal of this lab is to give you practice using R to create and evaluate basic regression models. This includes both classical linear regression and analysis of variance. Even though there are certainly differences between these two scenarios, seeing them as aspects of a common technique can make life easier.

We'll be using `ggplot2` to make graphics and `Anova` from the `car` package in this lab so load those libraries.

```
library(ggplot2)
library(car)
```

The data: Growth of redside shiners

The data that we will be looking at today are from Houston and Belk (2006). The goal of the experiment was to determine whether observed differences in fish growth resulted from environmental or genetic variation.

The data are on GitHub or in my Pick Up folder in the file *redside_shiner.csv*. Save this file to the data folder in your working directory for this lab and load it as the data frame `shiners` with the command

```
shiners <- read.csv("data/redside_shiner.csv")
```

Run `str(shiners)` or click on the blue circle with the white triangle icon next to the data in the *Environment* pane to get a list of the variables and their types. Observations in this data set are individual fish. The variables are:

- `obs` is the observation number
- `loc` is location
- `block` is a part of the experimental design
- `temp` is temperature where the fish were grown in deg C, an experimental treatment
- `food` is frequency of feeding, another experimental treatment
- `smass` is the starting mass
- `emass` is the mass at the end of the experiment
- `ssl` is the starting standard length
- `esl` is the ending standard length
- `days` is the length of time that the fish was in the experiment

Use `summary(shiners)` and `View(shiners)` to get an idea of the data. There are a fair number of observations with missing data, but the data set is still sufficiently large if we drop the observations with missing entries, and it makes the analysis much easier.

```
shiners <- na.omit(shiners)
```

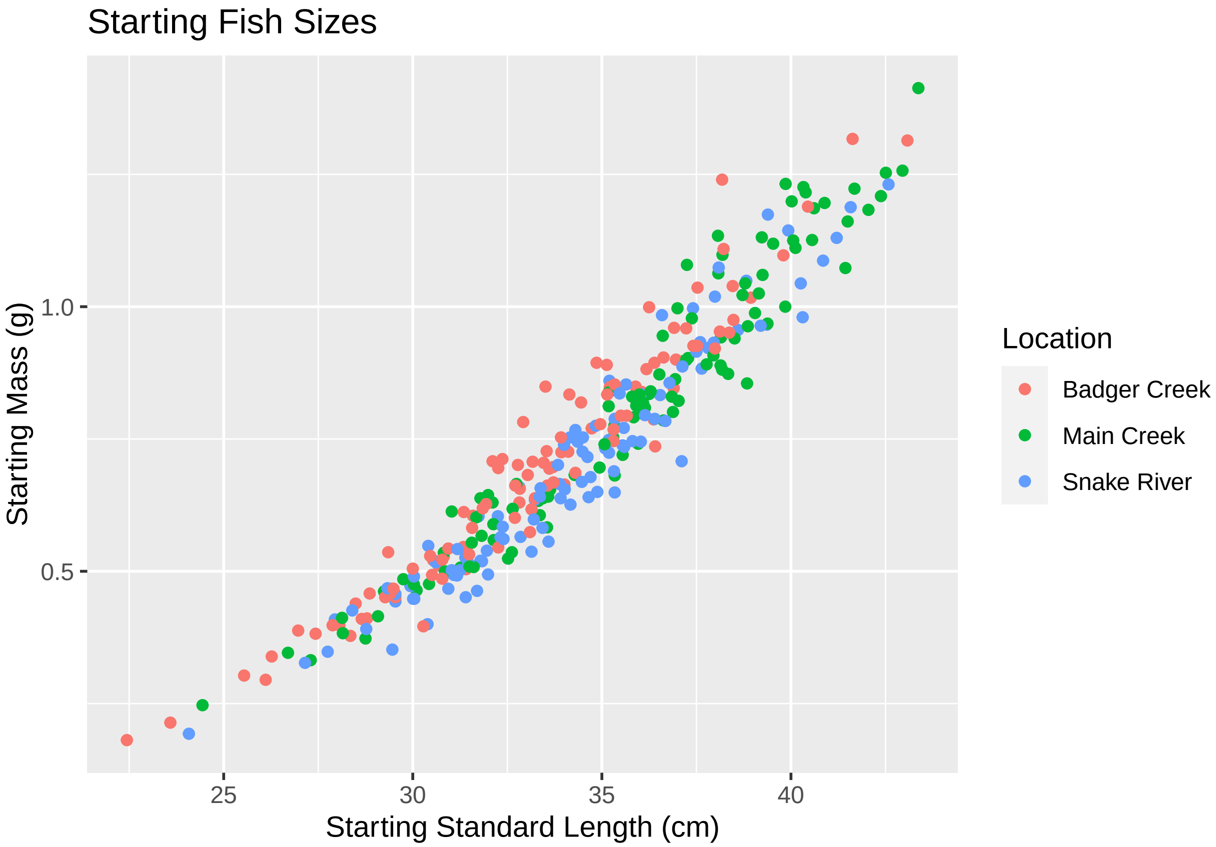
Since `temp` only takes three values, it makes sense to convert it to a factor.

```
shiners$temp <- factor(shiners$temp)
```

Starting populations

Let’s examine the fish as they were at the beginning of the experiment. Make a plot showing `smass`, `ssl` and `loc`.

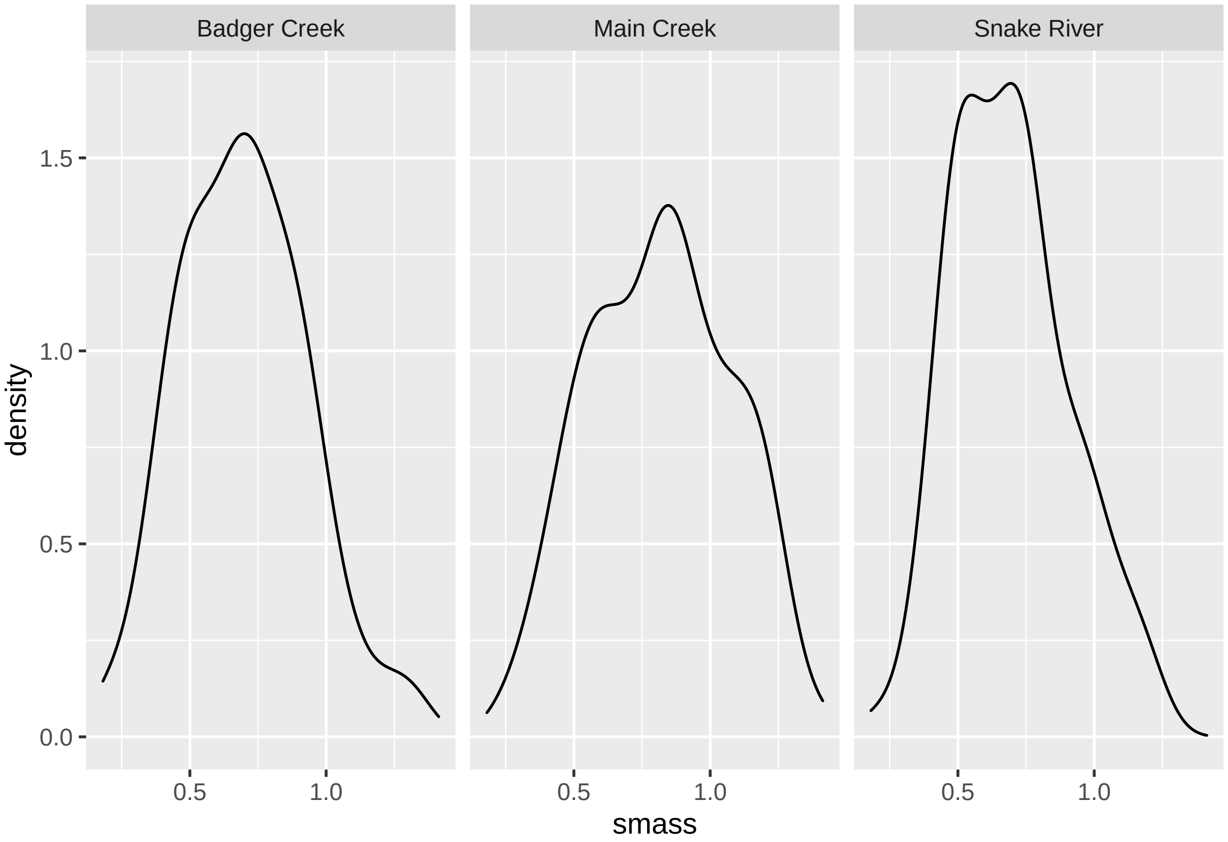
```
ggplot(shiners, aes(ssl, smass, color = loc))+
  geom_point()+
  labs(x = "Starting Standard Length (cm)", y = "Starting Mass (g)",
       title = "Starting Fish Sizes", color = "Location")
```



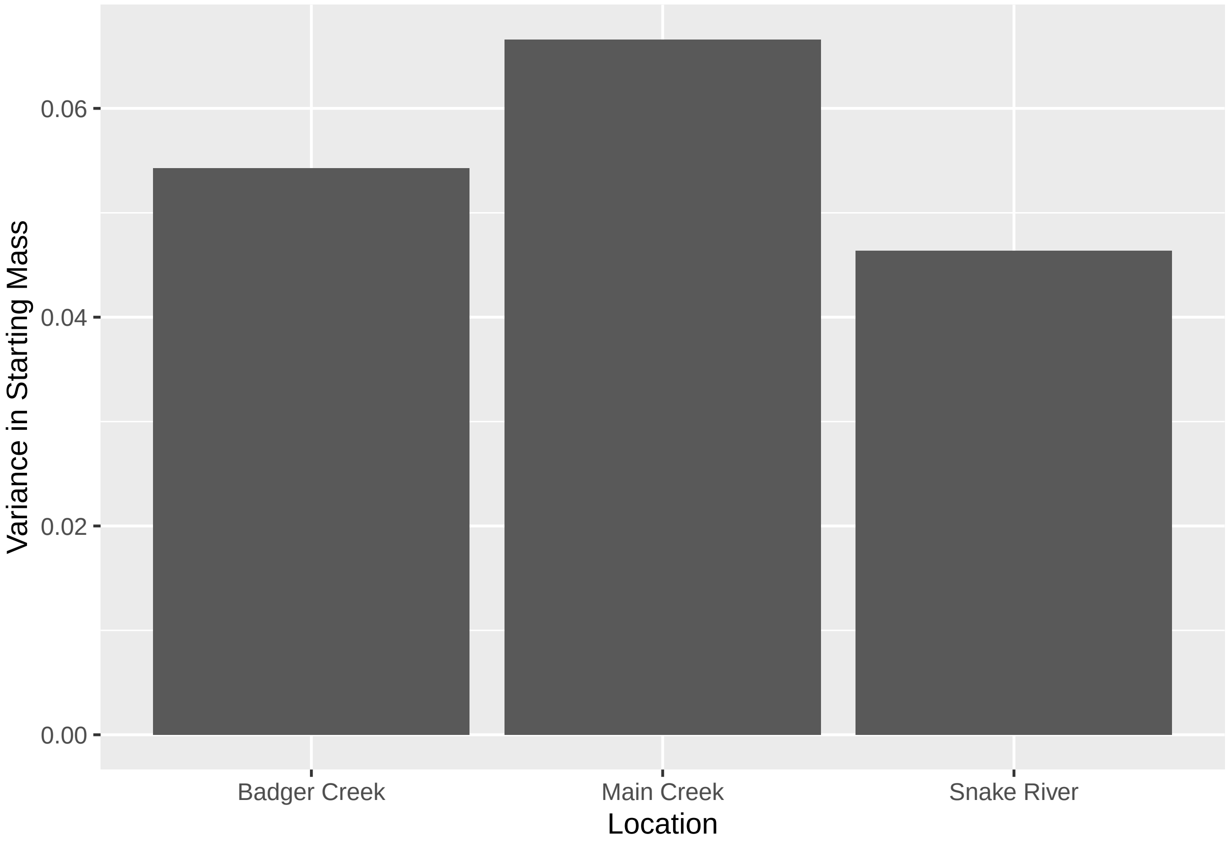
(1) Does it appear that the fish from any one location are larger or smaller than the others either in mass, length or both?

We can do a one-way anova to answer this question for mass. Start by examining the normality and homogeneity of variance in `smass` for the `loc` groups to confirm that anova is legitimate.

```
ggplot(shiners, aes(smass))+
  geom_density()+
  facet_wrap(~loc)
```



```
ggplot(shiners, aes(loc, smass))+
  stat_summary(fun=var, geom = "bar")+
  labs(x="Location",y="Variance in Starting Mass")
```



```
tapply(shiners$smass, shiners$loc, var)
```

```
## Badger Creek   Main Creek  Snake River 
##   0.05426096   0.06659484   0.04638570
```

We perform the regression with the `lm` function.

```
lm1 <- lm(smass~loc, data = shiners)
summary(lm1)
```

```
## 
## Call:
## lm(formula = smass ~ loc, data = shiners)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.57014 -0.17914 -0.00155  0.15749  0.62045 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      0.696552   0.023044  30.227  < 2e-16 ***
## locMain Creek    0.120588   0.032436   3.718 0.000238 ***
## locSnake River -0.002076   0.032436  -0.064 0.949015    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2361 on 316 degrees of freedom
## Multiple R-squared:  0.05636,    Adjusted R-squared:  0.05039 
## F-statistic: 9.437 on 2 and 316 DF,  p-value: 0.0001046
```

The last line of the output gives us the p value for the F test ($p < 0.001$), and we can conclude that there is some difference in starting mass. The coefficients of the model tell us that while the Beaver Creek and Snake River fish do not appear different, even after correcting the p value for multiple comparisons (there are $\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3$ comparisons, so multiplying the coefficient of `locMainCreek`'s p value of 0.000238 by 3 gives the Bonferroni correction, which tends to be conservative) we have $p < 0.001$ for the comparison of Main Creek and Beaver Creek. Mean starting mass for the Main Creek fish appears to be about 0.1g greater than the mean masses for the fish from the other locations. With only three levels to the single factor in this model it is possible to see the individual comparisons in the coefficients table, but in general after a significant F test we will want to perform post-hoc tests with R.

```
pairwise.t.test(shiners$smass, shiners$loc, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  shiners$smass and shiners$loc
##
##             Badger Creek Main Creek
## Main Creek  0.00071      -
## Snake River 1.00000      0.00052
##
## P value adjustment method: bonferroni
```

(2) Are the Main Creek fish also longer on average at the start of the experiment?

An aside on dimensions

Length and mass are both measuring size. Since mass is proportional to volume, and volume is measured in cubic cm while length is measured in cm, it makes sense to expect that the relationship between mass and length is roughly cubic.

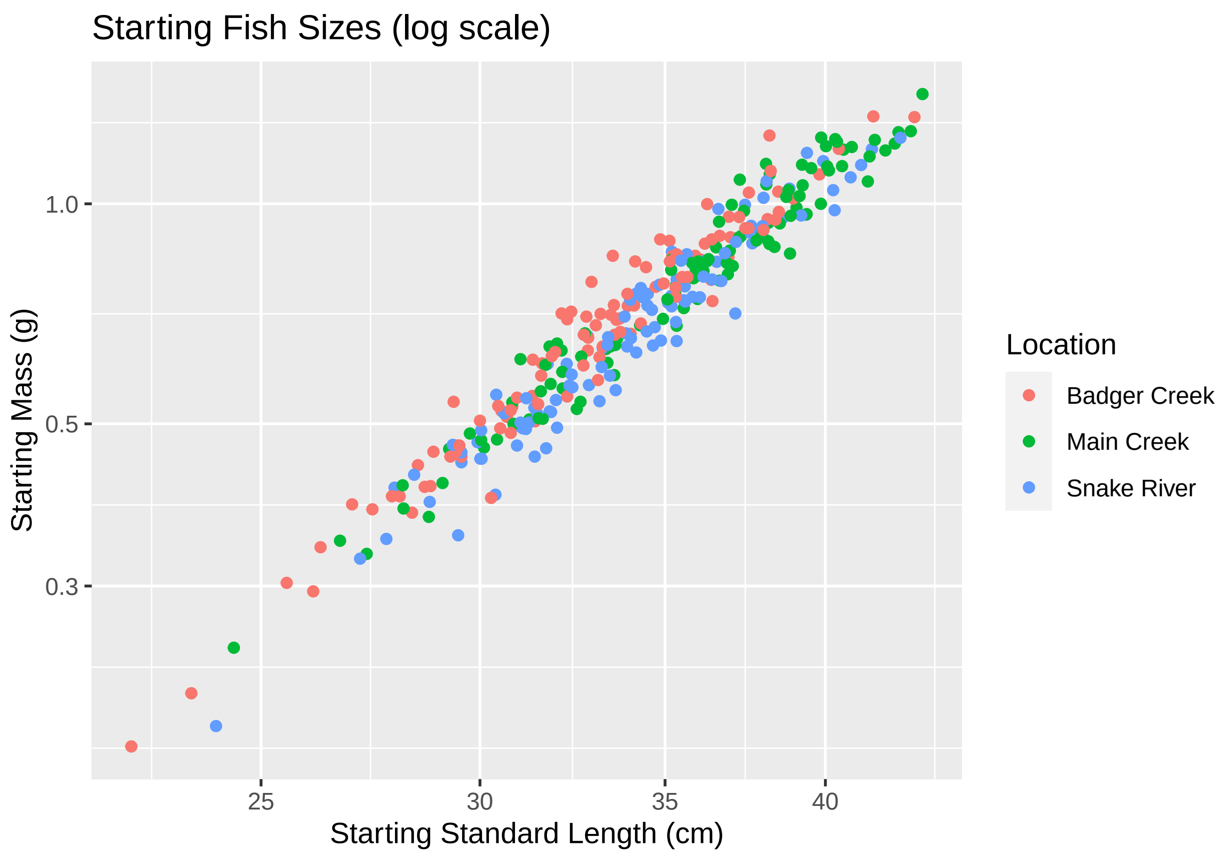
$$\text{mass} = k \cdot \text{length}^3$$

where k encompasses density and aspect ratio. Significant deviations from this exponent of 3 could potentially have interpretations in terms of the growth processes of the fish but we won't speculate on that here. Taking the log of this equation gives

$$\log(\text{mass}) = \log(k) + 3\log(\text{length})$$

Plotting this relationship can either be done by transforming the variables, or using a log scale. Let's examine the fish as they were at the beginning of the experiment. Make a plot showing `smass`, `ssl` and `loc`.

```
ggplot(shiners, aes(ssl, smass, color = loc))+
  geom_point()+
  labs(x = "Starting Standard Length (cm)", y = "Starting Mass (g)",
       title = "Starting Fish Sizes (log scale)", color="Location")+
  scale_x_log10() + scale_y_log10()
```



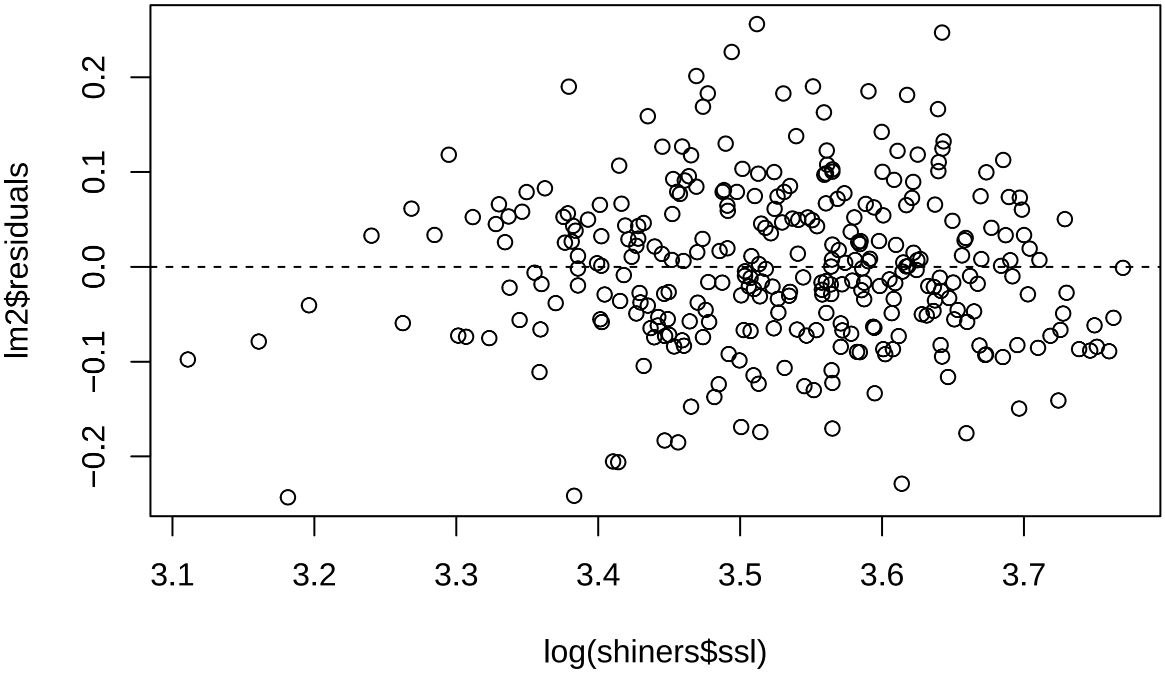
This does look like a more linear pattern than the plot made on the direct scales above. We can use a regression to confirm this relationship.

```
lm2 <- lm(log(smass)~log(ssl), data = shiners)
summary(lm2)
```

```
##
## Call:
## lm(formula = log(smass) ~ log(ssl), data = shiners)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.243144 -0.058111 -0.001724  0.053747  0.256110
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.85681    0.14432  -75.23   <2e-16 ***
## log(ssl)      2.97194    0.04086   72.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08451 on 317 degrees of freedom
## Multiple R-squared:  0.9435, Adjusted R-squared:  0.9433
## F-statistic:  5291 on 1 and 317 DF,  p-value: < 2.2e-16
```

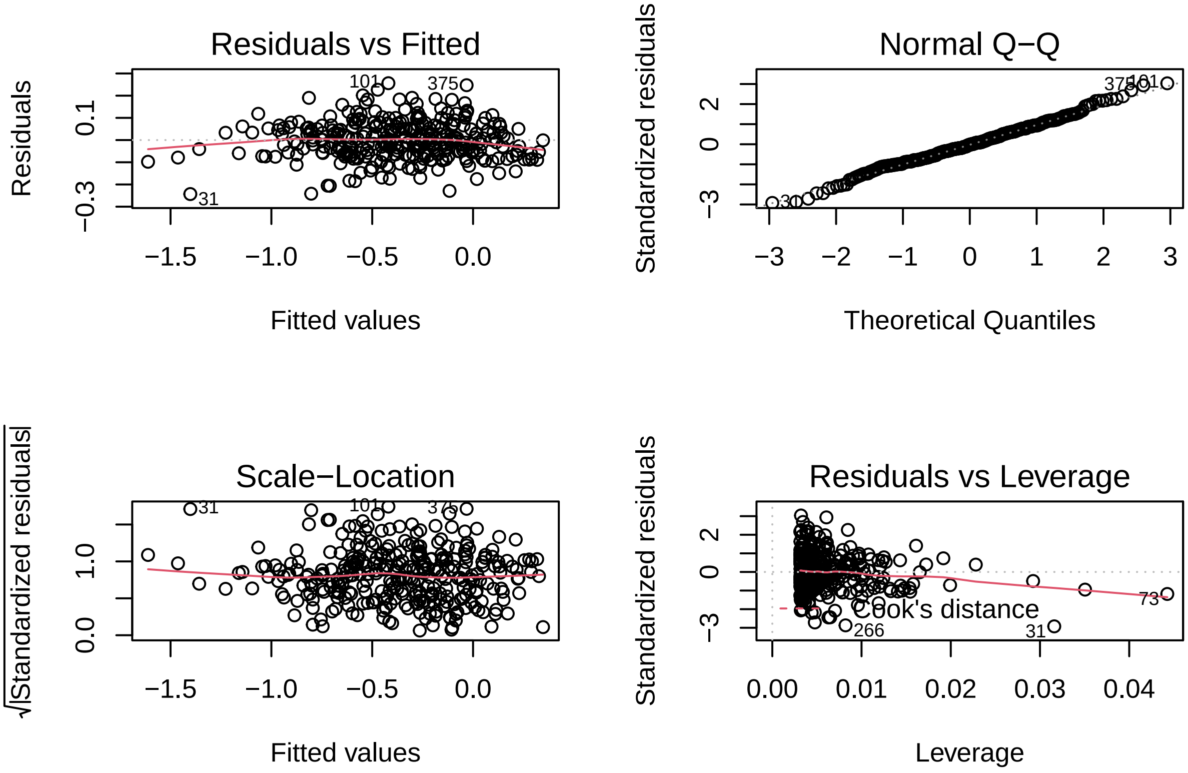
To evaluate this regression, we can plot the residuals against the explanatory variable.

```
plot(lm2$residuals~log(shiners$ssl))
abline(0,0, lty="dashed")
```



A linear model can also be evaluated by the four diagnostic plots produced by R. When you run the following line look at the console for a prompt to scroll through the four plots that follow.

```
par(mfrow=c(2,2)) # to show all four plots at once
plot(lm2)
```



```
par(mfrow=c(1,1)) # to see one plot at a time again
```

We examine the first and third of these diagnostic plots for evidence of heteroscedatsicity. The second helps us asses normality of the residuals, and the fourth helps us asses the influence of any detected deviations from these assumptions. There do not appear to be any significant issues with the assumptions underlying this regression.

We might wonder if location has any connection on this relationship. It can be introduced to the regression additively, so that it can only change the intercept, which is log(k):

```
lm2a <- lm(log(smass)~log(ssl)+loc, data = shiners)
anova(lm2a)
```

```
## Analysis of Variance Table
##
## Response: log(smass)
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## log(ssl)    1 37.793  37.793 6191.433 < 2.2e-16 ***
## loc         2  0.341   0.171   27.963 6.624e-12 ***
## Residuals 315  1.923   0.006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm2a)
```

```
##
## Call:
## lm(formula = log(smass) ~ log(ssl) + loc, data = shiners)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.245071 -0.049466 -0.002939  0.051087  0.212074
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -10.96088    0.13688 -80.079  < 2e-16 ***
## log(ssl)          3.01412    0.03909  77.109  < 2e-16 ***
## locMain Creek    -0.05409    0.01110  -4.872 1.75e-06 ***
## locSnake River   -0.07952    0.01079  -7.368 1.53e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07813 on 315 degrees of freedom
## Multiple R-squared:  0.952,  Adjusted R-squared:  0.9515
## F-statistic:  2082 on 3 and 315 DF,  p-value: < 2.2e-16
```

Or it can be an interaction term, so that it can potentially change the slope, which is in this case the exponent of approximately 3.

```
lm2b <- lm(log(smass)~log(ssl)*loc, data = shiners)
anova(lm2b)
```

```
## Analysis of Variance Table
##
## Response: log(smass)
##               Df Sum Sq Mean Sq   F value    Pr(>F)
## log(ssl)       1 37.793  37.793 6201.1396 < 2.2e-16 ***
## loc            2  0.341   0.171   28.0070 6.464e-12 ***
## log(ssl):loc   2  0.015   0.008    1.2469    0.2888
## Residuals    313  1.908   0.006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The same checks for normality and homogeneity of variance in `smass` across locations above are also relevant to this regression.

It appears that the differences in mass due to location that we noted earlier are also reflected in this regression, though only the additive effect is only significant at the $\alpha = 0.05$ level, not the interaction effect.

(3) Does this result change after the experiment? Perform the same regressions for location's effect on the relationship between log(`emass`) in terms of log(`esl`). Report on your results, including the diagnostic checks.

The Experiment

The goal is to evaluate growth, so it makes sense to create two new variables measuring the differences in mass and length at the beginning and end of the experiment. To create the mass difference variable:

```
shiners$mass.diff <- shiners$emass - shiners$smass
```

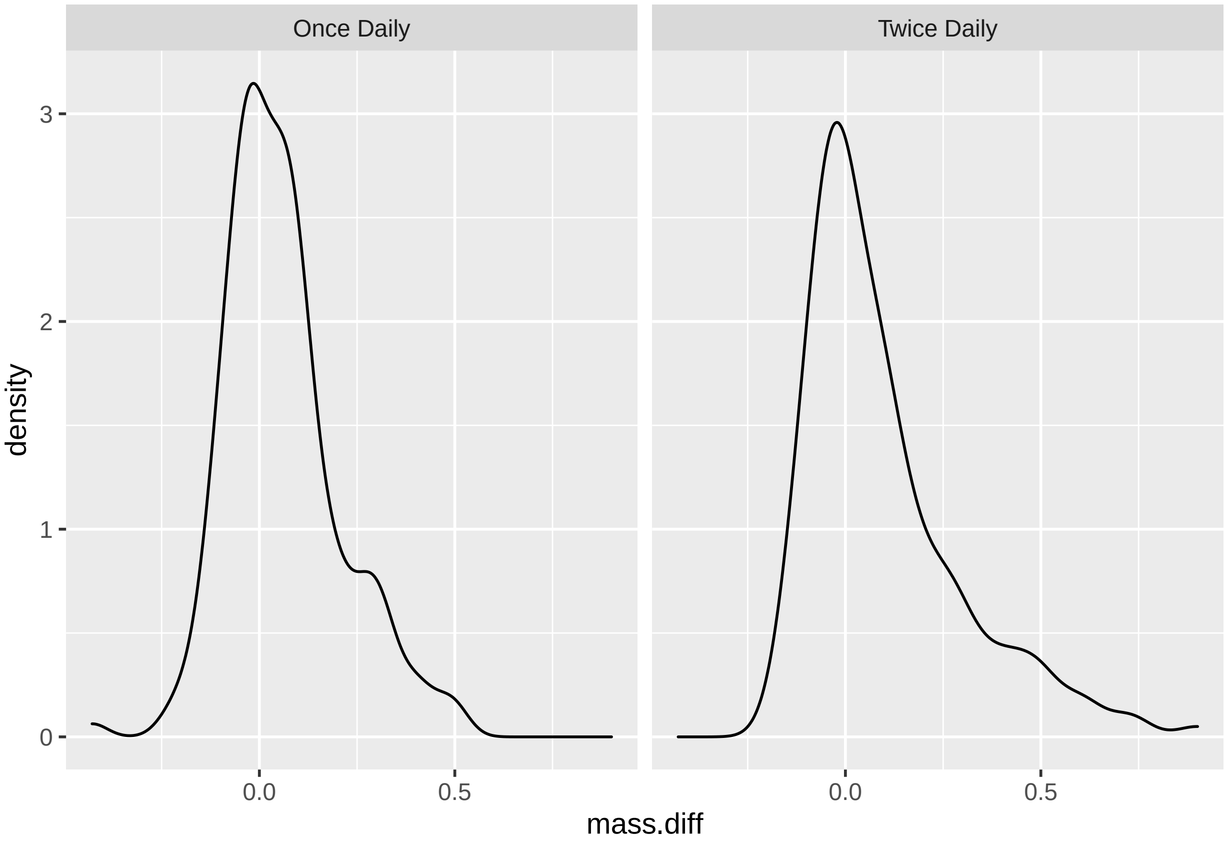
(4) Also create a length difference variable. Call it `sl.diff`.

From t tests to regression

Let's begin by asking if the mass difference is the same in the two feeding groups. We'd expect that the fish who were fed twice daily grew more, but why not test this?

First, check the distributions for these two groups. There appears to be some right skew but it isn't severe

```
ggplot(shiners, aes(mass.diff))+
  geom_density()+
  facet_wrap(~food)
```



To do a t test, we can use the `t.test` function:

```
t.test(mass.diff~food, data = shiners)
```

```
##
##  Welch Two Sample t-test
##
## data:  mass.diff by food
## t = -2.0781, df = 286.63, p-value = 0.03859
## alternative hypothesis: true difference in means between group Once Daily and group Twice Daily is n
## 95 percent confidence interval:
##  -0.080138926 -0.002175299
## sample estimates:
##  mean in group Once Daily mean in group Twice Daily
##                0.05833333                0.09949045
```

Alternatively, we can try to answer the same question with regression:

```
summary(lm(mass.diff~food, data = shiners))
```

```
##
## Call:
## lm(formula = mass.diff ~ food, data = shiners)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48633 -0.11883 -0.04233  0.07009  0.80151
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       0.05833    0.01383   4.218 3.22e-05 ***
## foodTwice Daily   0.04116    0.01971   2.088   0.0376 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.176 on 317 degrees of freedom
## Multiple R-squared:  0.01356,    Adjusted R-squared:  0.01045 
## F-statistic: 4.359 on 1 and 317 DF,  p-value: 0.03762
```

Note that the p value is almost exactly the same. In fact, if you run `t.test` with the option `var.equal=TRUE`, you will get exactly the p value from the regression. The advantage of the t test is its ability to deal with unequal variances.

(5) Do a t test or regression to determine if `food` has an effect on `sl.diff`. Confirm the suitability of the test and report the results.

Analysis

The results of the experiment stated in Houston and Belk (2006) are:

Individuals grew faster at higher temperatures and with more food, and there was a significant interaction between location and temperature. There was no difference in growth rates among the three populations at 10C and 17C. However, at 24C, individuals from the Snake River population grew significantly slower than those from Badger Creek and Main Creek.

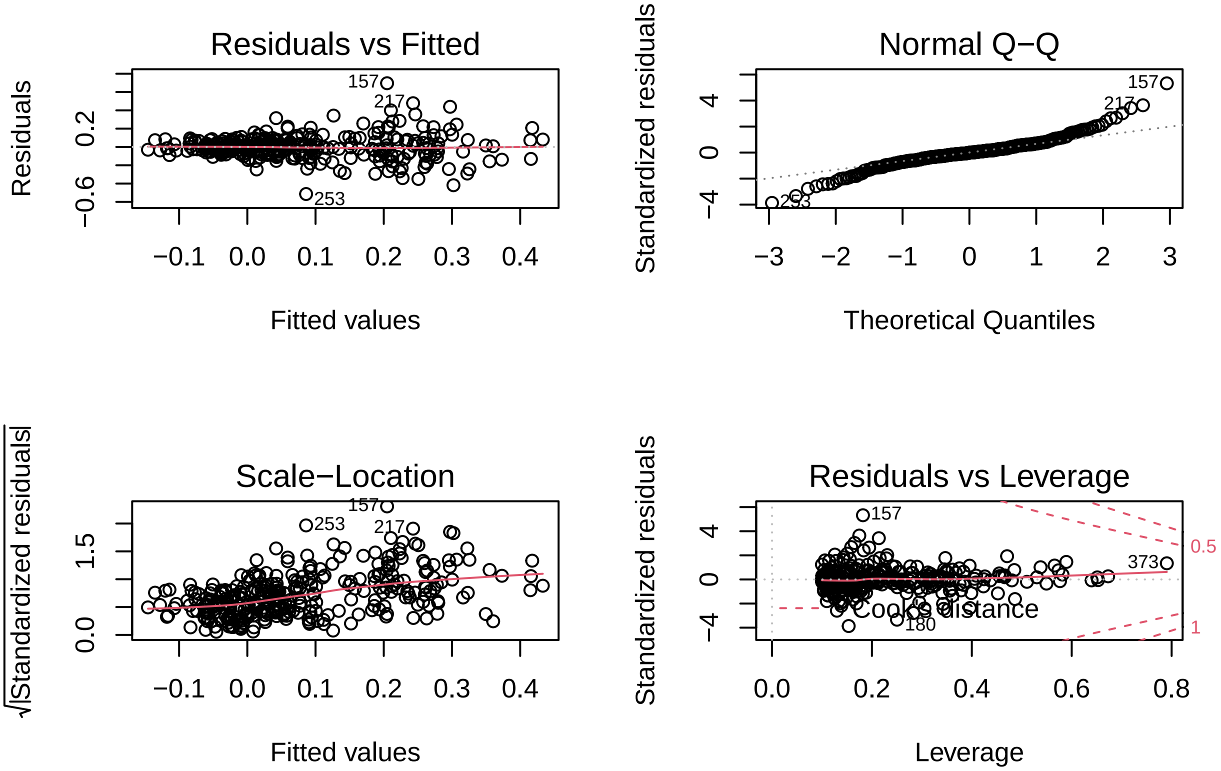
Let's come to this same conclusion. We'll fit a full interaction model. That's a lot of terms, but only a few of them are statistically significant. In many cases, especially observational studies, there is not enough data to test all of the interaction terms. When we have more terms in the model, we lose power. To test interactions only up to a specific level, you can use the symbol `^`. For example, to only test up to the two term interactions, replace `smass*block*loc*temp*food` with `(smass+block+loc+temp+food)^2` in the formula for `lm.mass`.

```
lm.mass <- lm(mass.diff ~ smass*block*loc*temp*food,
             data=shiners)
Anova(lm.mass)
```

```
## Anova Table (Type II tests)
## 
## Response: mass.diff
##                           Sum Sq  Df F value    Pr(>F)    
## smass                     0.0126   1  0.6043  0.437698    
## block                     0.0769   1  3.6798  0.056227 .  
## loc                       0.0081   2  0.1934  0.824257    
## temp                      3.2115   2 76.8592 < 2.2e-16 ***
## food                      0.1411   1  6.7518  0.009928 ** 
## smass:block               0.0066   1  0.3167  0.574101    
## smass:loc                 0.0295   2  0.7064  0.494418    
## block:loc                 0.0171   2  0.4087  0.664931    
## smass:temp                0.0441   2  1.0552  0.349689    
## block:temp                0.0395   2  0.9464  0.389552    
## loc:temp                  0.2021   4  2.4183  0.049203 *  
## smass:food                0.0042   1  0.1999  0.655176    
## block:food                0.0562   1  2.6910  0.102189    
## loc:food                  0.0439   2  1.0496  0.351636    
## temp:food                 0.0813   2  1.9462  0.144996    
## smass:block:loc           0.0198   2  0.4746  0.622693    
## smass:block:temp          0.0220   2  0.5269  0.591075    
## smass:loc:temp            0.0120   4  0.1437  0.965665
## block:loc:temp            0.0202   4  0.2414  0.914677
## smass:block:food          0.0433   1  2.0740  0.151097
## smass:loc:food            0.1063   2  2.5437  0.080631 .
## block:loc:food            0.0036   2  0.0863  0.917322
## smass:temp:food           0.0070   2  0.1680  0.845456
## block:temp:food           0.0572   2  1.3687  0.256355
## loc:temp:food             0.0508   4  0.6073  0.657729
## smass:block:loc:temp      0.1129   4  1.3514  0.251494
## smass:block:loc:food      0.0499   2  1.1938  0.304814
## smass:block:temp:food     0.0766   2  1.8325  0.162178
## smass:loc:temp:food       0.1359   4  1.6267  0.168072
## block:loc:temp:food       0.0174   4  0.2078  0.933929
## smass:block:loc:temp:food 0.0162   4  0.1943  0.941229
## Residuals                 5.1604 247
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Before we continue, we should make some diagnostic plots. We can start with the standard diagnostics from plotting `lm.mass`.

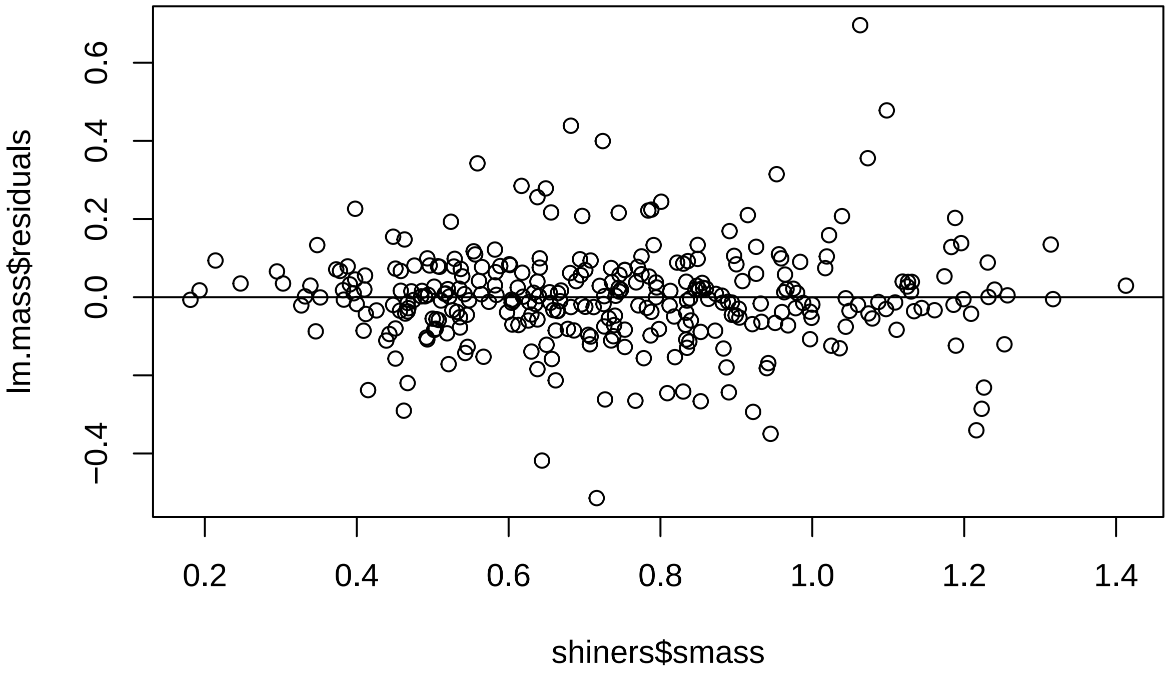
```
par(mfrow=c(2,2))
plot(lm.mass)
```



```
par(mfrow=c(1,1))
```

There do appear to be some deviations from both heteroscedasticity (visible in the first plot) and normality (visible in the second plot). Let's also look at the residuals and the continuous variable, `smass`.

```
plot(lm.mass$residuals~shiners$smass)
abline(h=0)
```



There doesn't seem to be an issue with this variable alone. To be thorough, we should look at all of the others. Instead we'll revisit these data when we have some more tools at our disposal, but for now let's continue with the present analysis.

In order to probe the effects of the significant factors, we can build a simpler model that only includes these terms marked as significant in the full model.

```
lm.mass2 <- lm(mass.diff ~  loc+temp+food+loc:temp, data=shiners)
Anova(lm.mass2)
```

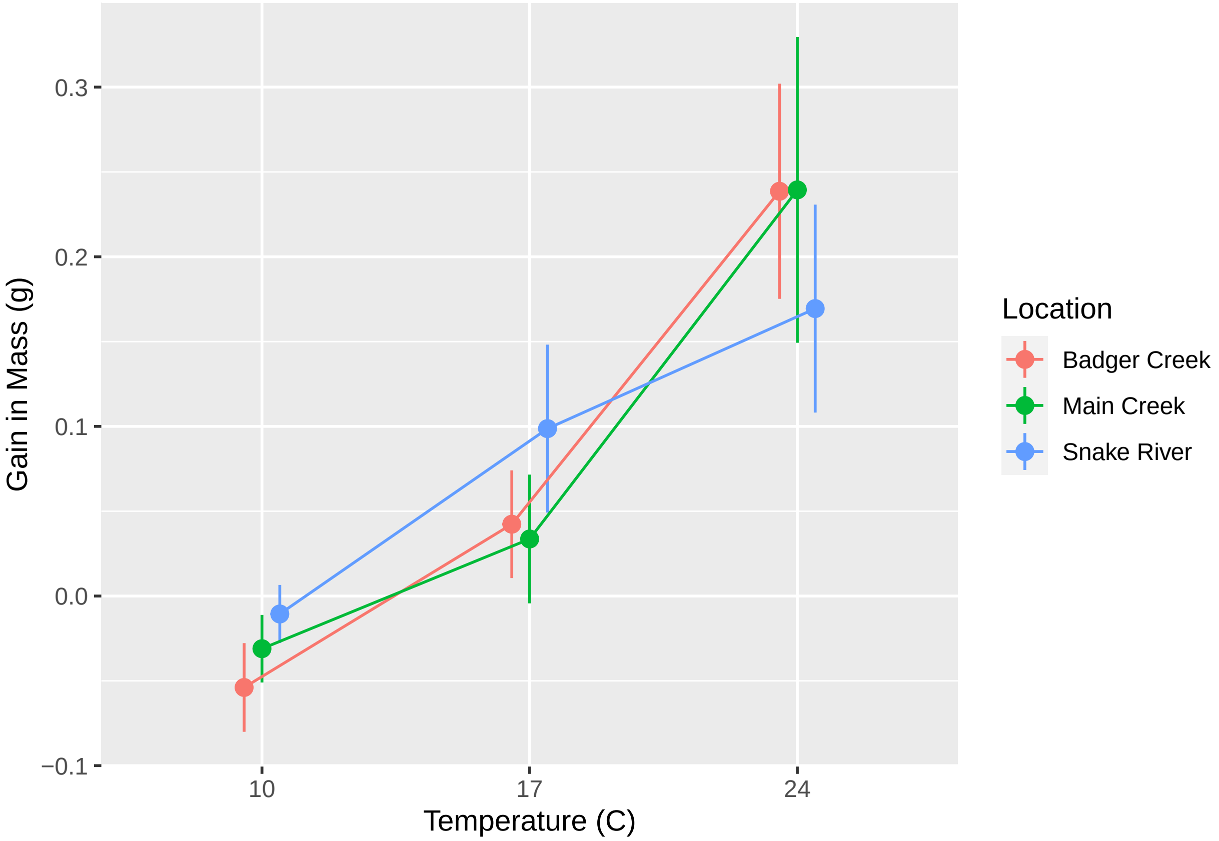
```
## Anova Table (Type II tests)
##
## Response: mass.diff
##           Sum Sq  Df F value  Pr(>F)
## loc       0.0064   2  0.1569 0.85487
## temp      3.3135   2 81.7272 < 2e-16 ***
## food      0.1268   1  6.2564 0.01289 *
## loc:temp  0.2298   4  2.8341 0.02473 *
## Residuals 6.2640 309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm.mass2)
```

```
##
## Call:
## lm(formula = mass.diff ~ loc + temp + food + loc:temp, data = shiners)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57746 -0.07023 -0.00287  0.05779  0.64101
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.07502    0.02518  -2.979  0.00312 **
## locMain Creek            0.02451    0.03334   0.735  0.46271
## locSnake River           0.04389    0.03333   1.317  0.18888
## temp17                   0.09736    0.03356   2.901  0.00399 **
## temp24                   0.29182    0.03431   8.504 7.96e-16 ***
## foodTwice Daily          0.03998    0.01599   2.501  0.01289 *
## locMain Creek:temp17   -0.03036    0.04747  -0.640  0.52291
## locSnake River:temp17  0.01496    0.04804   0.311  0.75571
## locMain Creek:temp24   -0.02130    0.04801  -0.444  0.65754
## locSnake River:temp24 -0.11123    0.04753  -2.340  0.01990 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1424 on 309 degrees of freedom
## Multiple R-squared:  0.3709, Adjusted R-squared:  0.3526
## F-statistic: 20.24 on 9 and 309 DF,  p-value: < 2.2e-16
```

As reported in the paper, it appears that the fish grew faster at higher temperatures and with more food (the coefficients on `temp17`, `temp24` and `foodTwice Daily` are significant and positive), and the individuals from the Snake River grew more slowly at 24C (the coefficient on `locSnake River:temp24` is significant and negative). We can see this last effect in the following plot.

```
ggplot(shiners, aes(temp, mass.diff, color = loc, group = loc))+
  stat_summary(fun.data = mean_cl_normal,
               position = position_dodge(width = 0.2))+
  stat_summary(fun = mean, geom = "line",
               position = position_dodge(width = 0.2))+
  labs(x = "Temperature (C)", y = "Gain in Mass (g)", color = "Location")
```



(6) Conduct the same analysis, but for `sl.diff`. Does it tell the same story?

References

Houston, Derek D., and Mark C. Belk. 2006. "Geographic Variation in Somatic Growth of Redside Shiner." *Transactions of the American Fisheries Society* 135 (3): 801–10. https://doi.org/10.1577/T05-082.1.