Lab # 11: Correlation and Model Fitting

Objectives:

1. Correlations between variables
2. Data Manipulation, creation of squares
3. Model fitting with regression
4. Comparison of models

Correlations between variables

Throughout the semester we have learned various methods of trying to identify relationships between variables. When they were categorical our only option was to check the corresponding contingency table and try the chi-square analysis. When we were trying to compare a numerical variable with a categorical one we used TTest and ANOVA. In this section we will explore the idea of comparing two numerical variables, for which we have a one to one correspondence between each value in our dataset. We will do that with the use of the correlation coefficient (Pearson) mathematically and a simple scatterplot visually.

We will then use those results to build a model that will predict our output variable using the input variables as best as possible.

The example we will use today is the dataset Larvae.csv which you can find on our MOODLE page and contains the measurements of the Headwidth, Length and Mass, including the gender for the Sirex Noctilio Larvae that we have seen in the past. Our output variable here will be Mass and we will try to correlate that to Headwidth and Length respectively. Apparently, Mass is the hardest measurement of the three, since both Headwidth and Length are performed automatically by a modified photometer-microscope, so a model to infer Mass based on the other two would be useful.

First let's import our dataset by using the by now well-known proc import command as follows:

```
PROC IMPORT OUT= Larvae
            DATAFILE= "put the complete path here…\Larvae.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
```

The dataset is quite large, so we will modify a bit the proc print procedure to only output the first 10 observations only

```
PROC PRINT DATA = Larvae (obs=10);
RUN;
```

What we would like to do then is compute the correlation coefficients for the pairs Mass with Length and Mass with Headwidth respectively:

The procedure for that is sas is **PROC COR.** The following code will compute those correlations and output them in a nice matrix way

```
Title "Correlations Mass vs Length and Mass vs Headwidth";
PROC CORR Data=Larvae;
Var Mass;
With Length Headwidth;
RUN;
```

We declare the output variable by using "**VAR** Mass" and the variables we want to compare it with (inputs) by using "**With** Length Headwidth". After a table with simple statistics for each variable, we get the t he pearson correlations between Mass and the other two in a near looking table.

Pearson Correlation Coefficients, N = 4895 Prob > |r| under H0: Rho=0	
	Mass
Length	0.83863 <.0001
Headwidth	0.80712 <.0001

As we see, both correlation coefficients are quite high (close to 1) and thus we can assume that Mass is positively correlated with Length and it is also positively correlated with Headwidth.
Now if one wants to just compute all possible correlations amongst variables, the code is actually just:

```
Title "Correlations amongst all variables";
PROC CORR Data=Larvae;
RUN;
```

As you see if you don't declare what is your input and what is your output, SAS is going to compute all pairwise correlations.

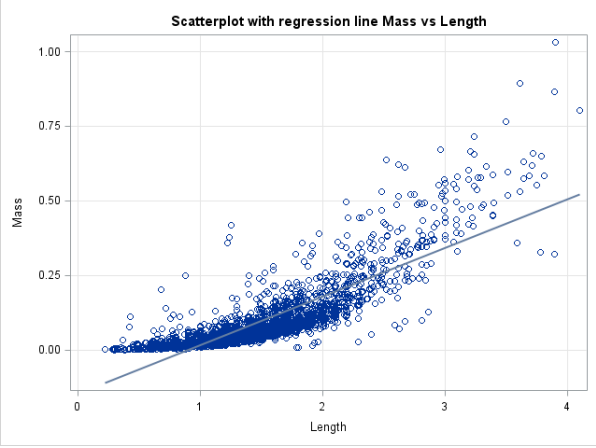
Pearson Correlation Coefficients, N = 4895 Prob > |r| under H0: Rho=0			
	Length	**Headwidth**	**Mass**
Length	1.00000	0.85269 <.0001	0.83863 <.0001
Headwidth	0.85269 <.0001	1.00000	0.80712 <.0001
Mass	0.83863 <.0001	0.80712 <.0001	1.00000

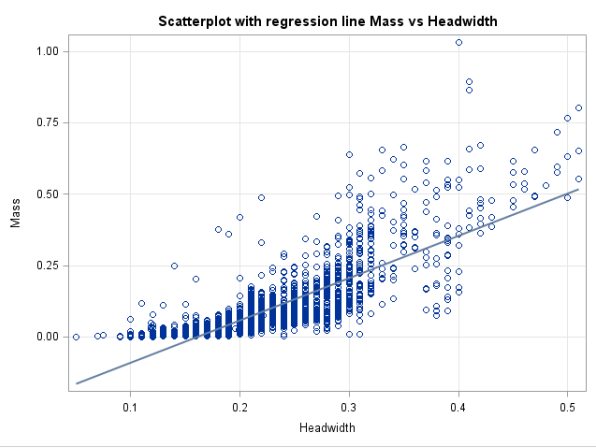
Armed with the knowledge that Mass is correlated with the other two variables, let's create the corresponding scatterplots that will showcase that connection using SGSCATTER. First we will see a relationship between Mass and Length by using the code which will also output the best fit regression line:

```
Title "Scatterplot with regression line Mass vs Length";
PROC SGSCATTER Data=Larvae;
Plot Mass*Length/ Reg= (nogroup degree=1) grid;
RUN;
```

Then let's explore the relationship between Mass and Headwidth by the following code

```
Title "Scatterplot with regression line Mass vs Headwidth";
PROC SGSCATTER Data=Larvae;
Plot Mass*Headwidth/ Reg= (nogroup degree=1) grid;
RUN;
```



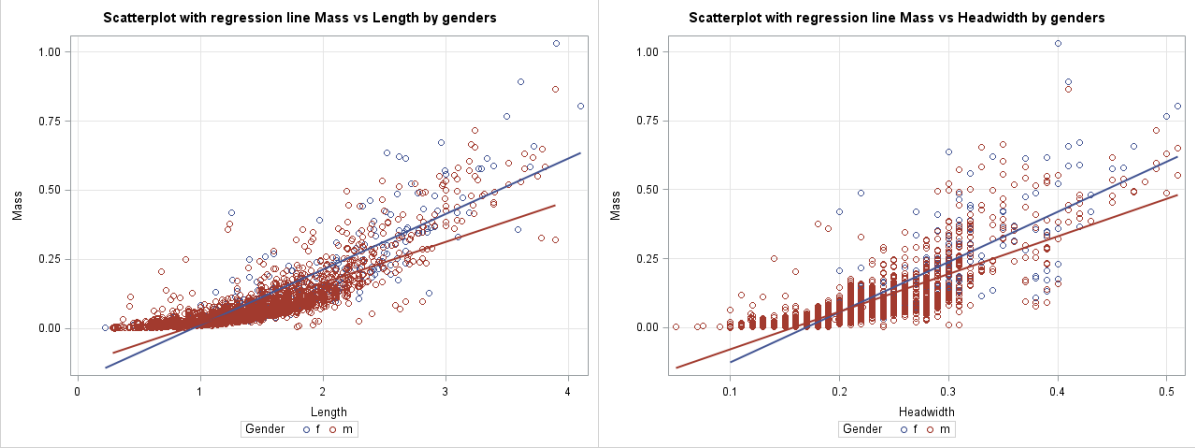


Since we have information about the gender of the larvae, we could reflect that in our analysis by creating colored scatterplots and different fits. You can do that by extending the codes using the **group=gender** add on and removing the "nogroup" statement from the regression.

```
Title "Scatterplot with regression line Mass vs Length by genders";
PROC SGSCATTER Data=Larvae;
Plot Mass*Length/ group=Gender Reg= (degree=1) grid;
RUN;
```

Similarly for Mass compared to Headwidth:

```
Title "Scatterplot with regression line Mass vs Headwidth by genders";
PROC SGSCATTER Data=Larvae;
Plot Mass*Headwidth/ group=Gender Reg= (degree=1) grid;
RUN;
```

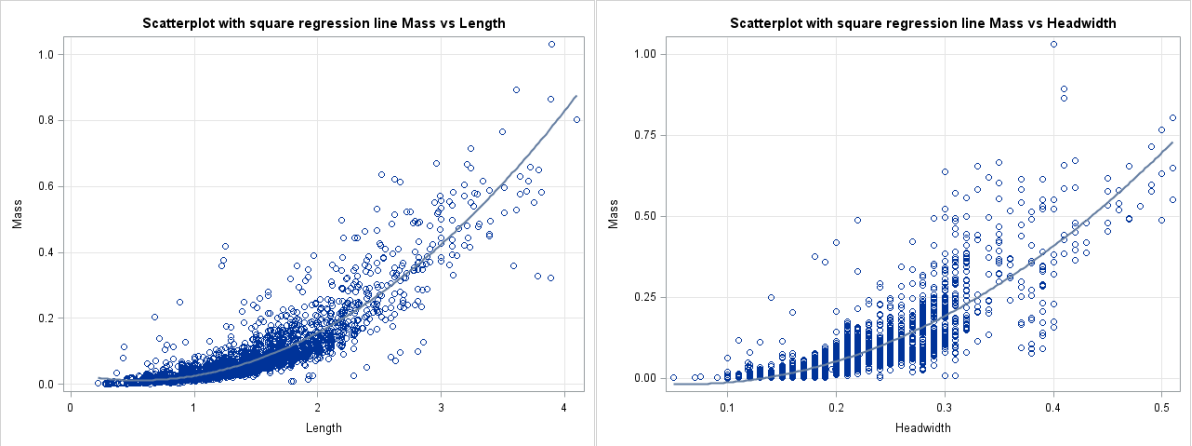


Unfortunately, the multitude of red points overshadows the blue making this harder to read. But still this is a much better looking picture.

Before we move forward, just by looking at the shape of the scatterplots, it is almost obvious that a higher degree polynomial would fit a little bit better. Let's try that with a second degree polynomial for both independent of genders.

```
Title "Scatterplot with square regression line Mass vs Length";
PROC SGSCATTER Data=Larvae;
Plot Mass*Length/ Reg= (nogroup degree=2) grid;
RUN;
```

```
Title "Scatterplot with square regression line Mass vs Headwidth";
PROC SGSCATTER Data=Larvae;
Plot Mass*Headwidth/ Reg= (nogroup degree=2) grid;
RUN;
```



Besides being IMMENSLY more aesthetically pleasing, this also gives as an idea as to what the right regression model might be. Namely, perhaps it is the squares of Length and Headwidth that we should be using to predict the mass, not the Length and Headwidth themselves.

Data Manipulations, creation of Squares of columns

In the previous section we noticed that having the square of Length and Headwidth as input might be a good idea. Thus we will instruct SAS to expand our dataset by adding two new columns, called Length2 and Headwidth2 corresponding to the squares of Length and Headwidth respectively. Basically we will create a new dataset called Larvae2 extending the old one as follows:

```
Data Larvae2;
SET Larvae;
Length2= Length*Length;
Headwidth2=Headwidth*Headwidth;
RUN;
```

Again, we can view the first 10 observations with

```
PROC PRINT DATA = Larvae2 (obs=10);
RUN;
```

The new columns of the squares of Length and Headwidth are now added. Using similar methods we can create any algebraic expression for our variables or even between variables and save them as new columns.

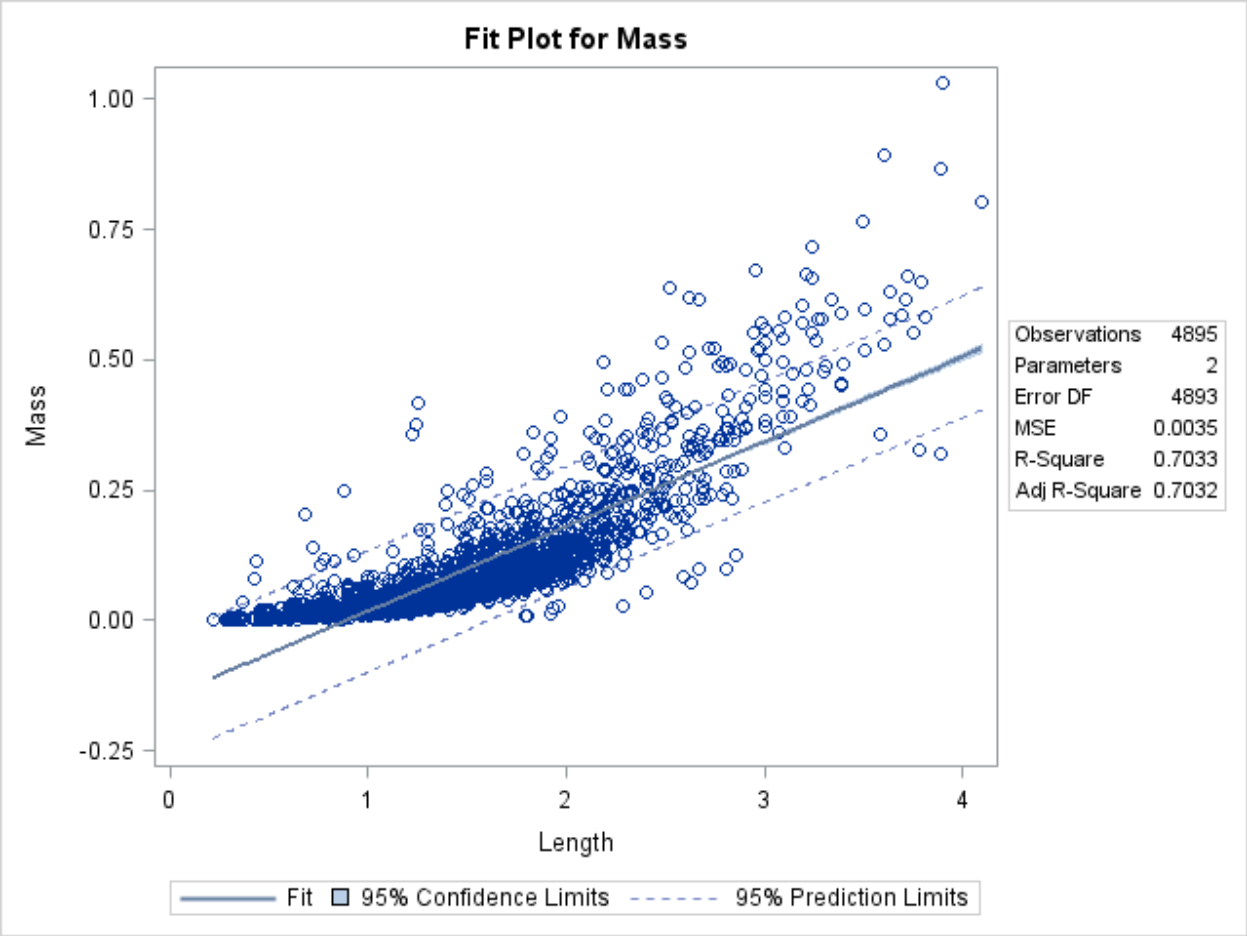
Model fitting with regression

Suppose now that we want to find the best regression model that fits our dataset, and use that to predict the mass given the other information. We will be using the extended dataset Larvae2. We start by a simple regression between Mass and Length followed by one for Mass and Width.

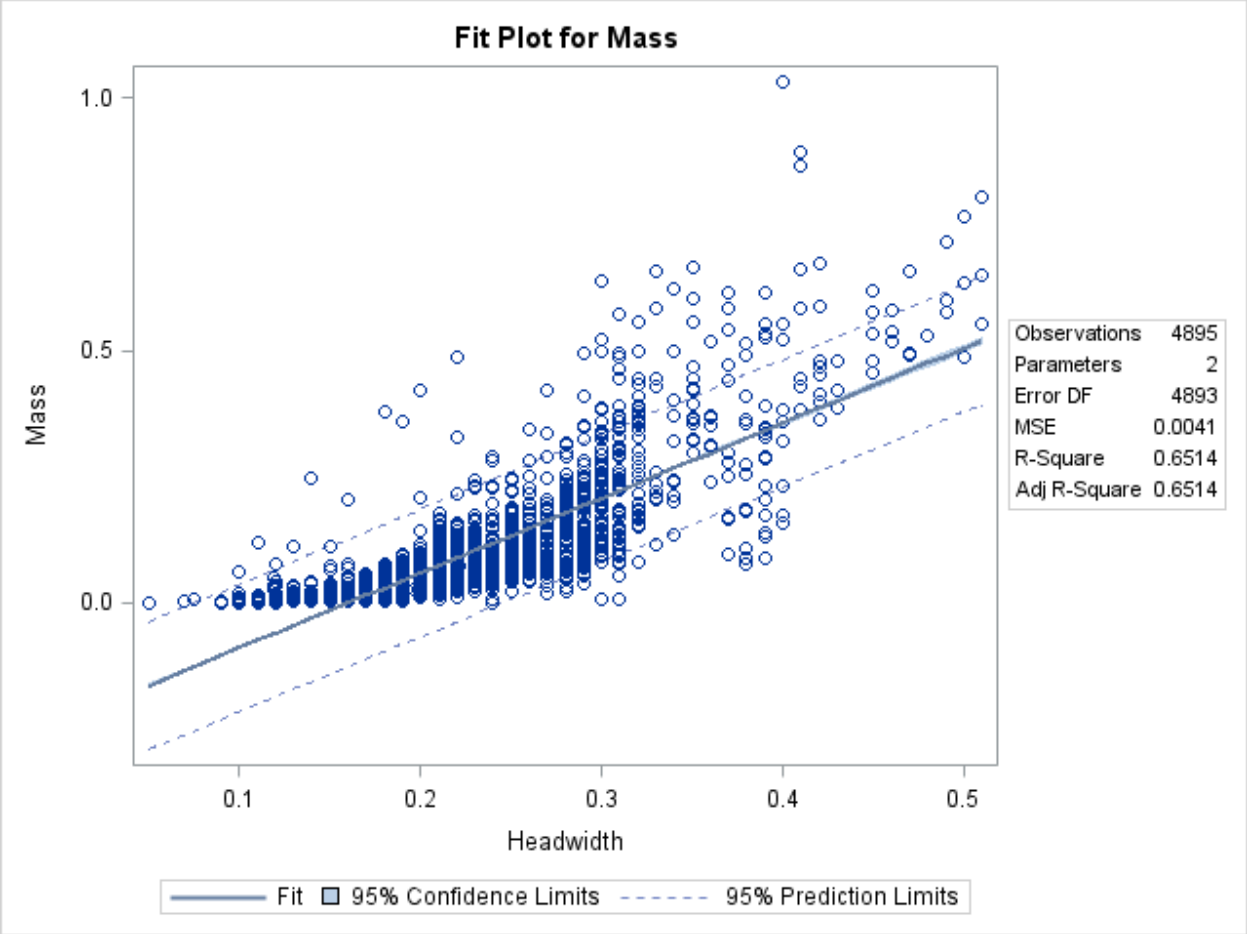
```
Title "Linear Regression Mass vs Length";
PROC REG Data=Larvae2;
Model Mass=Length;
RUN;
```

```
Title "Linear Regression Mass vs Headwidth";
PROC REG Data=Larvae2;
Model Mass=Headwidth;
RUN;
```

Giving us the following plots



And



Looking at the R-square the regression using the length performs better than that of the Headwidth since 0.7033 is larger than 0.6514. (The closer to 1 the better)

Looking at the MSE the regression using the length again performs better than that of the Headwidth since 0.0035 is smaller than 0.0041. (The closer to 0 the better)

Thus if we were to use one linear model it should be the one with Length. By looking at both the ANOVAs we can also say that both Length and Headwith are connected to Mass, although our previous analysis with the correlations is a stronger indicator.

Finally the corresponding formulas for the regressions are:

$$\boldsymbol{Mass = -0.14601 + 0.1625 * Length + \varepsilon}$$

And:

$$\boldsymbol{Mass = -0.23842 + 1.48251 * Headwidth + \varepsilon}$$

Comparison of Models:

The rest of this section is a preview of higher level classes. It also serves to point out that the statistical analysis that we did in regression translates to other models not just straight lines. We will attempt various models to predict Mass from the other variables and output a measure to compare them by.

First let us run various models by using different combinations in the line mode Mass= …

a) Mass as a linear function of Length and Headwidth.

This is an example of a multilinear model thus we are aiming for a relationship of the form:

$$Mass = \beta_0 + \beta_1 * Length + \beta_2 * Headwidth + \varepsilon$$

The appropriate code is

```
Title "Linear Regression Mass vs Length and Headwidth";
PROC REG Data=Larvae2;
Model Mass=Length Headwidth;
RUN;
```

The result is very similar to the one we get from classic regression. The parameter estimates contains the coefficients we want which are read as follows:

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	-0.20214	0.00322	-62.74	<.0001
Length	**1**	0.10679	0.00273	39.07	<.0001
Headwidth	**1**	0.61936	0.02591	23.90	<.0001

$$Mass = -0.20214 + 0.10679 * Length + 0.61936 * Headwidth + \varepsilon$$

What we also need to note here is the R-square value found in the table:

Root MSE	0.05613	**R-Square**	0.7343
Dependent Mean	0.08785	**Adj R-Sq**	0.7342
Coeff Var	63.89541		

This will be our way of comparing different models. The one with the highest R-square is the one we need.

b) Mass as a function of Length and Length squared.

The formula then is

$$\boldsymbol{Mass = \beta_0 + \beta_1 * Length + \beta_2 * Length^2 + \varepsilon}$$

The corresponding code is:

```
Title "Mass vs Length and Length2";
PROC REG Data=Larvae2;
Model Mass=Length Length2;
RUN;
```

Which yields:

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	0.03243	0.00376	8.61	<.0001
Length	**1**	-0.07522	0.00453	-16.59	<.0001
Length2	**1**	0.06857	0.00126	54.34	<.0001

And thus:

$$\boldsymbol{Mass = 0.03243 - 0.07522 * Length + 0.06857 * Length^2 + \varepsilon}$$

The R-square value is then:

Root MSE	0.04684	**R-Square**	0.8150
Dependent Mean	0.08785	**Adj R-Sq**	0.8149
Coeff Var	53.32208		

Let us compare these two new models and add the linear regressions from before:

Model	Formula	R-square
1	$\boldsymbol{Mass = -0.14601 + 0.1625 * Length + \varepsilon}$	**0.7033**
2	$\boldsymbol{Mass = -0.23842 + 1.48251 * Headwidth + \varepsilon}$	**0.6514**
3	$\boldsymbol{Mass = -0.20214 + 0.10679 * Length + 0.61936 * Headwidth + \varepsilon}$	**0.7343**
4	$\boldsymbol{Mass = 0.03243 - 0.07522 * Length + 0.06857 * Length^2 + \varepsilon}$	**0.8150**

As we mentioned earlier the one with the highest R-square value is the one we should be focusing on. So we say that the last model is the one that predicts Mass the best.