Linear Regression Analysis of the Women Dataset

Jon Kinsey

Wed Dec 31 12:54:13 2014

```
# This data set gives the average heights and weights for American
# women aged 30-39. There are 15 observations and 2 variables: height, weight
data(women)
str(women)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ height: num  58 59 60 61 62 63 64 65 66 67 ...
##  $ weight: num  115 117 120 123 126 129 132 135 139 142 ...
```

```
summary(women)
```

```
##      height         weight
##  Min.   :58.0   Min.   :115.0
##  1st Qu.:61.5   1st Qu.:124.5
##  Median :65.0   Median :135.0
##  Mean   :65.0   Mean   :136.7
##  3rd Qu.:68.5   3rd Qu.:148.0
##  Max.   :72.0   Max.   :164.0
```

```
#
fit <- lm(weight ~ height, data=women)
summary(fit)
```

```
##
## Call:
## lm(formula = weight ~ height, data = women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7333 -1.1333 -0.3833  0.7417  3.1167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
## height        3.45000    0.09114   37.85 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.525 on 13 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9903
## F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

```
#
# From the output, we see that the prediction equation is
# Weight = -87.52 + 3.45*Height
```

```r
# From the Pr(>|t|) column, we see that the regression coefficient (3.45)
# is significantly different from zero (p < 0.001) and indicates that there's
# an expected increase of 3.45 pounds of weight for every 1 inch increase in height.
# The multiple R-squared (0.991) indicates that the model accounts for 99.1 percent
# of the variance in weights. The residual standard error (1.53 lbs.) can be
# thought of as the average error in predicting weight from height using this model.
#
women$weight
```

```
##  [1] 115 117 120 123 126 129 132 135 139 142 146 150 154 159 164
```

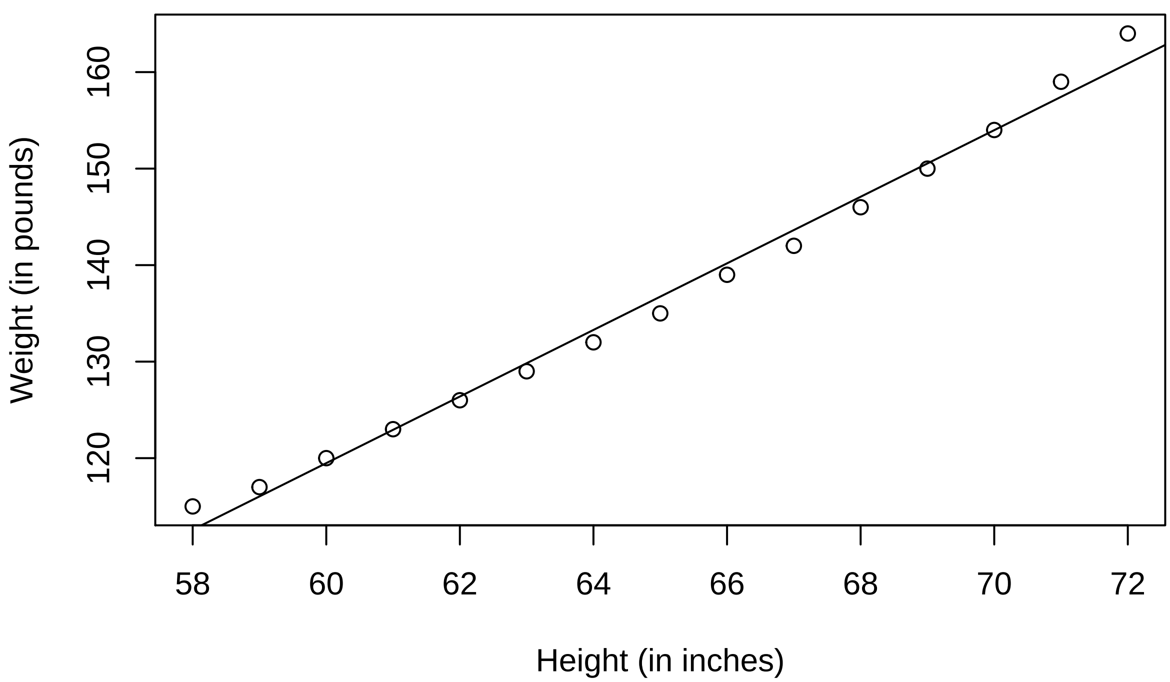
```r
fitted(fit)
```

```
##        1        2        3        4        5        6        7        8
## 112.5833 116.0333 119.4833 122.9333 126.3833 129.8333 133.2833 136.7333
##        9       10       11       12       13       14       15
## 140.1833 143.6333 147.0833 150.5333 153.9833 157.4333 160.8833
```

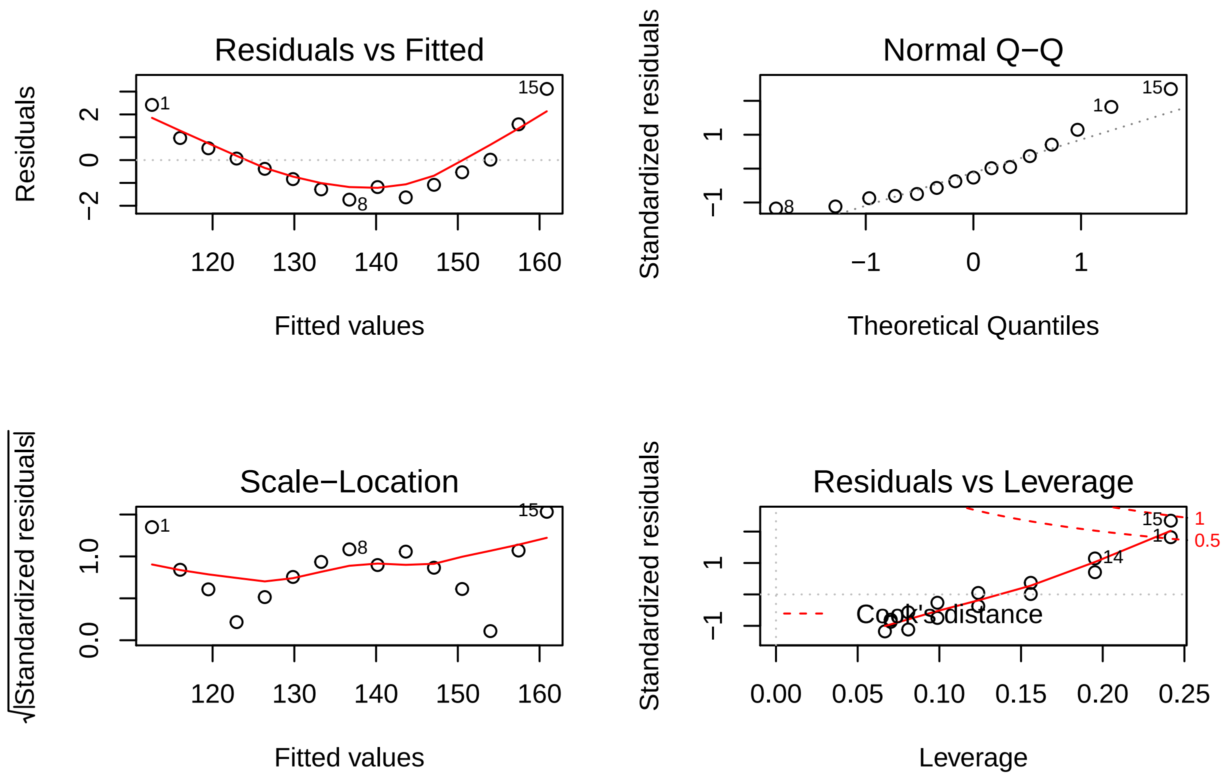
```r
residuals(fit)
```

```
##           1           2           3           4           5           6
##  2.41666667  0.96666667  0.51666667  0.06666667 -0.38333333 -0.83333333
##           7           8           9          10          11          12
## -1.28333333 -1.73333333 -1.18333333 -1.63333333 -1.08333333 -0.53333333
##          13          14          15
##  0.01666667  1.56666667  3.11666667
```

```r
#
plot(women$height,women$weight,
     xlab="Height (in inches)",
     ylab="Weight (in pounds)")
abline(fit)
```



```r
par(mfrow=c(2,2))
# The plot shows that the largest residuals occur for low and high heights.
# This suggests we might be able to improve your prediction using a regression
# with a quadratic term (that is, X**2).
plot(fit)
```

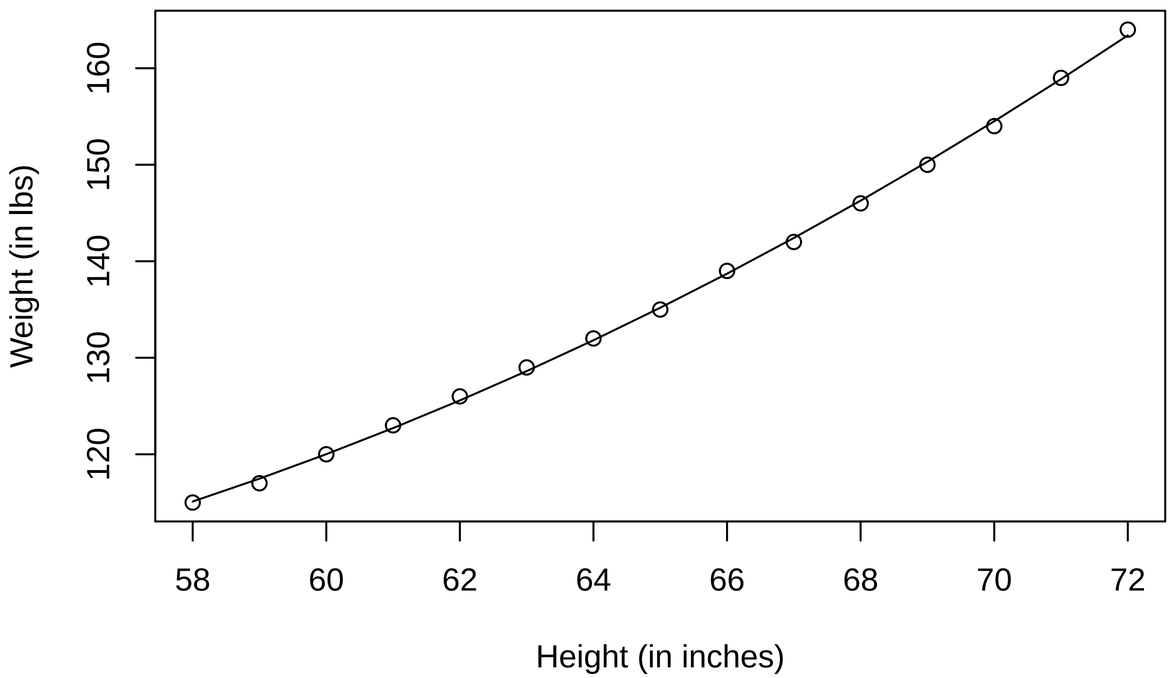



```r
par(mfrow=c(1,1))     # reset the graphics defaults
#
# Lets try a polynomial fit now.
#
fit2 <- lm(weight ~ height + I(height^2), data=women)
# Note: The I function treats the contents within the parentheses as an R regular
# expression. You need this because the ^ operator has a special meaning in
# formulas that we don't want to invoke here.
#
summary(fit2)
```

```
##
## Call:
## lm(formula = weight ~ height + I(height^2), data = women)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50941 -0.29611 -0.00941  0.28615  0.59706
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 261.87818   25.19677  10.393 2.36e-07 ***
```

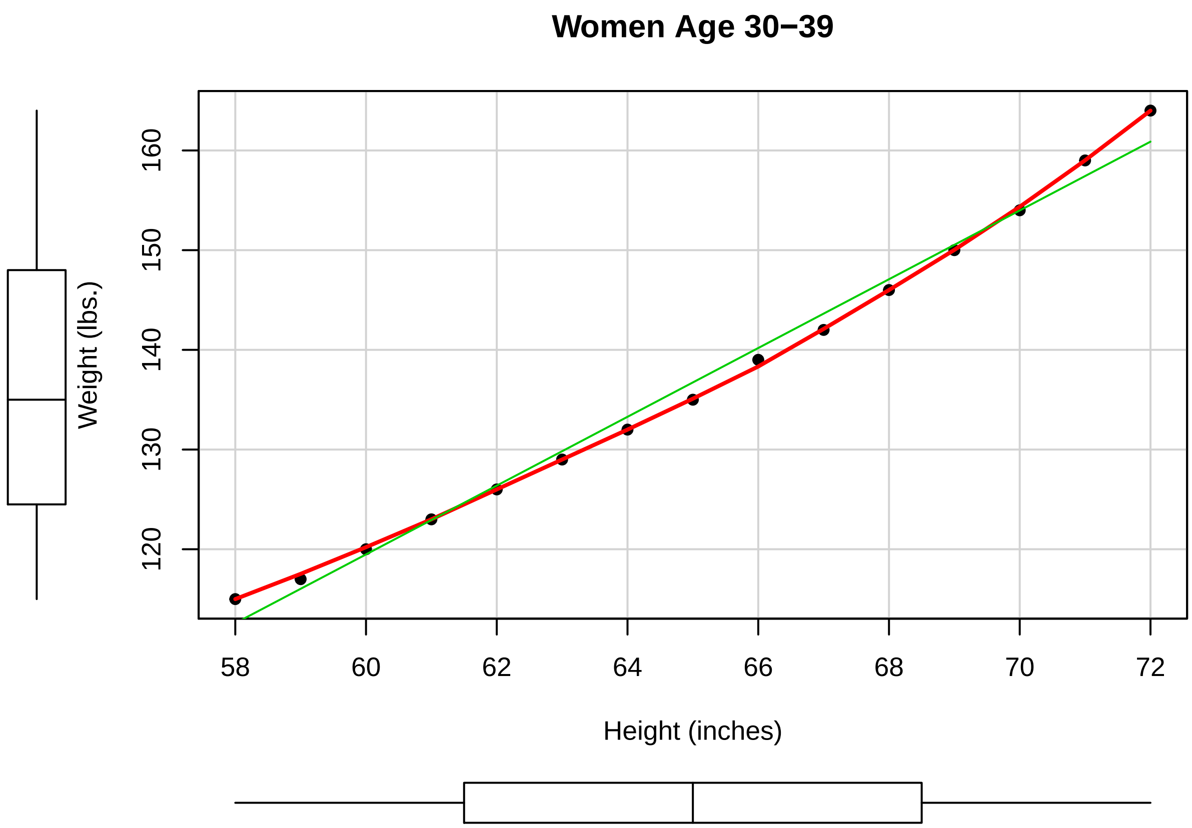
```
## height       -7.34832    0.77769  -9.449 6.58e-07 ***
## I(height^2)   0.08306    0.00598  13.891 9.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3841 on 12 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic: 1.139e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

```
# This gives us: Weight = 261.88 - 7.35*Height + 0.0833*Height**2
# The significance of the squared term (t = 13.89, p < .001) suggests that
# inclusion of the quadratic term improves the model fit. This is evident
# when we plot the fit.
plot(women$height,women$weight,
     xlab="Height (in inches)",
     ylab="Weight (in lbs)")
lines(women$height,fitted(fit2))
```




```
#
# Lets try a cubic equation
fit3 <- lm(weight ~ height + I(height^2) +I(height^3), data=women)
#
# the scatterplot() function in the car package provides a simple and convenient
# method of plotting a bivariate relationship.
library(car)
scatterplot(weight ~ height,
            data=women,
            spread=FALSE, lty.smooth=2,
            pch=19,
            main="Women Age 30-39",
            xlab="Height (inches)",
            ylab="Weight (lbs.)")
```

```
## Warning in plot.window(...): "lty.smooth" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "lty.smooth" is not a graphical
## parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "lty.smooth"
## is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "lty.smooth"
## is not a graphical parameter

## Warning in box(...): "lty.smooth" is not a graphical parameter

## Warning in title(...): "lty.smooth" is not a graphical parameter
```




```
# This enhanced plot provides the scatter plot of weight with height, box plots
# for each variable in their respective margins, the linear line of best fit,
# and a smoothed (loess) fit line. The spread=FALSE options suppress spread and
# asymmetry information. The lty.smooth=2 option specifies that the loess fit
# be rendered as a dashed line. The pch=19 options display points as filled
# circles (the default is open circles). You can tell at a glance that the two
# variables are roughly symmetrical and that a curved line will fit the data
# points better than a straight line.
```