More About OLS Regression (Part II)

Simple Linear Regression

Let us understand the `lm()` function and the other functions that are useful when fitting linear models with the help of a simple regression example.

The dataset `women` in the base installation of R provides the height and weight for a set of 15 women, aged 30 to 39. Suppose we want to predict weight from height. Having an equation for predicting weight for height can also help us identify overweight or underweight women.

```
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 of `lm()` above, we see that prediction equation is

$$\hat{weight} = -87.52 + 3.45 \times Height$$

Here is the actual weight data of the women in the sample

```
women$weight
```

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

The `fitted()` function lists the predicted values from the fitted model, based on the OLS regression.

```
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
```

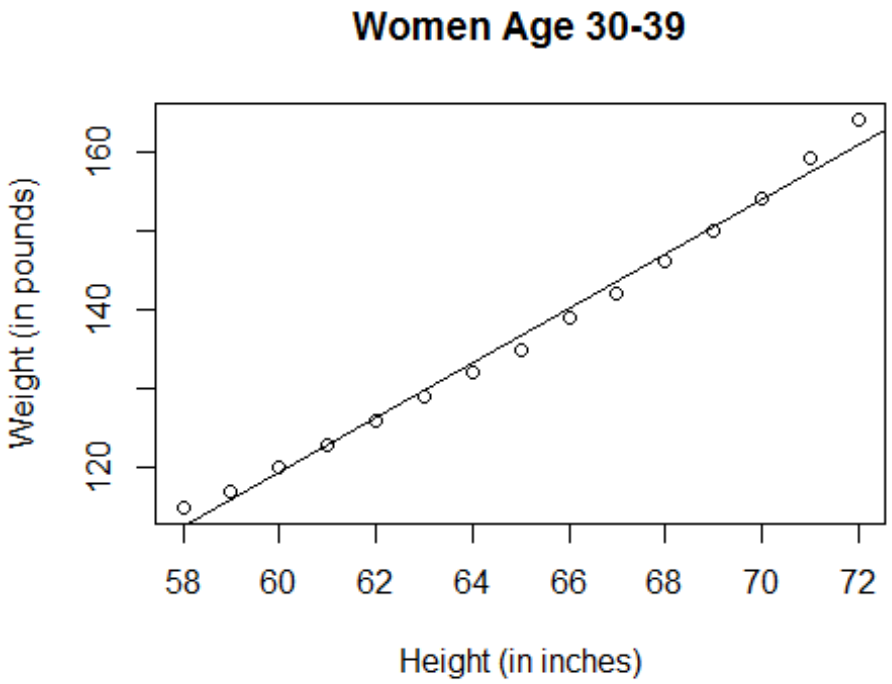
The `residuals()` function gives the residual values from the fitted model, based on OLS regression

```
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
```

And here is a scatter plot with the regression line for weight predicted from height, based on the above OLS model.

```
plot(women$height,women$weight,
     main="Women Age 30-39",
     xlab="Height (in inches)",
     ylab="Weight (in pounds)")
# add the line of best fit
abline(fit)
```



Interpretation

Recall from the `lm()` output above, that the prediction equation is

$$\hat{weight} = -87.52 + 3.45 \times Height$$

- Since a height of `0` is impossible, we cannot give a physical interpretation to the intercept. It merely becomes an adjustment constant.
- The regression coefficients (3.45) is significantly different from zero, ($p < 0.001$) and indicates that there is an expected increase of 3.45 pounds of weight for every one inch increase in height
- The multiple R-squared (0.991) indicates that the model accounts for 99.1% of the variance in weights.

- The residual standard error (1.53 pounds) can be thought of as the average error in predicting weight from height using this model.
- The F statistic tests whether the predictor variables, taken together, predict the response variable above chance levels. (Since there is only one predictor variable in simple regression, in this example, the F test is equivalent to the t-test for teh regresion coefficient for height.)

For demonstration purposes, we have printed out the acutal, predicted and residual values above. Evidently the largest residuals occur for low and high heights, which can also be seen in the above scatter plot.