Linear and Robust Regression of the Cats Dataset from the MASS Package

Jon Kinsey

Sat Jan 3 14:16:28 2015

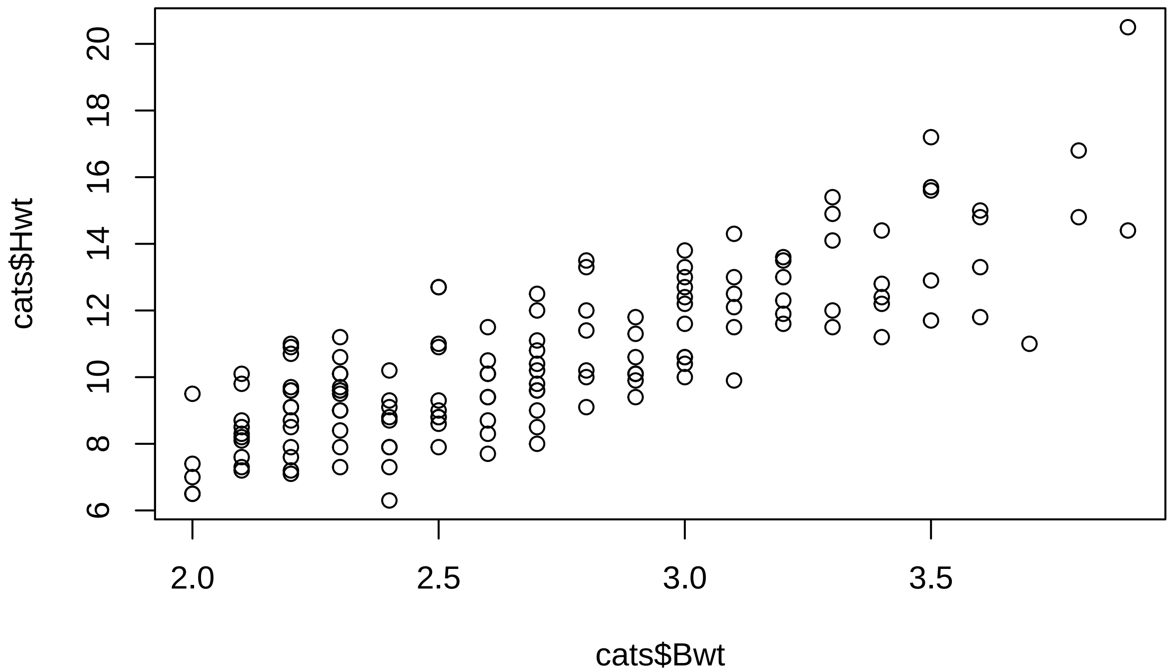
```r
# The heart and body weights of samples of male and female cats used for
# digitalis experiments. The cats were all adult, over 2 kg body weight.
# This data frame contains the following columns:
# Sex : Factor with evels "F" and "M".
# Bwt : body weight in kg.
# Hwt : heart weight in g.
# Reference:
# R. A. Fisher (1947) The analysis of covariance method for the relation between
# a part and the whole, Biometrics 3, 65-68.
#
library(MASS)
data(cats)
str(cats)
```

```
## 'data.frame':    144 obs. of  3 variables:
##  $ Sex: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Bwt: num  2 2 2 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
##  $ Hwt: num  7 7.4 9.5 7.2 7.3 7.6 8.1 8.2 8.3 8.5 ...
```

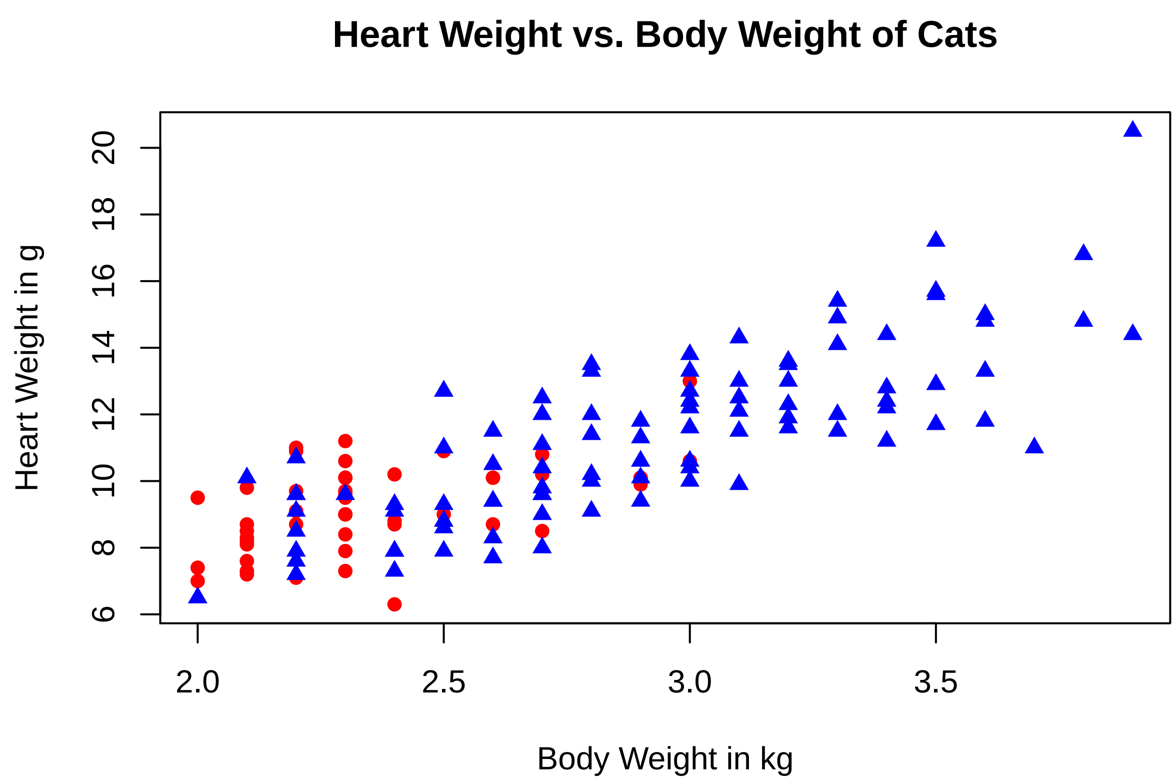
```r
summary(cats)
```

```
##  Sex         Bwt             Hwt
##  F:47   Min.   :2.000   Min.   : 6.30
##  M:97   1st Qu.:2.300   1st Qu.: 8.95
##         Median :2.700   Median :10.10
##         Mean   :2.724   Mean   :10.63
##         3rd Qu.:3.025   3rd Qu.:12.12
##         Max.   :3.900   Max.   :20.50
```

```r
#
plot(cats$Bwt, cats$Hwt)
```




```r
# Here's a more detailed plot separating males and females :
with(cats, plot(Bwt, Hwt, type="n", xlab="Body Weight in kg",
                ylab="Heart Weight in g",
                main="Heart Weight vs. Body Weight of Cats"))
with(cats,points(Bwt[Sex=="F"],Hwt[Sex=="F"],pch=16,col="red"))
with(cats,points(Bwt[Sex=="M"],Hwt[Sex=="M"],pch=17,col="blue"))
```



```r
# A Pearson product-moment correlation coefficient can be calculated using
# the cor( ) function
with(cats, cor(Bwt, Hwt))
```

```
## [1] 0.8041274
```

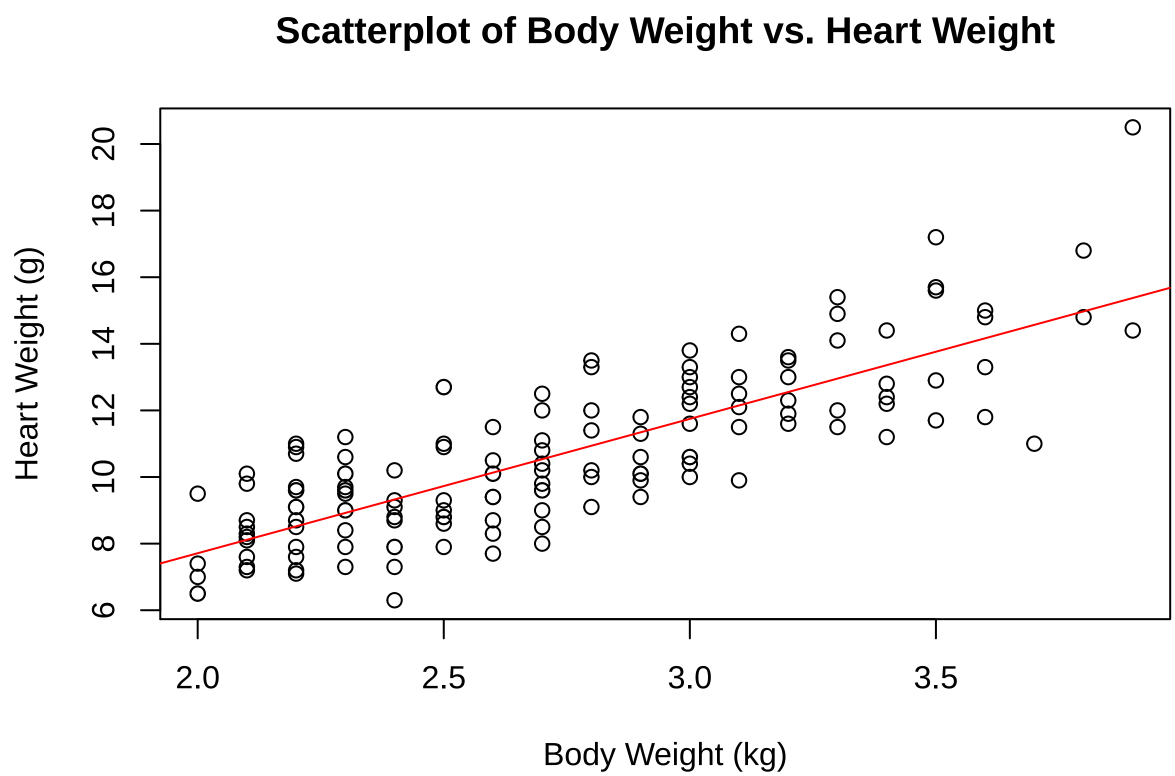
```r
# Pearson's r = .804 indicates a strong positive relationship.
#
# The linear regression fit coefficients are:
lm(Hwt ~ Bwt, data=cats)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats)
##
## Coefficients:
## (Intercept)          Bwt
##     -0.3567       4.0341
```

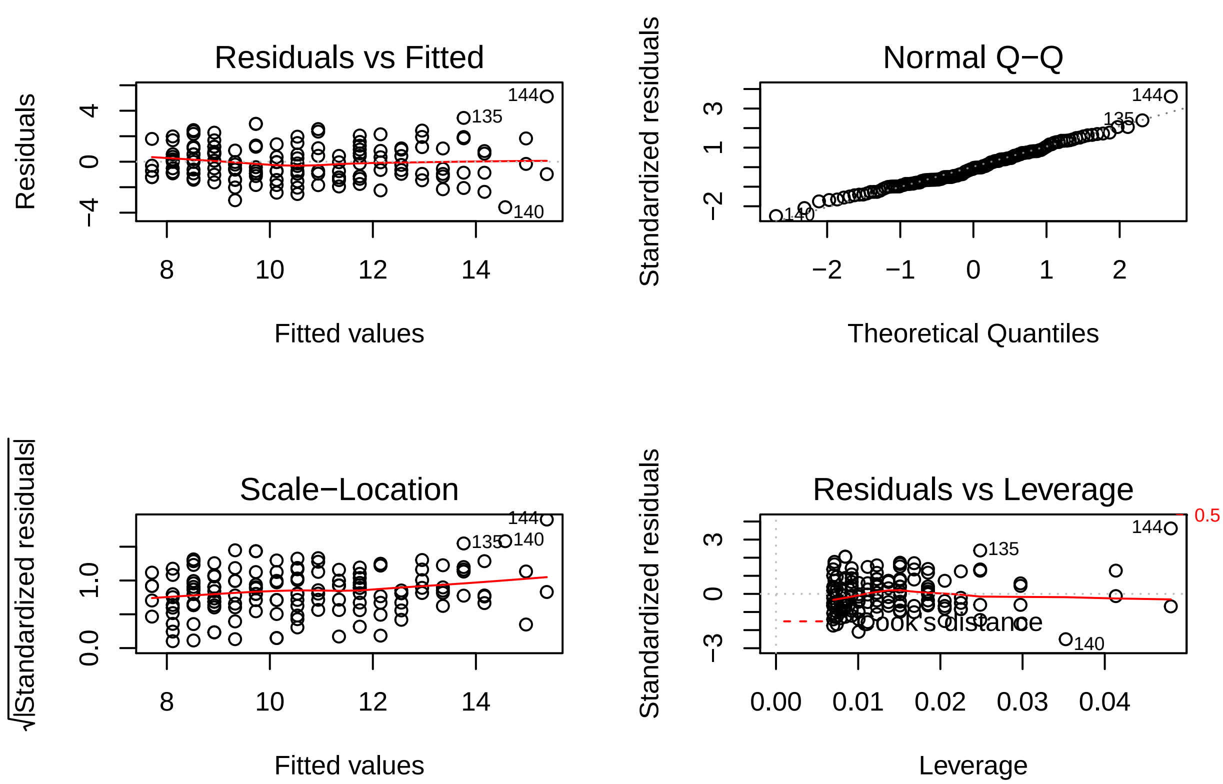
```r
# So the fitted regression equation is Hwt = 4.0341 (Bwt) - 0.3567.
lmfit <- lm(Hwt ~ Bwt, data=cats)
summary(lmfit)
```

```
##
## 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
```

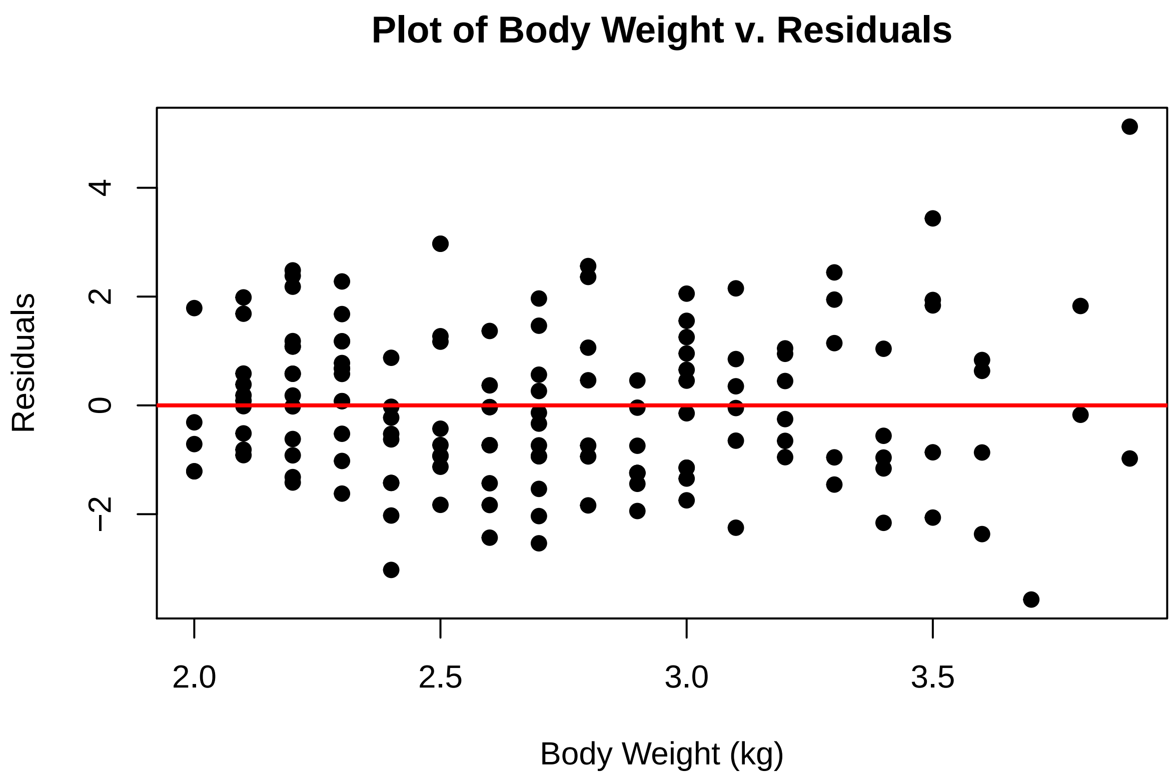
```r
# The three stars for Bwt indicates that the significance of the Bwt
# coefficient is between 0 and 0.001.
#
# Next, we plot the regression line on a scatterplot with the data
plot(cats$Hwt ~ cats$Bwt,
     xlab = "Body Weight (kg)", ylab = "Heart Weight (g)",
     main="Scatterplot of Body Weight vs. Heart Weight")
abline(lmfit, col="red")
```



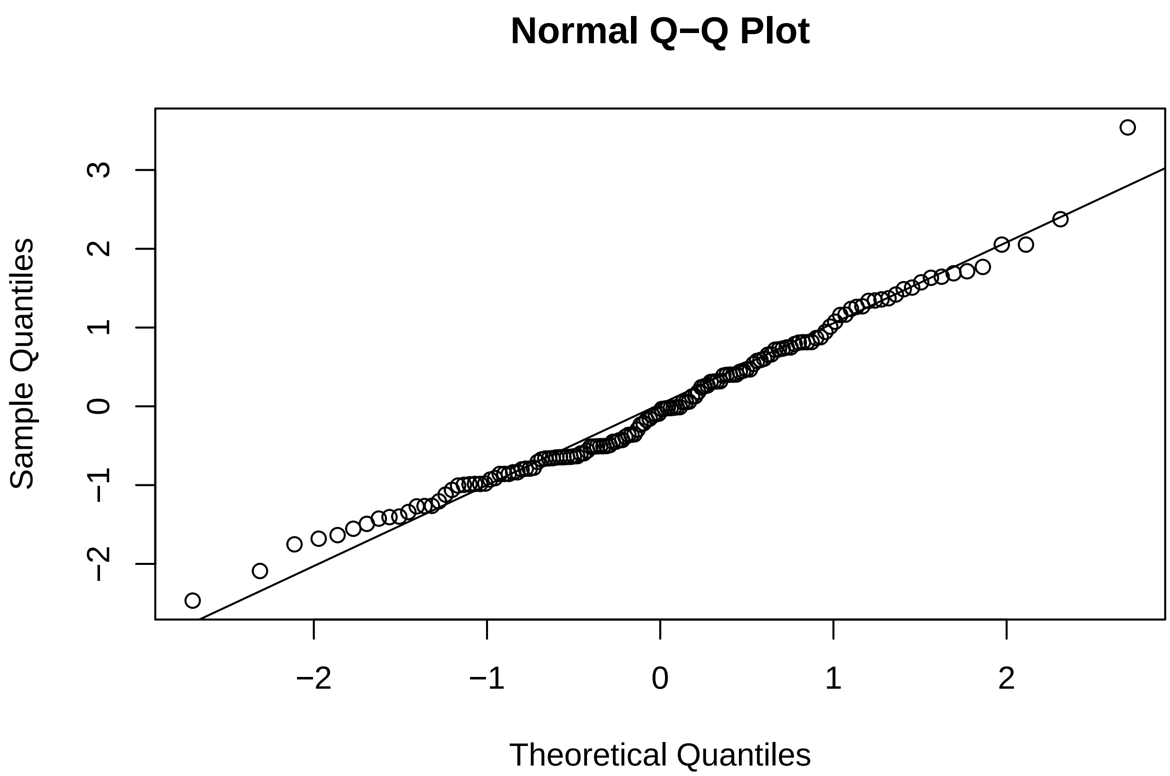

```
#
par(mfrow=c(2,2)) # configure output window for 4 plots, 2x2
plot(lmfit)
```



```r
#
# plot of the residuals
par(mfrow=c(1,1))
res = lmfit$residuals
plot(cats$Bwt, res, pch = 19, xlab = "Body Weight (kg)", ylab = "Residuals",
     main = "Plot of Body Weight v. Residuals")
abline(h = 0, col = "red", lwd = 2)
```




```r
#
res = scale(res)
# Let's replot the Q-Q plot
# Note: An ideal Q-Q plot has points falling more or less on the diagonal line,
# indicating that our residuals are approximately normally distributed.
qqnorm(res)
qqline(res)
```




```
#
# find the expected heart weight for a cat that weighs 3 kg
newObs <- data.frame(Bwt=3)
predict(lmfit, newObs, interval="predict")
```

```
##        fit      lwr      upr
## 1 11.74553 8.861263 14.62979
```

```
# going back to the heart-body plot we see that 11.75 gms for a 3kg cat
# looks about right
#
# Outlier Analysis
# Now, lets look at case 144 which appears to be an outlier
cats[144,]
```

```
##     Sex Bwt  Hwt
## 144   M 3.9 20.5
```

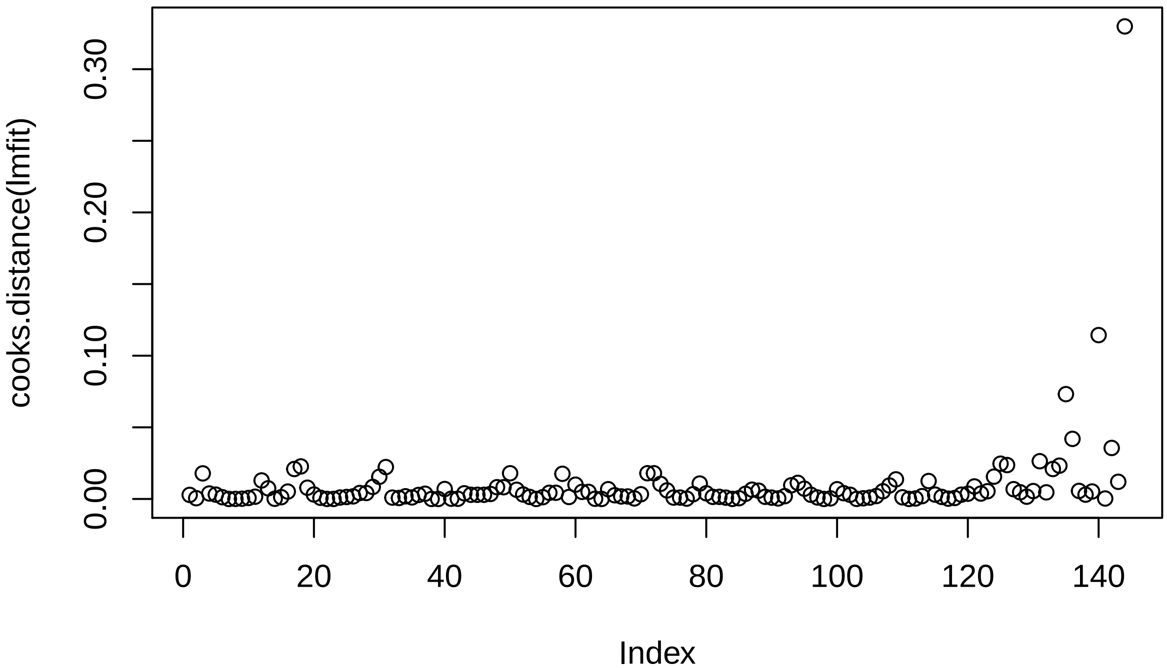
```
lmfit$fitted[144]
```

```
##      144
## 15.37618
```

```
lmfit$residuals[144]
```

```
##      144
## 5.123818
```

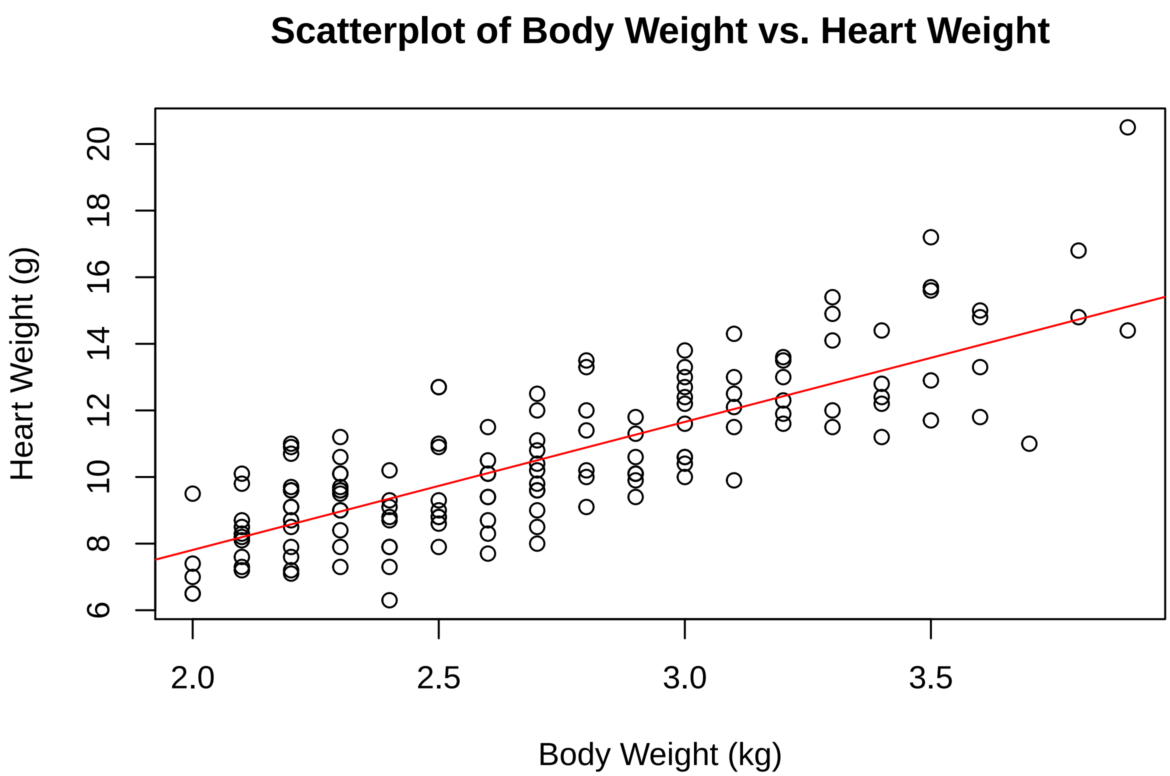
```
# The observed value of the heart weight was 20.5gm for a 3.9kg cat, but the
# fitted value was only 15.38gm, giving a residual of 5.12gm.
# The residual standard error (from the summary above) was 1.452, so converting
# this to a standardized residual gives 5.124/1.452=3.53. This is a substantial value.
# A commonly used measure of influence is Cook's Distance, which can be visualized
# for all the cases in the model as follows...
par(mfrow=c(1,1))  # reset graphics
plot(cooks.distance(lmfit))
```




```
# The plots shows that case 144 appears much higher than the other cases.
# There are a number of ways to procede. One is to look at the regression
# coefficients without the outlying point in the model...
lmfit.without144 = lm(cats$Hwt ~ cats$Bwt, subset=(cats$Hwt<20.5))
lmfit.without144
```

```
##
## Call:
## lm(formula = cats$Hwt ~ cats$Bwt, subset = (cats$Hwt < 20.5))
##
## Coefficients:
## (Intercept)     cats$Bwt
##       0.118        3.846
```

```
# This fit gives us : Hwt = 3.846 (Bwt) + 0.118.
plot(cats$Hwt ~ cats$Bwt,
     xlab = "Body Weight (kg)", ylab = "Heart Weight (g)",
     main="Scatterplot of Body Weight vs. Heart Weight")
abline(lmfit.without144, col="red")
```

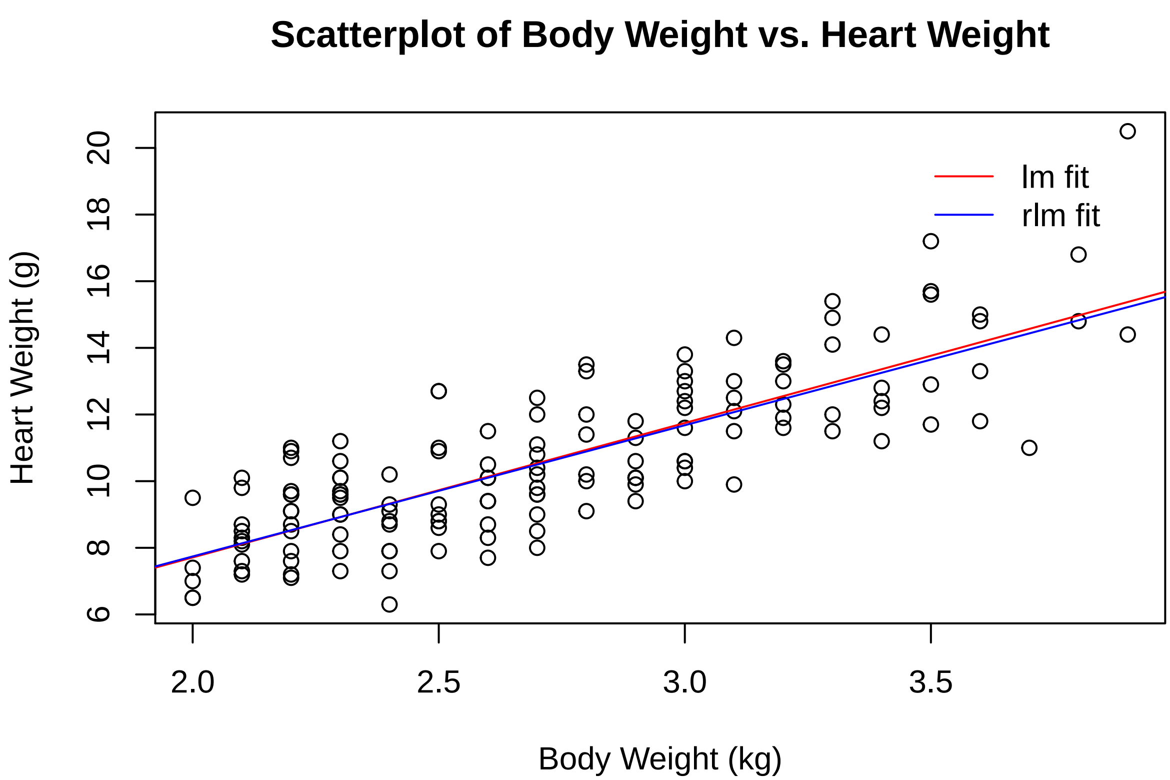



```
#
# Robust Regression
# Another is to use the rlm procedure in MASS that is robust in the face of
# outlying points. This robust regression is an alternative to least squares
# regression when data are contaminated with outliers or influential observations.
# Here, fitting is done by iterated re-weighted least squares (IWLS) while
# lm() uses ordinary least squares.
rlmfit <- rlm(cats$Hwt ~ cats$Bwt)
rlmfit
```

```
## Call:
## rlm(formula = cats$Hwt ~ cats$Bwt)
## Converged in 5 iterations
##
## Coefficients:
## (Intercept)    cats$Bwt
##  -0.1361777   3.9380535
##
## Degrees of freedom: 144 total; 142 residual
## Scale estimate: 1.52
```

```
# This fit gives us : Hwt = 3.938 (Bwt) - 0.1362.
# Now, lets replot the data with the lm and rlm fits to compare
plot(cats$Hwt ~ cats$Bwt,
     xlab = "Body Weight (kg)", ylab = "Heart Weight (g)",
     main="Scatterplot of Body Weight vs. Heart Weight")
abline(lmfit, col="red")
abline(rlmfit, col="blue")
legend("topright", inset=0.05, bty="n",
```

```
       legend = c("lm fit","rlm fit"),
       lty = c(1, 1),       # 1 = "solid" ; 2 = "dashed"
       col = c("red", "blue")  )
```




```
# Only a slight difference is evident in the fits
```