

# Simple Linear Regression Handout

## 1 Initialization

```
> library(NCStats)
```

## 2 Salmon Sperm Example

### 2.1 Data Preparation

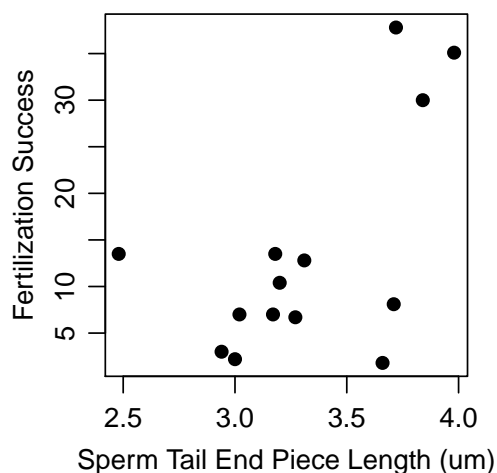
You must change the directory to where the following file is located.

```
> ss <- read.table("SalmonSperm.txt",head=TRUE)
> str(ss)

'data.frame':      14 obs. of  3 variables:
 $ step.len : num  2.48 2.94 3 3.02 3.17 3.18 3.2 3.27 3.31 3.66 ...
 $ fert.succ: num  13.5 3 2.2 7 7 13.5 10.4 6.7 12.8 1.8 ...
 $ mat      : Factor w/ 2 levels "Adult","Parr": 1 2 2 1 2 1 1 1 1 2 ...
```

The next two commands save the name to be used for axes labels into objects to save typing in later commands.

```
> xlbl <- "Sperm Tail End Piece Length (um)"
> ylbl <- "Fertilization Success"
> plot(fert.succ~step.len,data=ss,xlab=xlbl,ylab=ylbl,pch=19)
```



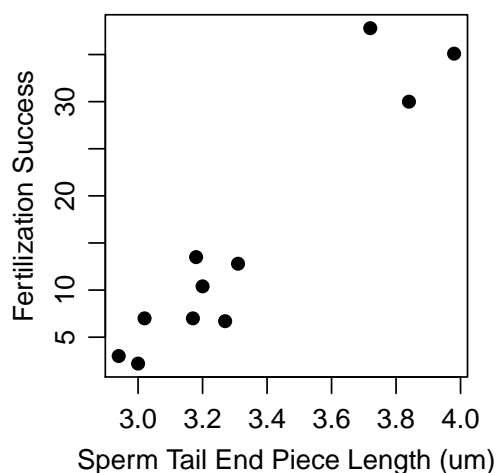
The following command was not evaluated in R because it can be used to interactively click on points on the plot above. When you hit the “STOP” button after clicking on the last point then it will give you a list

of the row numbers for the selected points. This command must be issued while the plot above is still active in R. I used this command to identify the individuals that were apparent outliers.

```
> identify(fert.succ~step.len,data=ss)
```

I removed the three outliers from the data set with the commands below. I removed these outliers only to simplify this analysis so that you could remain focused on the concepts of simple linear regression. It is generally NOT good practice to simply remove outliers without considering them further.

```
> ss1 <- ss[-c(1,10,11),]
> plot(fert.succ~step.len,data=ss1,xlab=xlbl,ylab=ylbl,pch=19)
```



## 2.2 Lecture Support I – Model Fitting and Simple Predictions

```
> lm1 <- lm(fert.succ~step.len,data=ss1)
> summary(lm1)
```

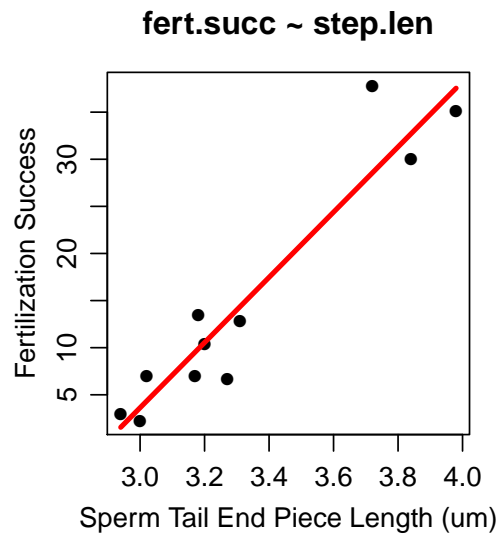
```
Call:
lm(formula = fert.succ ~ step.len, data = ss1)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-6.269  -2.475  -1.424   2.068   9.257
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -100.205     13.015  -7.699 3.00e-05
step.len       34.610       3.889   8.901 9.35e-06
```

```
Residual standard error: 4.366 on 9 degrees of freedom
Multiple R-squared:  0.898,    Adjusted R-squared:  0.8866
F-statistic: 79.22 on 1 and 9 DF,  p-value: 9.35e-06
```

```
> fitPlot(lm1,xlab=xlbl,ylab=ylbl)
```



```
> predict(lm1, data.frame(step.len=3.5))
```

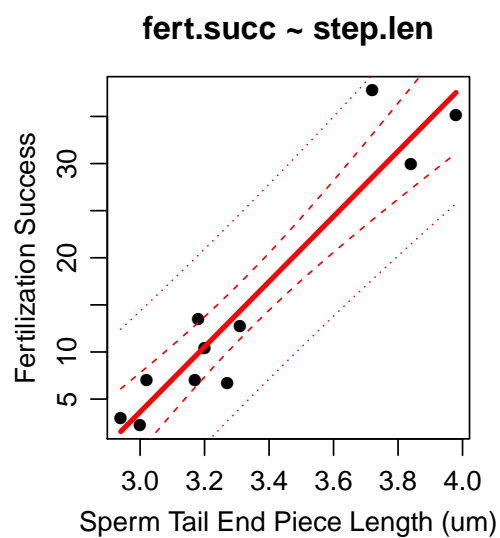
```
1
20.92912
```

## 2.3 Lecture Support II – Sampling Variability

```
> confint(lm1)
```

```
                2.5 %    97.5 %
(Intercept) -129.64815 -70.76202
step.len      25.81336  43.40619
```

```
> fitPlot(lm1, xlab=xlbl, ylab=ylbl, interval="both")
```



```
> predict(lm1,data.frame(step.len=3.5),interval="c")
```

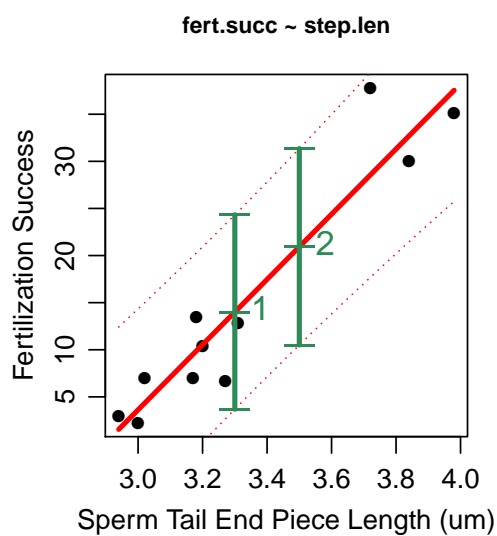
```
      fit      lwr      upr
1 20.92912 17.5967 24.26153
```

```
> predict(lm1,data.frame(step.len=3.5),interval="p")
```

```
      fit      lwr      upr
1 20.92912 10.50502 31.35321
```

```
> predictionPlot(lm1,data.frame(step.len=c(3.3,3.5)),interval="p",xlab=xlbl,ylab=ylbl)
```

```
  obs step.len      fit      lwr      upr
1   1      3.3 14.00716  3.687506 24.32682
2   2      3.5 20.92912 10.505016 31.35321
```



## 2.4 Lecture Support III – Model Comparisons

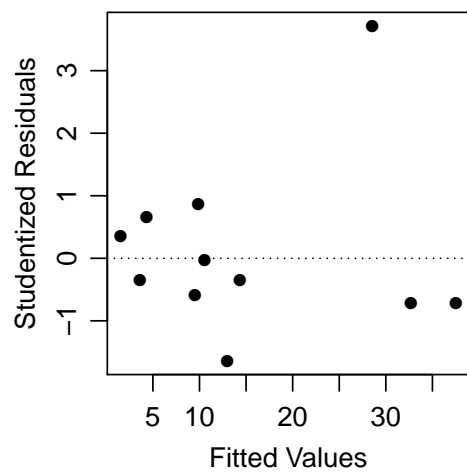
```
> anova(lm1)
```

```
Analysis of Variance Table
```

```
Response: fert.succ
      Df Sum Sq Mean Sq F value    Pr(>F)
step.len  1 1510.23  1510.23   79.22 9.35e-06
Residuals  9  171.58    19.06
```

## 2.5 Lecture Support IV – Assumption Checking

```
> residPlot(lm1)
```

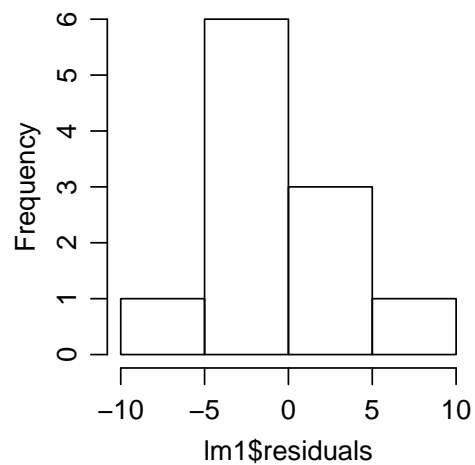


```
> adTest(lm1$residuals)
```

```
Anderson-Darling normality test
```

```
data: lm1$residuals  
A = 0.4022, p-value = 0.2962
```

```
> hist(lm1$residuals,main="")
```



```
> outlierTest(lm1)
```

```
No Studentized residuals with Bonferonni p < 0.05  
Largest |rstudent|:  
  rstudent unadjusted p-value Bonferonni p  
12 3.717896      0.0058892      0.064781
```

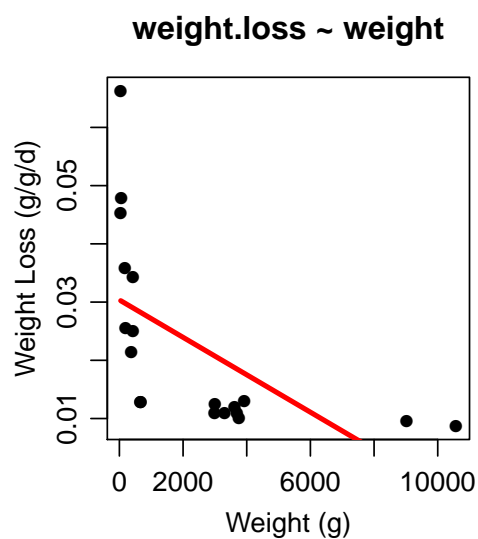
### 3 Petrels Example

The following assumes you have loaded the appropriate packages (as shown in Section 1) and changed the working directory to the folder containing the external data file.

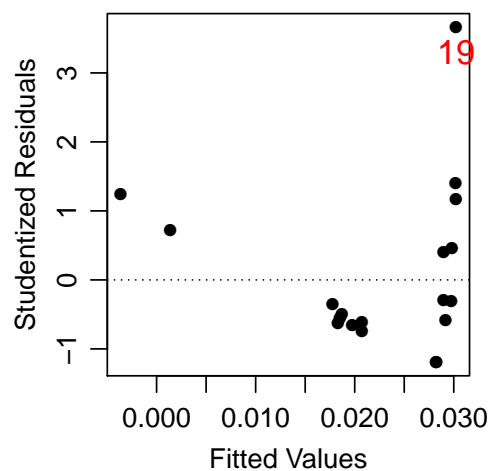
```
> petrels <- read.table("petrels.txt",head=TRUE)
> str(petrels)

'data.frame':      19 obs. of  4 variables:
 $ species      : Factor w/ 13 levels "Diomedea chrysostoma",...: 2 2 4 4 1 1 3 3 3 9 ...
 $ sex          : Factor w/ 4 levels "both","female",...: 3 2 3 2 3 2 3 2 1 3 ...
 $ weight       : int  10577 9022 3922 3694 3751 3624 3305 3000 2996 668 ...
 $ weight.loss: num  0.0087 0.0096 0.013 0.011 0.01 0.012 0.011 0.0125 0.0109 0.0128 ...

> lm1 <- lm(weight.loss~weight,data=petrels)
> fitPlot(lm1,xlab="Weight (g)",ylab="Weight Loss (g/g/d)")
```



```
> residPlot(lm1)
```



```
> with(petrels,max(weight)/min(weight))
```

```
[1] 251.8333
```

The following command is an interactive command and, thus, the result cannot be shown below. Because the max:min ratio for the weight variable was greater than 10 I set the *lambda.x* slider bar to 0 (for natural log) and then manipulated the *lambda.y* variable to find an appropriate transformation for the response variable. This process led me to use the natural log for *weight.loss* also.

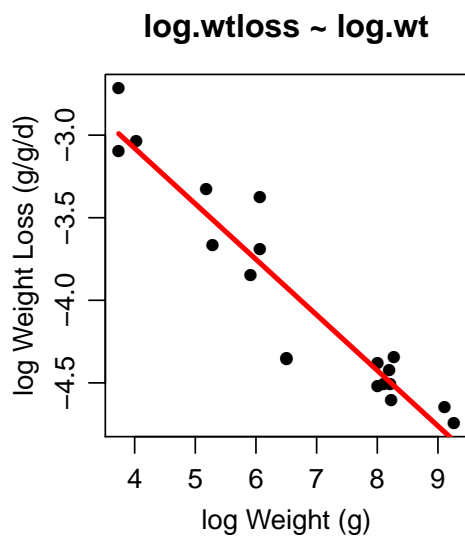
```
> transChooser(lm1)
```

```
> petrels$log.wt <- log(petrels$weight)
```

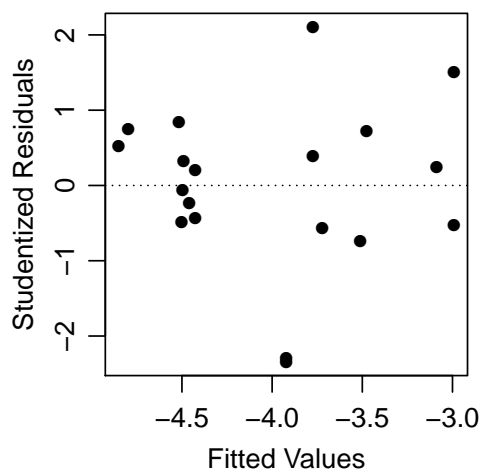
```
> petrels$log.wtloss <- log(petrels$weight.loss)
```

```
> lm2 <- lm(log.wtloss~log.wt,data=petrels)
```

```
> fitPlot(lm2,xlab="log Weight (g)",ylab="log Weight Loss (g/g/d)")
```



```
> residPlot(lm2)
```



```

> adTest(lm2$residuals)

Anderson-Darling normality test

data:  lm2$residuals
A = 0.3881, p-value = 0.3514

> anova(lm2)

Analysis of Variance Table

Response: log.wtloss
      Df Sum Sq Mean Sq F value    Pr(>F)
log.wt   1  6.5113   6.5113  140.65 1.204e-09
Residuals 17  0.7870   0.0463

> summary(lm2)

Call:
lm(formula = log.wtloss ~ log.wt, data = petrels)

Residuals:
    Min       1Q   Median       3Q      Max
-0.4368 -0.1033  0.0447  0.1279  0.4008

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.73403    0.19792  -8.761 1.04e-07
log.wt       -0.33632    0.02836 -11.860 1.20e-09

Residual standard error: 0.2152 on 17 degrees of freedom
Multiple R-squared:  0.8922,    Adjusted R-squared:  0.8858
F-statistic: 140.6 on 1 and 17 DF,  p-value: 1.204e-09

> confint(lm2)

                2.5 %      97.5 %
(Intercept) -2.1516113 -1.3164546
log.wt       -0.3961507 -0.2764885

> p.log.wtloss <- predict(lm2,data.frame(log.wt=log(5000)),interval="c")
> p.log.wtloss

      fit      lwr      upr
1 -4.598532 -4.746569 -4.450495

> exp(p.log.wtloss)*exp(anova(lm2)[2,3]/2)

      fit      lwr      upr
1 0.01030234 0.008884726 0.01194614

```