Biometry 2016 - Lab 10

Multiple Linear Regression Filled in Script made by Auriel April 7 2016

```
### Libraries
library(ggplot2) ### regular plotting with ggplot
library(ggfortify) ### this gives us autoplot
library(MASS) ### this will let us do stepwise regression in an automated fashion
library(ggthemes) # gives us theme_few()
library(car) # gives us durbin.watson()
```

If you want to know what version of R you are using

```
R.version$version.string
```

```
## [1] "R version 3.2.3 (2015-12-10)"
```

if you want the citation for your current version of R

Its good practice, no matter what software you use (excel, R, JMP, whatever) to cite the software, and the version you used, that way when/if someone goes back to re do your analysis they can use what you used.

```
citation()
```

```
##
## To cite R in publications use:
##
##   R Core Team (2015). R: A language and environment for
##   statistical computing. R Foundation for Statistical Computing,
##   Vienna, Austria. URL https://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2015},
##     url = {https://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating R, please
## cite it when using it for data analysis. See also
## 'citation("pkgname")' for citing R packages.
```

if you want the citation for a package

Typically you would only cite packages you use for your analysis, not graphing, but people invest a lot of time into all packages and citations help them gain some record that their work is being used and appreciated.

```r
citation("ggplot2")
```

```
##
## To cite ggplot2 in publications, please use:
##
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2009.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2009},
##     isbn = {978-0-387-98140-6},
##     url = {http://had.co.nz/ggplot2/book},
##   }
```

Open the file "sto_den.csv".

First set working directory

```r
setwd("~/Biometry_Materials/20160331_biometry_lab_10_mult_linear_regression")

### Yours will probably be different

dat <- read.csv("sto_den.csv")
```

Make sure things are ready in correctly and there aren't any crazy values

```r
head(dat)
```

```
##   POOL         SITE SPNUM  DEPTH SUBSTRAT AREAM2 COND  DO HABCOV DENSIOM
## 1    1 Bowman           2 18.720    5.000   49.7   19 5.1      7    74.3
## 2    2 Bowman           3 10.900    6.200   80.4   27 6.0      5    85.6
## 3    3 Bowman           4 14.698    5.849  253.7   29 4.8     10    73.5
## 4    1 Haw              7 17.163    5.875   63.2   32 4.4     20    60.3
## 5    2 Haw              5 18.085    6.169   44.7   30 5.1     20    68.3
## 6    3 Haw              5 18.154    4.764  185.5   29 4.7     25    72.5
##   MAXDEPTH STO_SM_DEN STO_L_DEN
## 1     48.0      0.000     0.000
## 2     28.0      0.000     0.000
## 3     42.5      0.032     0.071
## 4     38.5      0.616     0.092
## 5     36.0      0.626     0.000
## 6     34.5      0.000     0.016
```

```r
str(dat)
```

```
## 'data.frame':    16 obs. of  13 variables:
##  $ POOL      : int  1 2 3 1 2 3 1 2 3 1 ...
##  $ SITE      : Factor w/ 4 levels "Bowman      ",..: 1 1 1 2 2 2 3 3 3 4 ...
##  $ SPNUM     : int  2 3 4 7 5 5 13 7 10 3 ...
##  $ DEPTH     : num  18.7 10.9 14.7 17.2 18.1 ...
##  $ SUBSTRAT  : num  5 6.2 5.85 5.88 6.17 ...
##  $ AREAM2    : num  49.7 80.4 253.7 63.2 44.7 ...
##  $ COND      : int  19 27 29 32 30 29 30 39 49 55 ...
##  $ DO        : num  5.1 6 4.8 4.4 5.1 4.7 4 NA NA 5.8 ...
##  $ HABCOV    : int  7 5 10 20 20 25 3 10 7 40 ...
##  $ DENSIOM   : num  74.3 85.6 73.5 60.3 68.3 72.5 62.9 36.1 44.6 94 ...
##  $ MAXDEPTH  : num  48 28 42.5 38.5 36 34.5 40.5 25 24.5 33 ...
##  $ STO_SM_DEN: num  0 0 0.032 0.616 0.626 ...
##  $ STO_L_DEN : num  0 0 0.071 0.092 0 0.016 0.108 0 0.034 0 ...
```

```
summary(dat)
```

```
##       POOL                 SITE       SPNUM            DEPTH      
##  Min.   :1.000   Bowman      :3   Min.   : 2.000   Min.   : 4.15  
##  1st Qu.:1.750   Haw         :3   1st Qu.: 4.750   1st Qu.:10.68  
##  Median :2.500   Hurricane   :3   Median : 6.000   Median :14.42  
##  Mean   :2.875   Indian      :7   Mean   : 6.625   Mean   :15.42  
##  3rd Qu.:3.250                    3rd Qu.: 8.500   3rd Qu.:18.10  
##  Max.   :7.000                    Max.   :13.000   Max.   :32.19  
##                                                                   
##     SUBSTRAT         AREAM2            COND             DO       
##  Min.   :4.328   Min.   : 13.20   Min.   :19.00   Min.   :4.000  
##  1st Qu.:4.969   1st Qu.: 37.02   1st Qu.:29.75   1st Qu.:4.475  
##  Median :5.548   Median : 56.90   Median :43.00   Median :5.100  
##  Mean   :5.458   Mean   : 94.36   Mean   :40.31   Mean   :5.171  
##  3rd Qu.:5.856   3rd Qu.:108.92   3rd Qu.:50.00   3rd Qu.:5.950  
##  Max.   :6.551   Max.   :283.00   Max.   :55.00   Max.   :6.200  
##                                                   NA's   :2      
##      HABCOV         DENSIOM         MAXDEPTH       STO_SM_DEN    
##  Min.   : 3.00   Min.   :36.10   Min.   :20.00   Min.   :0.0000  
##  1st Qu.: 9.25   1st Qu.:66.95   1st Qu.:31.75   1st Qu.:0.0000  
##  Median :22.50   Median :76.53   Median :37.75   Median :0.3730  
##  Mean   :27.94   Mean   :74.83   Mean   :39.59   Mean   :0.3939  
##  3rd Qu.:40.00   3rd Qu.:86.28   3rd Qu.:41.38   3rd Qu.:0.6185  
##  Max.   :70.00   Max.   :99.00   Max.   :81.50   Max.   :1.6760  
##                                                                  
##    STO_L_DEN     
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0610  
##  Mean   :0.1135  
##  3rd Qu.:0.1190  
##  Max.   :0.5240  
## 
```

This file is based on a study in which we wished to examine the relationship between physical variables and the density of stonerollers (*Campostoma anomalum*). Stonerollers are small minnows occurring abundantly in Arkansas streams.We measured density of small fish (<80mm) and large fish (>80mm) separately. We sampled fish density in pools and measured physical variables in those same pools.

POOL = the pool ID
SITE = the stream name

the above two columns are categories, we aren't interested in those

SPNUM = species richness
DEPTH = depth in cm
SUBSTRAT = Size of substrate from 1 (sand) to 6 (boulders)
COND =Measure of dissolved ions in the water DO = Dissolved oxygen
HABCOV = Index of available cover
DENSIOM = Measure of canopy openness
MAXDEPTH = Maximum pool depth
STO_SM_DEN = Density of small stonerollers
STO_L_DEN = Density of large stonerollers

Run a regression using all predictor variables against small stoneroller density.

Example = lm(data=dat, RESPONSE ~ PREDICTOR1 + PREDICTOR2 + PREDICTOR3)

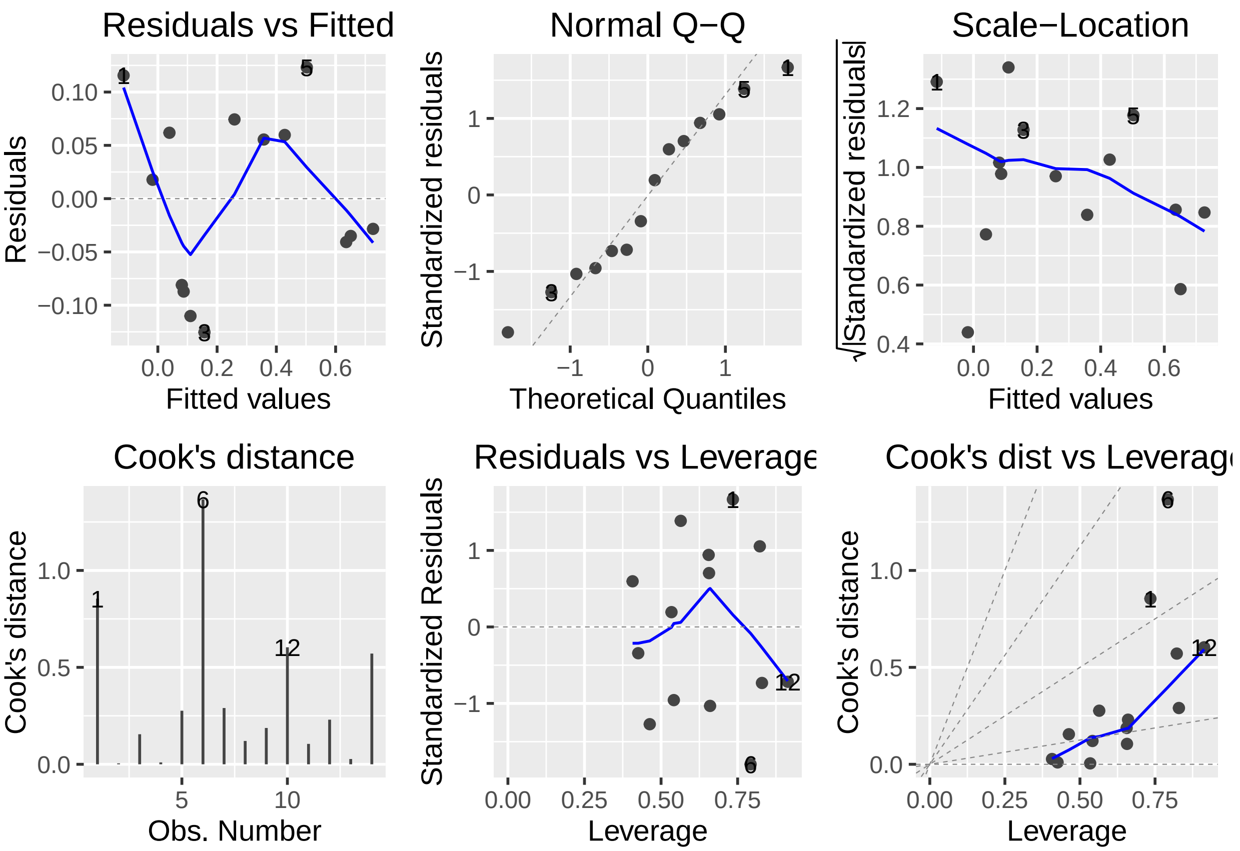
```
model <- lm(data=dat, STO_SM_DEN ~ SPNUM + DEPTH + SUBSTRAT + COND + DO + HABCOV + DENSIOM + MAXDEPTH)
summary(model)
```

```
##
## Call:
## lm(formula = STO_SM_DEN ~ SPNUM + DEPTH + SUBSTRAT + COND + DO +
##     HABCOV + DENSIOM + MAXDEPTH, data = dat)
##
## Residuals:
##        1        2        3        4        5        6        7       10
##  0.11556  0.01774 -0.12538 -0.03504  0.12314 -0.11012 -0.04072 -0.08716
##       11       12       13       14       15       16
##  0.07433 -0.02837  0.05547 -0.08099  0.06181  0.05974
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.061259   0.859543  -0.071   0.9459
## SPNUM        0.024246   0.023906   1.014   0.3570
## DEPTH        0.016044   0.023758   0.675   0.5294
## SUBSTRAT     0.100015   0.103416   0.967   0.3779
## COND         0.024066   0.010225   2.354   0.0653 .
## DO           0.214084   0.110557   1.936   0.1106
## HABCOV      -0.003017   0.004917  -0.614   0.5663
## DENSIOM     -0.031360   0.009107  -3.444   0.0184 *
## MAXDEPTH    -0.002107   0.009770  -0.216   0.8378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1345 on 5 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.9143, Adjusted R-squared:  0.7771
## F-statistic: 6.665 on 8 and 5 DF,  p-value: 0.02573
```

Here we are interested in what coefficients are significant, and what our Multiple R-squared value is.

Only Densiom is significant Our Multiple R-Squared is 0.91 which is CRAZY GOOD

```r
autoplot(model, which=1:6, ncol=3, label.size=3)
```



arguments of autoplot model = the model that you got from lm() above which = the graphs that you want, you don't need to change this ncol = the number of columns, you don't need to change this label.size = this makes the labels a bit bigger, you don't need to change this

What are the assumptions of multiple regression?

Normality Homogeneity of Variance Independence

How can we check for them?

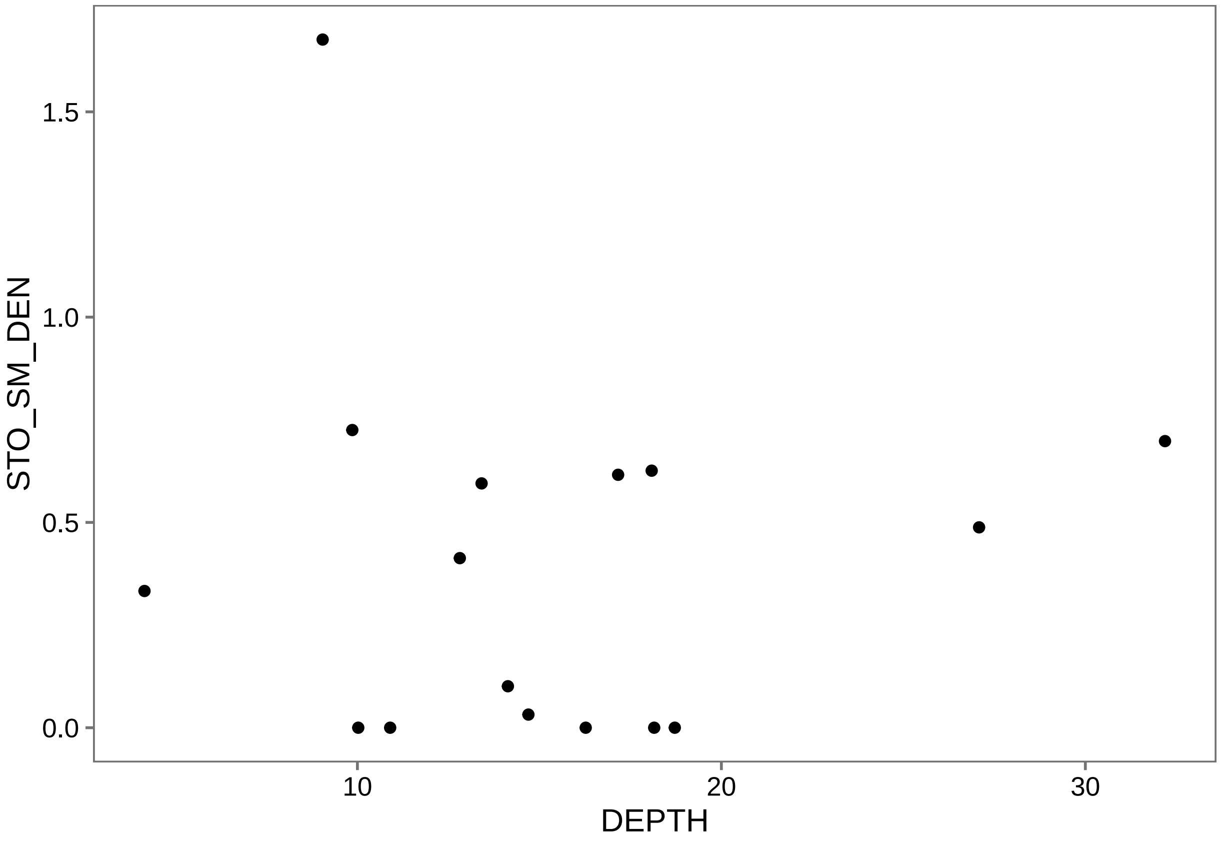
Normality - scatter plots can help, we can also look at the Normal Q-Q plot in the autoplot graph

Homogeneity of variance - scatter plots, and sometimes box plots can help, but since we are dealing with continuous variables box plots are often unhelpful, the residuals vs fitted plot in autoplot is very helpful

Independence - first set is to make sure that the experimental design is set up correctly. You can also run a Durbin's D test to test for independence and look for clumping or anti-clumping in scatter plots

Scatter plot in ggplot

```
ggplot()+geom_point(data=dat, aes(y=STO_SM_DEN, x=DEPTH))+theme_few()
```



You want to look at each predictor variable independently vs the response.

Durbin Watson's D

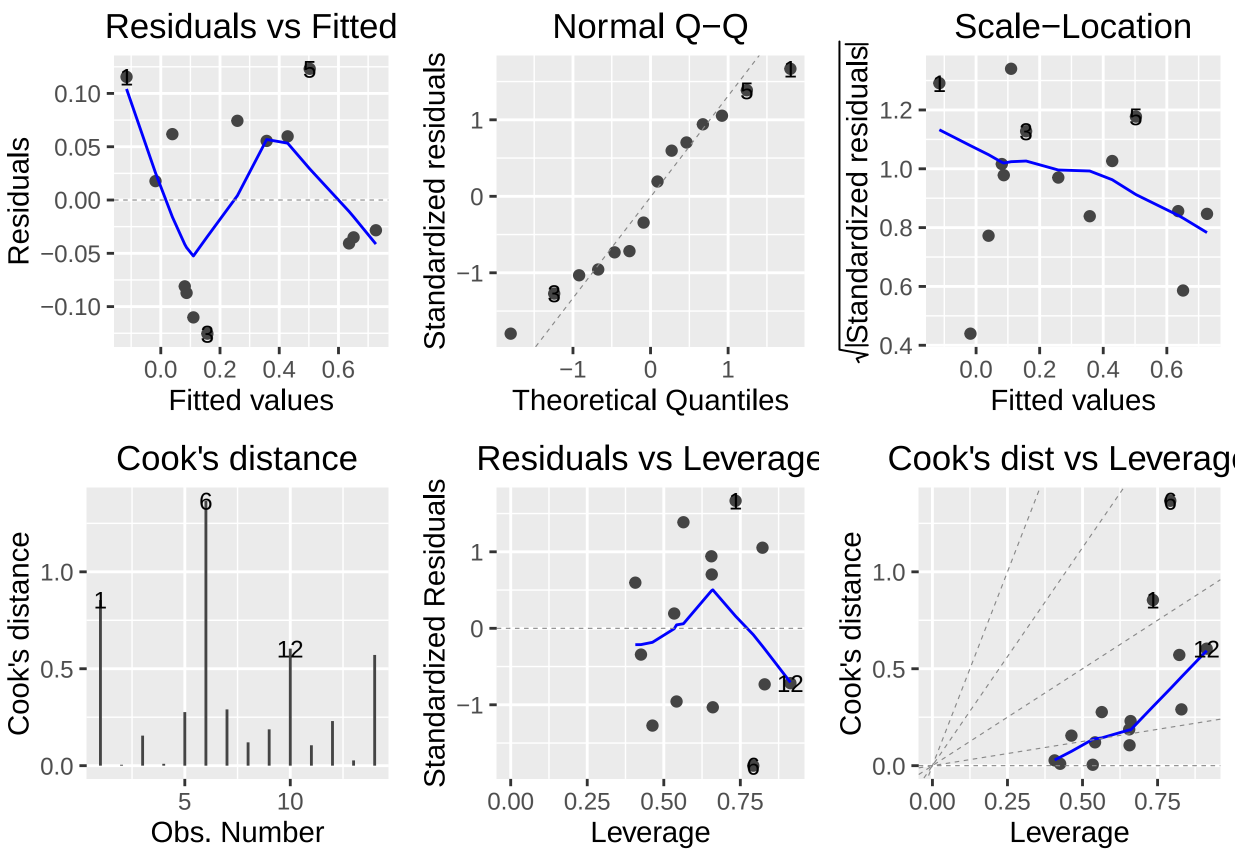
```
durbinWatsonTest(model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.2385217      2.290096   0.868
##  Alternative hypothesis: rho != 0
```

Big p value means that we are good to go. If we had a small p value we would have issues of independence

How do your residuals look?

```
autoplot(model, which=1:6, ncol=3, label.size=3)
```



```
### second column, top graph
```

```
ss <- summary(model)
ss$r.squared ### r-squared
```

```
## [1] 0.9142715
```

```
ss$adj.r.squared ### adjusted r-squared
```

```
## [1] 0.7771058
```

```
ss$coefficients ### regression coefficients
```

```
##                 Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) -0.061258604 0.859542561 -0.07126884 0.94594664
## SPNUM        0.024245588 0.023905818  1.01421286 0.35701722
## DEPTH        0.016043692 0.023757873  0.67530001 0.52944853
## SUBSTRAT     0.100014683 0.103415578  0.96711429 0.37790453
## COND         0.024066269 0.010224563  2.35377003 0.06525383
## DO           0.214083911 0.110557009  1.93641192 0.11057799
## HABCOV      -0.003016856 0.004916658 -0.61359887 0.56631796
## DENSIOM     -0.031359791 0.009106584 -3.44363941 0.01836313
## MAXDEPTH    -0.002106993 0.009769693 -0.21566627 0.83776961
```

What does the R2 value represent?

It means how much variation is being explained by the data If your rˆ2 is 0.91 that means that your predictors are explaining 91% of the variation in your response variable.

The adjusted R2?

Adjusted R squared is a method of ranking models, the actual value here is not important, you are just interested in if it is bigger or smaller then another model you are considering (like AIC)

How are they calculated?

See your textbook

What do the p values for the regression coefficients mean?

There is a p value for each predictor and that is telling you if you if the relationship for that covariate, is significantly different then zero. These relationships can change depending on the other predictors in the model. So if you are looking at the relationship with DO and you remove MAXDEPTH and rerun the model it could change.

In the ANOVA table, what does the p-value mean?

This value, the p value at the bottom of summary(model) is the significant of the entire model. This may or may not be of interest. Often its the p value of the predictors and the Rˆ2 value that we are really interested in.

Second Data Set

Open the file "multiple_regression.csv".

```r
setwd("~/Biometry_Materials/20160331_biometry_lab_10_mult_linear_regression")
dat <- read.csv("muliple_regression.csv")
```

These data are from exclusions of crayfish and fish in the Little Mulberry River.

BIOFILM is the size of the effect of fish and crayfish on stream algal communities (Positive numbers indicate grazing decreased biofilm, negative indicate increases due to grazing).

Check assumptions

(see above)

Perform a step-wise multiple regression predicting BIOFILM from the other variables.

Use a forward and a backward selection procedure in the step-wise regression.
Compare the resulting models.
Are they similar?

First we will do this the automated way

```
model <- lm(data = dat, biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
    canopy_cov_2 + max_dep_2)
both <- stepAIC(model, direction = "both")
```

```
## Start:  AIC=-17.39
## biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size + 
##     canopy_cov_2 + max_dep_2
## 
##                Df Sum of Sq    RSS     AIC
## - cray_density  1   0.00223 5.9726 -19.381
## - qdep          1   0.09783 6.0682 -19.000
## - max_dep_2     1   0.28410 6.2545 -18.274
## <none>                      5.9704 -17.390
## - canopy_cov_2  1   0.78678 6.7571 -16.419
## - pred          1   1.31220 7.2826 -14.622
## - pool_size     1   1.31639 7.2867 -14.608
## - csr_den_2     1   2.01397 7.9843 -12.414
## 
## Step:  AIC=-19.38
## biofilm ~ qdep + pred + csr_den_2 + pool_size + canopy_cov_2 + 
##     max_dep_2
## 
##                Df Sum of Sq    RSS     AIC
## - qdep          1   0.09672 6.0693 -20.995
## - max_dep_2     1   0.31811 6.2907 -20.136
## <none>                      5.9726 -19.381
## - canopy_cov_2  1   0.79502 6.7676 -18.382
## + cray_density  1   0.00223 5.9704 -17.390
## - pool_size     1   1.44906 7.4216 -16.168
## - pred          1   1.61733 7.5899 -15.630
## - csr_den_2     1   2.25651 8.2291 -13.689
## 
## Step:  AIC=-21
## biofilm ~ pred + csr_den_2 + pool_size + canopy_cov_2 + max_dep_2
## 
##                Df Sum of Sq    RSS     AIC
## - max_dep_2     1   0.27406 6.3434 -21.936
## <none>                      6.0693 -20.995
## + qdep          1   0.09672 5.9726 -19.381
## + cray_density  1   0.00112 6.0682 -19.000
## - canopy_cov_2  1   1.10258 7.1719 -18.989
## - pred          1   1.53152 7.6008 -17.595
## - pool_size     1   1.72391 7.7932 -16.995
## - csr_den_2     1   2.25282 8.3221 -15.419
## 
## Step:  AIC=-21.94
## biofilm ~ pred + csr_den_2 + pool_size + canopy_cov_2
## 
##                Df Sum of Sq    RSS     AIC
## <none>                       6.3434 -21.936
## + max_dep_2     1    0.2741  6.0693 -20.995
## - canopy_cov_2  1    0.8681  7.2115 -20.857
## + qdep          1    0.0527  6.2907 -20.136
## + cray_density  1    0.0293  6.3141 -20.046
## - csr_den_2     1    1.9864  8.3298 -17.397
## - pool_size     1    3.8818 10.2252 -12.477
## - pred          1    4.1550 10.4984 -11.844
```

```
model <- lm(data = dat, biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
    canopy_cov_2 + max_dep_2)
back <- stepAIC(model, direction = "backward")
```

```
## Start:  AIC=-17.39
## biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
##     canopy_cov_2 + max_dep_2
##
##                Df Sum of Sq    RSS     AIC
## - cray_density  1   0.00223 5.9726 -19.381
## - qdep          1   0.09783 6.0682 -19.000
## - max_dep_2     1   0.28410 6.2545 -18.274
## <none>                      5.9704 -17.390
## - canopy_cov_2  1   0.78678 6.7571 -16.419
## - pred          1   1.31220 7.2826 -14.622
## - pool_size     1   1.31639 7.2867 -14.608
## - csr_den_2     1   2.01397 7.9843 -12.414
##
## Step:  AIC=-19.38
## biofilm ~ qdep + pred + csr_den_2 + pool_size + canopy_cov_2 +
##     max_dep_2
##
##                Df Sum of Sq    RSS     AIC
## - qdep          1   0.09672 6.0693 -20.995
## - max_dep_2     1   0.31811 6.2907 -20.136
## <none>                      5.9726 -19.381
## - canopy_cov_2  1   0.79502 6.7676 -18.382
## - pool_size     1   1.44906 7.4216 -16.168
## - pred          1   1.61733 7.5899 -15.630
## - csr_den_2     1   2.25651 8.2291 -13.689
##
## Step:  AIC=-21
## biofilm ~ pred + csr_den_2 + pool_size + canopy_cov_2 + max_dep_2
##
##                Df Sum of Sq    RSS     AIC
## - max_dep_2     1   0.27406 6.3434 -21.936
## <none>                      6.0693 -20.995
## - canopy_cov_2  1   1.10258 7.1719 -18.989
## - pred          1   1.53152 7.6008 -17.595
## - pool_size     1   1.72391 7.7932 -16.995
## - csr_den_2     1   2.25282 8.3221 -15.419
##
## Step:  AIC=-21.94
## biofilm ~ pred + csr_den_2 + pool_size + canopy_cov_2
##
##                Df Sum of Sq     RSS     AIC
## <none>                       6.3434 -21.936
## - canopy_cov_2  1    0.8681  7.2115 -20.857
## - csr_den_2     1    1.9864  8.3298 -17.397
## - pool_size     1    3.8818 10.2252 -12.477
## - pred          1    4.1550 10.4984 -11.844
```

```
model <- lm(data = dat, biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
    canopy_cov_2 + max_dep_2)
forward <- stepAIC(model, direction = "forward")
```

```
## Start:  AIC=-17.39
## biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
##     canopy_cov_2 + max_dep_2
```

compare the three methods

```
both$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
##     canopy_cov_2 + max_dep_2
##
## Final Model:
## biofilm ~ pred + csr_den_2 + pool_size + canopy_cov_2
##
##
##             Step Df    Deviance Resid. Df Resid. Dev       AIC
## 1                                      16   5.970358 -17.38993
## 2 - cray_density  1 0.002230269        17   5.972589 -19.38096
## 3         - qdep  1 0.096717869        18   6.069306 -20.99543
## 4    - max_dep_2  1 0.274055626        19   6.343362 -21.93548
```

```
back$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
##     canopy_cov_2 + max_dep_2
##
## Final Model:
## biofilm ~ pred + csr_den_2 + pool_size + canopy_cov_2
##
##
##             Step Df    Deviance Resid. Df Resid. Dev       AIC
## 1                                      16   5.970358 -17.38993
## 2 - cray_density  1 0.002230269        17   5.972589 -19.38096
## 3         - qdep  1 0.096717869        18   6.069306 -20.99543
## 4    - max_dep_2  1 0.274055626        19   6.343362 -21.93548
```

```r
forward$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
##     canopy_cov_2 + max_dep_2
##
## Final Model:
## biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
##     canopy_cov_2 + max_dep_2
##
##
##   Step Df Deviance Resid. Df Resid. Dev       AIC
## 1                         16   5.970358 -17.38993
```

Perform a manual iterative selection process by including all the independent variables, then removing the one with the largest p-value.

```r
model <- lm(data = dat, biofilm ~ qdep + pred + cray_density + csr_den_2 + pool_size +
    canopy_cov_2 + max_dep_2)

summary(model)
```

```
##
## Call:
## lm(formula = biofilm ~ qdep + pred + cray_density + csr_den_2 +
##     pool_size + canopy_cov_2 + max_dep_2, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21818 -0.18866 -0.06043  0.17960  1.60800
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.307904   1.669595  -0.783   0.4449
## qdep          0.006758   0.013199   0.512   0.6156
## pred          0.471862   0.251626   1.875   0.0791 .
## cray_density -0.034809   0.450254  -0.077   0.9393
## csr_den_2     2.865522   1.233438   2.323   0.0337 *
## pool_size     0.010108   0.005382   1.878   0.0787 .
## canopy_cov_2 -0.036642   0.025235  -1.452   0.1658
## max_dep_2     0.018705   0.021437   0.873   0.3958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6109 on 16 degrees of freedom
## Multiple R-squared:  0.5934, Adjusted R-squared:  0.4155
## F-statistic: 3.335 on 7 and 16 DF,  p-value: 0.02179
```

```r
AIC(model)
```

```
## [1] 52.71912
```

R^2 = .5934

Since cray_density has the biggest p value I drop it

```r
model <- lm(data=dat, biofilm ~ qdep + pred + csr_den_2 + pool_size + canopy_cov_2 + max_dep_2)

summary(model)
```

```
##
## Call:
## lm(formula = biofilm ~ qdep + pred + csr_den_2 + pool_size +
##     canopy_cov_2 + max_dep_2, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21044 -0.19443 -0.05969  0.17640  1.61530
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.370665   1.415645  -0.968   0.3465
## qdep          0.006713   0.012794   0.525   0.6066
## pred          0.479681   0.223568   2.146   0.0466 *
## csr_den_2     2.894068   1.141949   2.534   0.0214 *
## pool_size     0.009966   0.004907   2.031   0.0582 .
## canopy_cov_2 -0.036317   0.024142  -1.504   0.1509
## max_dep_2     0.019130   0.020104   0.952   0.3547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5927 on 17 degrees of freedom
## Multiple R-squared:  0.5932, Adjusted R-squared:  0.4497
## F-statistic: 4.132 on 6 and 17 DF,  p-value: 0.009683
```

```r
AIC(model)
```

```
## [1] 50.72809
```

R^2 = 0.5932

AIC went down (good!)

qdep has the biggest p value I drop it

```r
model <- lm(data=dat, biofilm ~  pred + csr_den_2 + pool_size + canopy_cov_2 + max_dep_2)

summary(model)
```

```
## 
## Call:
## lm(formula = biofilm ~ pred + csr_den_2 + pool_size + canopy_cov_2 + 
##     max_dep_2, data = dat)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.13801 -0.20401 -0.03912  0.15884  1.62270 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  -0.789032   0.862537  -0.915   0.3724  
## pred          0.441872   0.207333   2.131   0.0471 *
## csr_den_2     2.658196   1.028389   2.585   0.0187 *
## pool_size     0.010568   0.004674   2.261   0.0364 *
## canopy_cov_2 -0.040440   0.022363  -1.808   0.0873 .
## max_dep_2     0.017558   0.019475   0.902   0.3792  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5807 on 18 degrees of freedom
## Multiple R-squared:  0.5866, Adjusted R-squared:  0.4718 
## F-statistic: 5.109 on 5 and 18 DF,  p-value: 0.004331
```

```
AIC(model)
```

```
## [1] 49.11362
```

R^2 = 0.58

AIC went down!

max_dep_2 is highest, DROP IT!

```
model <- lm(data=dat, biofilm ~ pred + csr_den_2 + pool_size + canopy_cov_2)

summary(model)
```

```
## 
## Call:
## lm(formula = biofilm ~ pred + csr_den_2 + pool_size + canopy_cov_2, 
##     data = dat)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.07850 -0.21300 -0.09659  0.05723  1.68221 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  -0.798072   0.858219  -0.930  0.36408   
## pred          0.560965   0.159013   3.528  0.00225 **
## csr_den_2     2.285280   0.936882   2.439  0.02470 * 
## pool_size     0.012987   0.003809   3.410  0.00294 **
## canopy_cov_2 -0.027464   0.017032  -1.613  0.12333   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5778 on 19 degrees of freedom
## Multiple R-squared:  0.568,  Adjusted R-squared:  0.477
## F-statistic: 6.245 on 4 and 19 DF,  p-value: 0.002204
```

```
AIC(model)
```

```
## [1] 48.17357
```

R^2 = 0.56

AIC went down!

all p values (except intercept, which we don't really care about, are under 0.15) so this is our final model

How do p-values change after each iteration?

Go back and look at the model summaries and check on this

Slope coefficients?

Go back and look at the model summaries and check on this

Return to "sto_den.csv".

```
setwd("~/Biometry_Materials/20160331_biometry_lab_10_mult_linear_regression")
dat <- read.csv("sto_den.csv")
```

Now I want to know which of the physical variables best explains the variation in density and how much variation the model explains overall.

First check for correlations among independent variables.

```
dat <- dat[!is.na(dat$DO),] # we have two missing DO values, so we are removing those rows

cor(dat[,c("SPNUM","DEPTH","SUBSTRAT","AREAM2","COND","DO","HABCOV","DENSIOM","MAXDEPTH","STO_SM_DEN")]
```

```
##                 SPNUM       DEPTH     SUBSTRAT        AREAM2
## SPNUM       1.00000000  0.471357629 -0.0459711269  0.4175807538
## DEPTH       0.47135763  1.000000000 -0.4773922635  0.3360976946
## SUBSTRAT   -0.04597113 -0.477392263  1.0000000000 -0.0008942959
## AREAM2      0.41758075  0.336097695 -0.0008942959  1.0000000000
## COND        0.25062751 -0.096791757 -0.4206081646 -0.3350939612
## DO         -0.42686834 -0.107145096 -0.1911768801 -0.5618089331
## HABCOV      0.11140539  0.008236916 -0.3815316337 -0.4646460511
## DENSIOM    -0.26763559 -0.206369443 -0.3466291735 -0.4260407405
## MAXDEPTH    0.56067083  0.933068828 -0.5247924119  0.3370580507
## STO_SM_DEN  0.74472079  0.413878262  0.0762326286  0.0835749387
##                   COND          DO       HABCOV     DENSIOM    MAXDEPTH
## SPNUM       0.25062751 -0.42686834  0.111405393 -0.26763559  0.56067083
## DEPTH      -0.09679176 -0.10714510  0.008236916 -0.20636944  0.93306883
## SUBSTRAT   -0.42060816 -0.19117688 -0.381531634 -0.34662917 -0.52479241
## AREAM2     -0.33509396 -0.56180893 -0.464646051 -0.42604074  0.33705805
## COND        1.00000000  0.32471696  0.861315614  0.66726343  0.06574085
## DO          0.32471696  1.00000000  0.327009479  0.80811403 -0.05176563
## HABCOV      0.86131561  0.32700948  1.000000000  0.64937587  0.05671468
## DENSIOM     0.66726343  0.80811403  0.649375866  1.00000000 -0.07622653
## MAXDEPTH    0.06574085 -0.05176563  0.056714685 -0.07622653  1.00000000
## STO_SM_DEN  0.11083322 -0.41511728 -0.033618803 -0.51545691  0.41773892
##             STO_SM_DEN
## SPNUM       0.74472079
## DEPTH       0.41387826
## SUBSTRAT    0.07623263
## AREAM2      0.08357494
## COND        0.11083322
## DO         -0.41511728
## HABCOV     -0.03361880
## DENSIOM    -0.51545691
## MAXDEPTH    0.41773892
## STO_SM_DEN  1.00000000
```

Select four candidate models and rank them according to AIC(corrected) values (smaller AIC = better/more parsimonious models).

```
model1 <- lm(data = dat, STO_SM_DEN ~ AREAM2 + COND + DO + HABCOV)
model2 <- lm(data = dat, STO_SM_DEN ~ MAXDEPTH + DO + COND)
model3 <- lm(data = dat, STO_SM_DEN ~ DENSIOM + DO + AREAM2)
model4 <- lm(data = dat, STO_SM_DEN ~ SPNUM + DEPTH + SUBSTRAT + COND + DO +
    HABCOV + DENSIOM + MAXDEPTH)


AIC(model1)
```

```
## [1] 9.850567
```

```
AIC(model2)
```

```
## [1] 6.834037
```

```
AIC(model3)
```

```
## [1] 8.700541
```

```
AIC(model4)
```

```
## [1] -10.84772
```

Do this separately for small and large fish, and then compare the two.

```
model1 <- lm(data = dat, STO_L_DEN ~ AREAM2 + COND + DO + HABCOV)
model2 <- lm(data = dat, STO_L_DEN ~ MAXDEPTH + DO + COND)
model3 <- lm(data = dat, STO_L_DEN ~ DENSIOM + DO + AREAM2)
model4 <- lm(data = dat, STO_L_DEN ~ SPNUM + DEPTH + SUBSTRAT + COND + DO +
    HABCOV + DENSIOM + MAXDEPTH)


AIC(model1)
```

```
## [1] -6.792955
```

```
AIC(model2)
```

```
## [1] -22.09144
```

```
AIC(model3)
```

```
## [1] -3.431584
```

```
AIC(model4)
```

```
## [1] -20.60188
```

Which variables did you use in your model?

I just strung some together at random, go through and make your own :)

Why?

What does the AIC value take into account?

See text book

For your top model, can you write out the linear model with the beta coefficients?