STAT3612_2312_DM_3RD_TUTORIAL

YOU Jia

2/14/2017

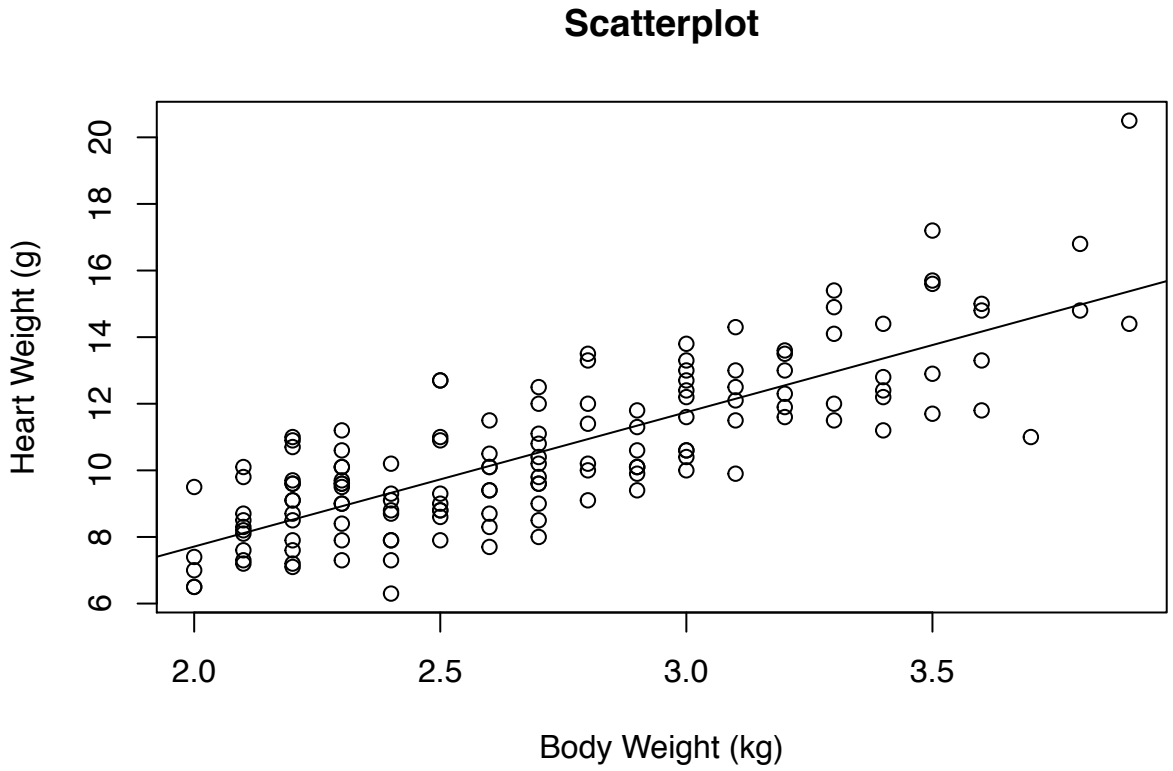
Simple Linear Regression

```
library(MASS)
attach(cats)
head(cats)
```

```
##   Sex Bwt Hwt
## 1   F 2.0 7.0
## 2   F 2.0 7.4
## 3   F 2.0 9.5
## 4   F 2.1 7.2
## 5   F 2.1 7.3
## 6   F 2.1 7.6
```

Let's do a simple linear regression using cats' body weights to predict heart weights

```
fit1 = lm(Hwt~Bwt,data=cats)
plot(Bwt,Hwt, xlab = "Body Weight (kg)", ylab="Heart Weight (g)", main="Scatterplot")
abline(fit1$coefficients)
```



```r
summary(fit1)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5694 -0.9634 -0.0921  1.0426  5.1238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923  -0.515    0.607
## Bwt           4.0341     0.2503  16.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 142 degrees of freedom
## Multiple R-squared:  0.6466, Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

Let's do all the procedures mannually. We initially set our design matrix X and corresponding response matrix y

Why do we need combine a column of "1"?

```r
X=cbind(1,Bwt)
y=Hwt
```

We aim to minimize the residual sum of squares:

$$RSS(\beta) = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = (\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\beta)$$

Differentiating with respect to β:

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2\mathbf{X}^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\beta)$$

Setting to zero leads to the normal equation:

$$\mathbf{X}^{\mathbf{T}}\mathbf{X}\hat{\beta} = \mathbf{X}^{\mathbf{T}}\mathbf{y}$$

Since $X^T X$ is invertible:

$$\hat{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

```r
beta_hat = solve(t(X)%*%X)%*%t(X)%*%y
beta_hat
```

```
##           [,1]
##     -0.3566624
## Bwt  4.0340627
```

```
fit1$coefficients
```

```
## (Intercept)         Bwt
##  -0.3566624   4.0340627
```

```
y_hat = X%*%beta_hat
y_hat[1:10]
```

```
##  [1] 7.711463 7.711463 7.711463 8.114869 8.114869 8.114869 8.114869 8.114869 8.114869 8.114869
```

```
fit1$fitted.values[1:10]
```

```
##        1        2        3        4        5        6        7        8        9       10
## 7.711463 7.711463 7.711463 8.114869 8.114869 8.114869 8.114869 8.114869 8.114869 8.114869
```

Residuals (Sum of residuals = 0)

```
res = y-y_hat
res[1:10]
```

```
##  [1] -0.71146296 -0.31146296  1.78853704 -0.91486923 -0.81486923 -0.51486923 -0.01486923  0.08513077
##  [9]  0.18513077  0.38513077
```

```
fit1$residuals[1:10]
```

```
##           1           2           3           4           5           6           7           8
## -0.71146296 -0.31146296  1.78853704 -0.91486923 -0.81486923 -0.51486923 -0.01486923  0.08513077
##           9          10
##  0.18513077  0.38513077
```

```
sum(res)
```

```
## [1] -3.077538e-12
```

Residual Standard Error:

$$\sigma^2 = \frac{\sum(y_i - \hat{y}_i)^2}{df}$$

```
n = dim(X)[1]
p = dim(X)[2]-1
df = n-(p+1)
df
```

```
## [1] 142
```

```
Sig2 = sum(res^2)/df
sqrt(Sig2)
```

```
## [1] 1.452373
```

Residual Sum of Squares and R-square

R-square measures the goodness-of-fit

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS} = 1 - \frac{var(y_i - \hat{y}_i)}{var(y_i)} = \frac{var(\hat{y}_i)}{var(y_i)}$$

```
RSS = sum(res^2)
TSS = sum((y-mean(y))^2)
Rsquare = 1-RSS/TSS
Rsquare
```

```
## [1] 0.6466209
```

```
1-var(res)/var(y)
```

```
##           [,1]
## [1,] 0.6466209
```

```
var(y_hat)/var(y)
```

```
##           [,1]
## [1,] 0.6466209
```

Model Inference

$$E[\hat{\beta}] = (\mathbf{X^T X})^{-1}\mathbf{X^T E[y]} = (\mathbf{X^T X})^{-1}\mathbf{X^T X}\beta = \beta$$

$$Cov[\hat{\beta}] = Cov[(\mathbf{X^T X})^{-1}\mathbf{X^T y}] = (\mathbf{X^T X})^{-1}\mathbf{X^T Cov[y][(X^T X)^{-1}X^T]^T} = \sigma^2(X^T X)^{-1}$$

Thus, we have

$$\hat{\beta} \backsim \mathbf{N}(\beta, \sigma^2(\mathbf{X^T X})^{-1})$$

```
beta_var = diag(Sig2*solve(t(X)%*%X))
beta_std = sqrt(beta_var)
beta_std
```

```
##                 Bwt
## 0.6922770 0.2502615
```

t statistics and p-values

$H_0 : \beta_j = 0$ vs $H_0 : \beta_j \neq 0$

Test statistic:

$$t_j = \frac{\hat{\beta}_j}{SE(\beta_j)} \sim t_{n-p}$$

```
t_stat = beta_hat / beta_std
t_stat
```

```
##           [,1]
##     -0.5152019
## Bwt 16.1193908
```

```
2*pt(t_stat[1],df)
```

```
## [1] 0.6072131
```

```
2*pt(-t_stat[2],df)
```

```
## [1] 6.969045e-34
```

F statistic

```
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: Hwt
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Bwt         1 548.09  548.09  259.83 < 2.2e-16 ***
## Residuals 142 299.53    2.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSR = sum((y_hat-mean(y))^2)
SSE = sum((y-y_hat)^2)
SSR
```

```
## [1] 548.0924
```

```
SSE
```

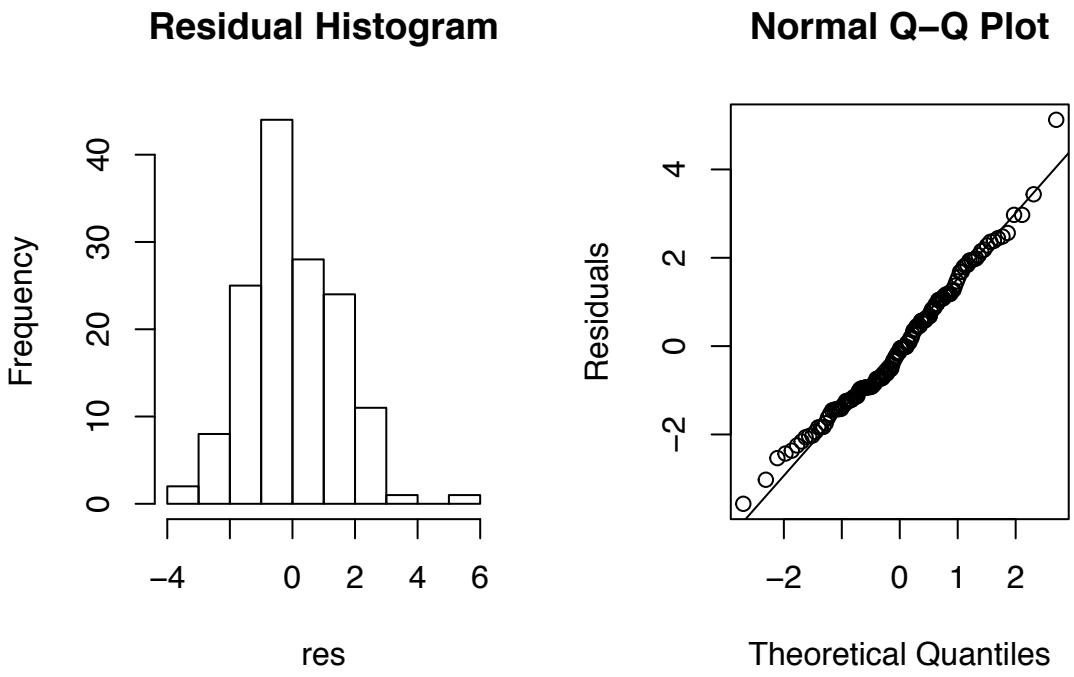
```
## [1] 299.5331
```

```
Fstat = (SSR/1)/(SSE/142)
Fstat
```

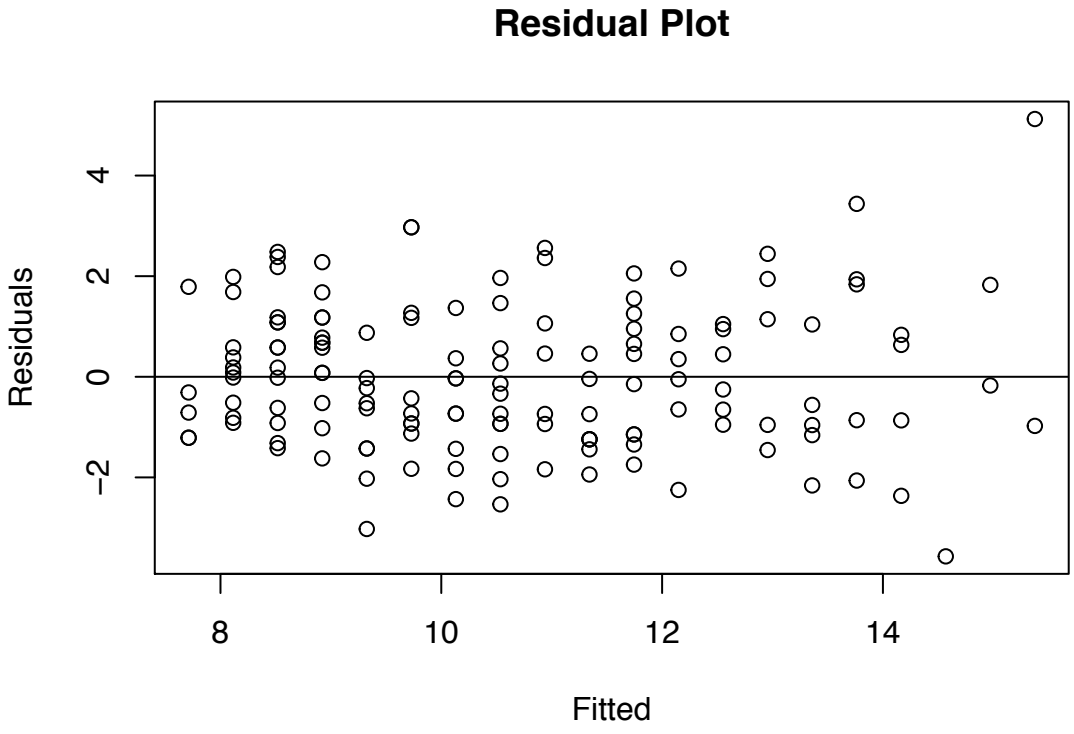
```
## [1] 259.8348
```

Regression Diagnostics

```
par(mfrow=c(1,2))
hist(res, main="Residual Histogram")
qqnorm(res, ylab="Residuals")
qqline(res)
```

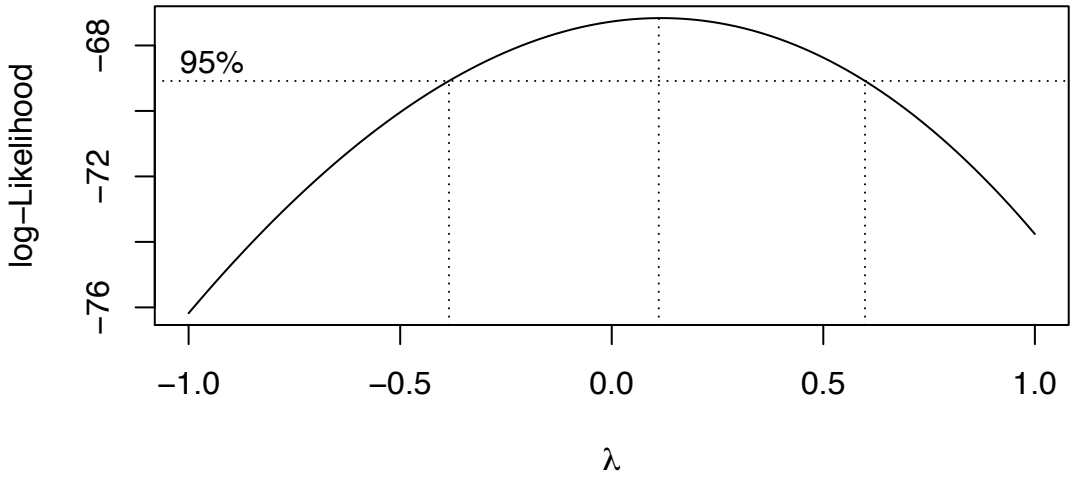



```
plot(y_hat, res, xlab="Fitted", ylab="Residuals", main="Residual Plot")
abline(h=0)
```



Find the optimal λ and the corresponding 95% confidence intervals

```
trans = boxcox(fit1, plotit=T, lambda=seq(-1,1 , by=0.05))
```



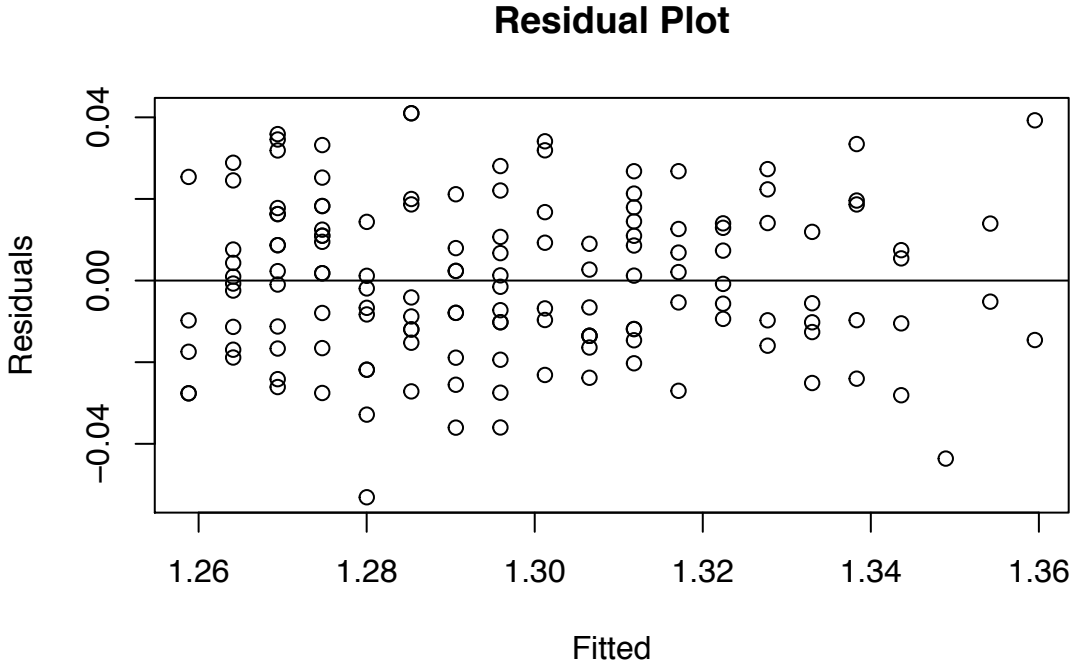

```
lambda = trans$x[trans$y == max(trans$y)]
tmp=trans$x[trans$y > max(trans$y) - qchisq(0.95, 1)/2]
lambda
```

```
## [1] 0.1111111
```

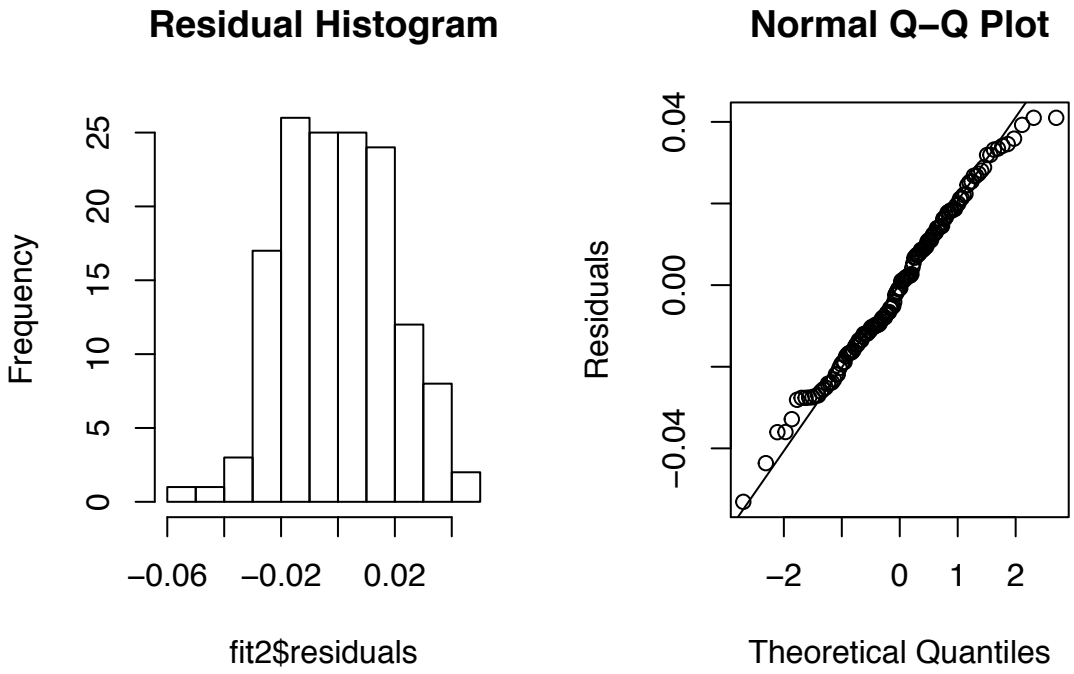
```
range(tmp)
```

```
## [1] -0.3737374  0.5959596
```

```
fit2 = lm(Hwt^lambda~Bwt,data=cats)
plot(fitted(fit2), residuals(fit2),
     xlab="Fitted", ylab="Residuals",
     main="Residual Plot")
abline(h=0)
```



```
par(mfrow=c(1,2))
hist(fit2$residuals, main="Residual Histogram")
qqnorm(fit2$residuals, ylab="Residuals")
qqline(fit2$residuals)
```



Variable Selection

First generate four new variables and combined into a new dataset

```
cats$Bwt2 = cats$Bwt^2
cats$Hwt2 = cats$Hwt^2
cats$Interact = cats$Hwt*cats$Bwt
cats$Ratio = cats$Hwt/cats$Bwt
head(cats)
```

```
##   Sex Bwt Hwt Bwt2  Hwt2 Interact    Ratio
## 1   F 2.0 7.0 4.00 49.00    14.00 3.500000
## 2   F 2.0 7.4 4.00 54.76    14.80 3.700000
## 3   F 2.0 9.5 4.00 90.25    19.00 4.750000
## 4   F 2.1 7.2 4.41 51.84    15.12 3.428571
## 5   F 2.1 7.3 4.41 53.29    15.33 3.476190
## 6   F 2.1 7.6 4.41 57.76    15.96 3.619048
```

Akaike information criterion (AIC) offers a relative estimate of the information lost when a given model is used to represent the process that generates the data. In doing so, it deals with the trade-off between the goodness of fit of the model and the complexity of the model.

$$AIC = -2log(L) + 2p$$

Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred.

$$BIC = -2log(L) + p * log(n)$$

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

Mallows's C_p is also used to address the overfitting of a regression model. It is applied in the context of model selection, where a number of predictor variables are available for predicting some outcome, and the goal is to find the best model involving a subset of these predictors. A small value of Cp means that the model is relatively precise.

$$Cp = \frac{SSE_p}{\hat{\sigma}^2} + 2p - n$$

The following is examples based on AIC. We use forward, backward and stepwise selection procedures to find our optimal model.

```
full = lm(Hwt~Bwt+Bwt2+Hwt2+Interact+Ratio,cats)
summary(full)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt + Bwt2 + Hwt2 + Interact + Ratio, data = cats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.088469 -0.011450 -0.003926  0.010108  0.086104
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.4732100  0.0767454 -71.316  < 2e-16 ***
## Bwt          3.9064433  0.0471254  82.895  < 2e-16 ***
## Bwt2        -0.7179196  0.0147177 -48.779  < 2e-16 ***
## Hwt2        -0.0029589  0.0007139  -4.145  5.9e-05 ***
## Interact     0.1954152  0.0060963  32.055  < 2e-16 ***
## Ratio        1.4013622  0.0102942 136.131  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02351 on 138 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 3.067e+05 on 5 and 138 DF,  p-value: < 2.2e-16
```

```r
null = lm(Hwt~1,cats)
step(null, scope = list(lower = null, upper = full), direction = "forward")
```

```
## Start:  AIC=257.26
## Hwt ~ 1
##
##            Df Sum of Sq    RSS     AIC
## + Hwt2      1    826.14  21.48 -269.98
## + Interact  1    781.58  66.04 -108.25
## + Bwt2      1    554.16 293.46  106.52
## + Bwt       1    548.09 299.53  109.47
## + Ratio     1    304.80 542.82  195.08
## <none>                  847.63  257.26
##
## Step:  AIC=-269.98
## Hwt ~ Hwt2
##
##            Df Sum of Sq    RSS     AIC
## + Bwt       1   1.83425 19.647 -280.83
## + Interact  1   1.66127 19.820 -279.57
## + Bwt2      1   1.06589 20.415 -275.31
## + Ratio     1   0.98606 20.495 -274.75
## <none>                  21.481 -269.98
##
## Step:  AIC=-280.83
## Hwt ~ Hwt2 + Bwt
##
##            Df Sum of Sq     RSS     AIC
## + Ratio     1   18.1792  1.4675 -652.42
## + Bwt2      1    4.8047 14.8420 -319.22
## <none>                  19.6467 -280.83
## + Interact  1    0.0874 19.5593 -279.48
##
## Step:  AIC=-652.42
## Hwt ~ Hwt2 + Bwt + Ratio
##
##            Df Sum of Sq     RSS     AIC
## + Bwt2      1   0.82330 0.64423 -768.97
## + Interact  1   0.07601 1.39151 -658.08
## <none>                  1.46752 -652.42
##
## Step:  AIC=-768.97
## Hwt ~ Hwt2 + Bwt + Ratio + Bwt2
##
##            Df Sum of Sq     RSS      AIC
## + Interact  1   0.56795 0.07628 -1074.22
## <none>                  0.64423  -768.97
##
## Step:  AIC=-1074.22
## Hwt ~ Hwt2 + Bwt + Ratio + Bwt2 + Interact
##
## Call:
## lm(formula = Hwt ~ Hwt2 + Bwt + Ratio + Bwt2 + Interact, data = cats)
```

```
##
## Coefficients:
## (Intercept)         Hwt2          Bwt        Ratio         Bwt2     Interact
##   -5.473210    -0.002959     3.906443     1.401362    -0.717920     0.195415
```

```
step(full, data = cats, direction = "backward")
```

```
## Start:  AIC=-1074.22
## Hwt ~ Bwt + Bwt2 + Hwt2 + Interact + Ratio
##
##            Df Sum of Sq     RSS      AIC
## <none>                  0.0763 -1074.22
## - Hwt2      1    0.0095  0.0858 -1059.32
## - Interact  1    0.5679  0.6442  -768.97
## - Bwt2      1    1.3152  1.3915  -658.08
## - Bwt       1    3.7982  3.8745  -510.62
## - Ratio     1   10.2434 10.3197  -369.55
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt + Bwt2 + Hwt2 + Interact + Ratio, data = cats)
##
## Coefficients:
## (Intercept)          Bwt         Bwt2         Hwt2     Interact        Ratio
##   -5.473210     3.906443    -0.717920    -0.002959     0.195415     1.401362
```

```
step(null, scope = list(upper = full), data = cats, direction = "both")
```

```
## Start:  AIC=257.26
## Hwt ~ 1
##
##            Df Sum of Sq    RSS     AIC
## + Hwt2      1    826.14  21.48 -269.98
## + Interact  1    781.58  66.04 -108.25
## + Bwt2      1    554.16 293.46  106.52
## + Bwt       1    548.09 299.53  109.47
## + Ratio     1    304.80 542.82  195.08
## <none>                  847.63  257.26
##
## Step:  AIC=-269.98
## Hwt ~ Hwt2
##
##            Df Sum of Sq    RSS     AIC
## + Bwt       1      1.83  19.65 -280.83
## + Interact  1      1.66  19.82 -279.57
## + Bwt2      1      1.07  20.42 -275.31
## + Ratio     1      0.99  20.49 -274.75
## <none>                    21.48 -269.98
## - Hwt2      1    826.14 847.63  257.26
##
## Step:  AIC=-280.83
## Hwt ~ Hwt2 + Bwt
##
##            Df Sum of Sq     RSS     AIC
## + Ratio     1    18.179   1.468 -652.42
## + Bwt2      1     4.805  14.842 -319.22
```

```
## <none>                   19.647 -280.83
## + Interact  1     0.087  19.559 -279.48
## - Bwt       1     1.834  21.481 -269.98
## - Hwt2      1   279.886 299.533  109.47
##
## Step:  AIC=-652.42
## Hwt ~ Hwt2 + Bwt + Ratio
##
##            Df Sum of Sq     RSS     AIC
## + Bwt2      1    0.8233  0.6442 -768.97
## + Interact  1    0.0760  1.3915 -658.08
## <none>                   1.4675 -652.42
## - Hwt2      1    9.1756 10.6431 -369.11
## - Ratio     1   18.1792 19.6467 -280.83
## - Bwt       1   19.0274 20.4949 -274.75
##
## Step:  AIC=-768.97
## Hwt ~ Hwt2 + Bwt + Ratio + Bwt2
##
##            Df Sum of Sq     RSS      AIC
## + Interact  1    0.5679  0.0763 -1074.22
## <none>                   0.6442  -768.97
## - Bwt2      1    0.8233  1.4675  -652.42
## - Bwt       1    4.3262  4.9705  -476.75
## - Hwt2      1    9.8793 10.5235  -368.73
## - Ratio     1   14.1978 14.8420  -319.22
##
## Step:  AIC=-1074.22
## Hwt ~ Hwt2 + Bwt + Ratio + Bwt2 + Interact
##
##            Df Sum of Sq     RSS      AIC
## <none>                   0.0763 -1074.22
## - Hwt2      1    0.0095  0.0858 -1059.32
## - Interact  1    0.5679  0.6442  -768.97
## - Bwt2      1    1.3152  1.3915  -658.08
## - Bwt       1    3.7982  3.8745  -510.62
## - Ratio     1   10.2434 10.3197  -369.55
##
## Call:
## lm(formula = Hwt ~ Hwt2 + Bwt + Ratio + Bwt2 + Interact, data = cats)
##
## Coefficients:
## (Intercept)         Hwt2          Bwt        Ratio         Bwt2     Interact
##   -5.473210    -0.002959     3.906443     1.401362    -0.717920     0.195415
```