Homework2

Rohan Sadale

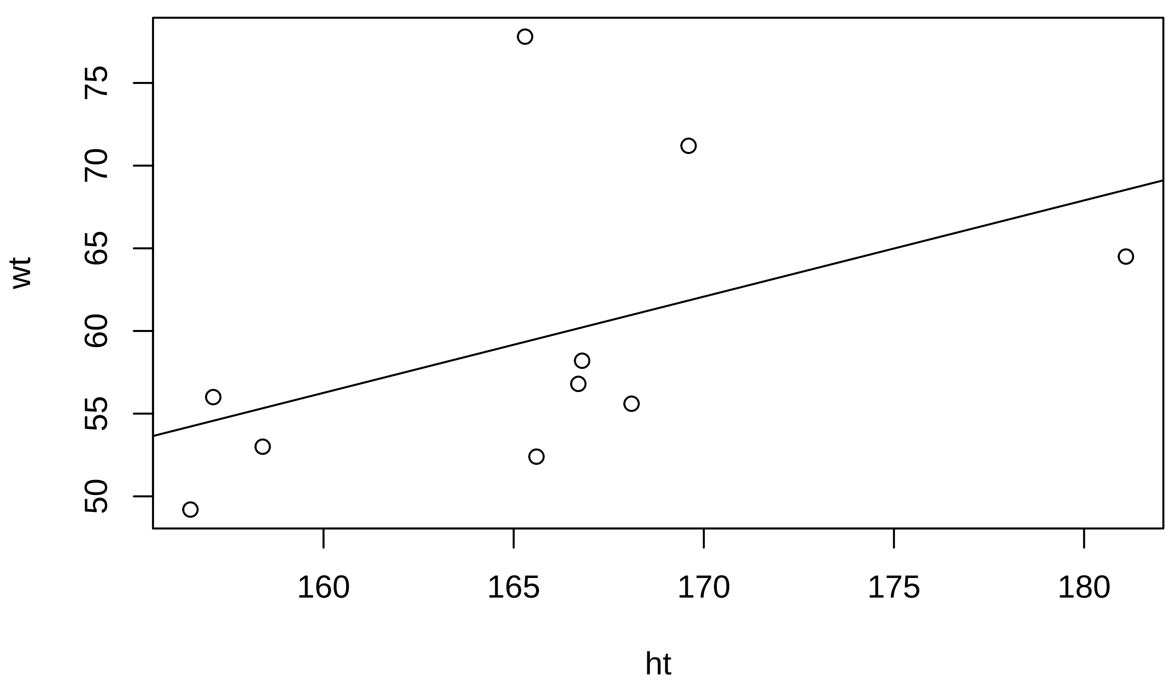
Feb 5, 2016

```
## Warning: package 'alr4' was built under R version 3.2.3

## Warning: package 'car' was built under R version 3.2.3

## Warning: package 'effects' was built under R version 3.2.3
```

2.1

- 2.1.1

```r
with(Htwt, plot(wt~ht))
abline(lm(wt~ht, data = Htwt))
```



Although with so less data (10 points) it is hard to decide the plausibility of Simple Linear Regression, we can see that with increasing height, the weight is increasing. Thus a simple linear regression model is plausible for this data.

- 2.1.2

```r
x_bar <- mean(Htwt$ht)
print(x_bar)
```

```
## [1] 165.52
```

```r
y_bar <- mean(Htwt$wt)
print(y_bar)
```

```
## [1] 59.47
```

```r
SXX <- sum((Htwt$ht - x_bar)**2)
print(SXX)
```

```
## [1] 472.076
```

```r
SYY <- sum((Htwt$wt - y_bar)**2)
print(SYY)
```

```
## [1] 731.961
```

```r
SXY <- sum((Htwt$ht - x_bar) * (Htwt$wt - y_bar))
print(SXY)
```

```
## [1] 274.786
```

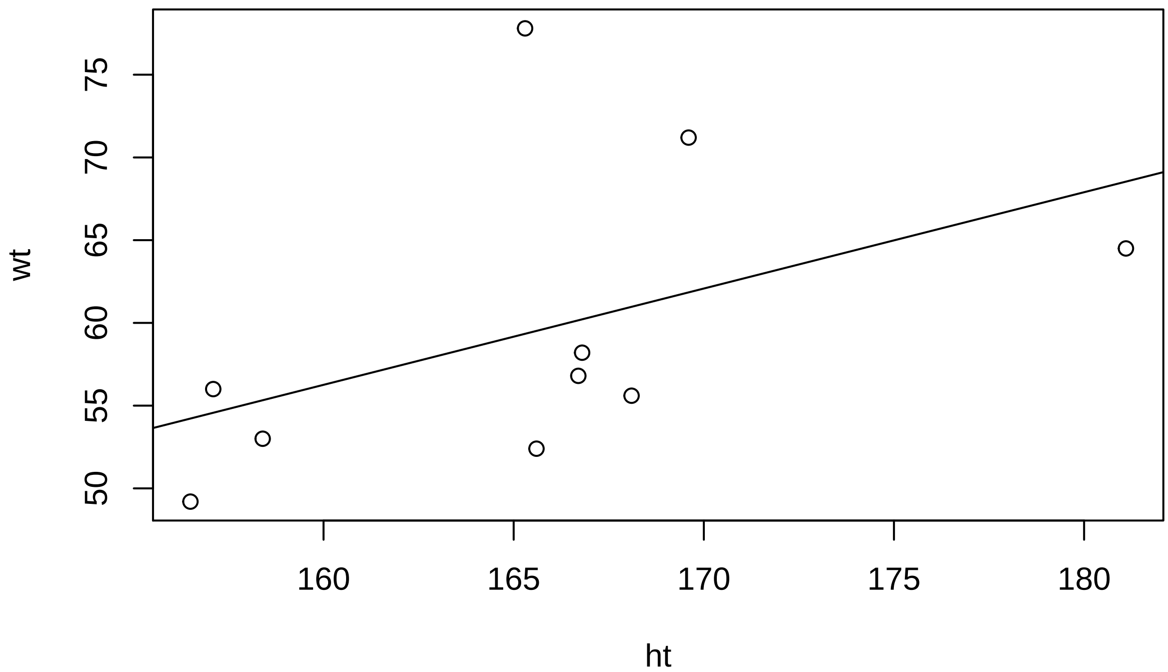
```r
slope <- SXY/SXX
print(slope)
```

```
## [1] 0.58208
```

```r
intercept <- y_bar - slope*x_bar
print(intercept)
```

```
## [1] -36.87588
```

```r
with(Htwt, plot(wt~ht))
abline(intercept, slope)
```



+ 2.1.3

```
RSS <- SYY - SXY^2/SXX
print(RSS)
```

```
## [1] 572.0136
```

```
variance <- RSS/(nrow(Htwt) - 2)

#estimated variance
print(variance)
```

```
## [1] 71.5017
```

```
slope <- SXY/SXX
print(slope)
```

```
## [1] 0.58208
```

```
intercept <- y_bar - slope*x_bar
print(intercept)
```

```
## [1] -36.87588
```

```r
se_intercept <- sqrt(variance * (1/nrow(Htwt) + x_bar^2/SXX))
print(se_intercept)
```

```
## [1] 64.4728
```

```r
se_slope <- sqrt(variance/SXX)
print(se_slope)
```

```
## [1] 0.3891815
```

```r
est_covariance <- -(variance)*x_bar/SXX
print(est_covariance)
```

```
## [1] -25.07003
```

```r
tval_beta0 <- intercept/se_intercept
print(tval_beta0)
```

```
## [1] -0.5719603
```

```r
tval_beta1 <- slope/se_slope
print(tval_beta1)
```

```
## [1] 1.495652
```

```r
p_values <- 2*pt(-abs(c(tval_beta0, tval_beta1)), 8)
print(p_values)
```

```
## [1] 0.5830589 0.1731089
```

2.2

- 2.2.1

The points above the line y=x suggest that rice price in 2009 was higher as compared to price in 2003 for the countries marked above the line.

The points below the line y=x suggest that rice price in 2009 was lower as compared to price in 2003 for the countries marked below the line.

- 2.2.2

Vilnius had the largest increase in rice price. Mumbai had largest descrease in rice price.

- 2.2.3

Yes, it does suggest that prices are lower in 2009. This is because the rate of increase in per unit of price (slope/beta1) is less for OLS than that of line y=x where beta1/slope is 1.

- 2.2.4

Fitting simple linear regression line is not suitable for the data because- a) There are some countries such as Vilnius and Budapest which can be considered as outliers and their presence can affect OLS significantly. b) The data distribution appears to be skewed. c) Also the presence of countries such as Mumbai (leverage points) can significantly affect the simple linear regression line and hence slope of line might change.

2.3

- 2.3.1 Log-scale transformation is preferable because

a) The distribution of points isn't skewed and the variability appears to be constant.
b) Points on the extreme right such as Mumbai no longer appear to be high leverage point.
c) Though countries Vilnius and Budapest still look outliers, they appear to be slighly close to the other data values and hence don't affect the simple linear regression line significantly.

- 2.3.2

$$log(E(y|x)) \approx E(log(y)|x) = \beta_0 + \beta_1 log(x)$$

Thus

$$E(y|x) = e^{\beta_0 + \beta_1 log(x)} = e^{\beta_0} e^{\beta_1 log(x)} = e^{log(\gamma_0)} (e^{log(x)})^{\beta_1} = \gamma_0 x^{\beta_1}$$

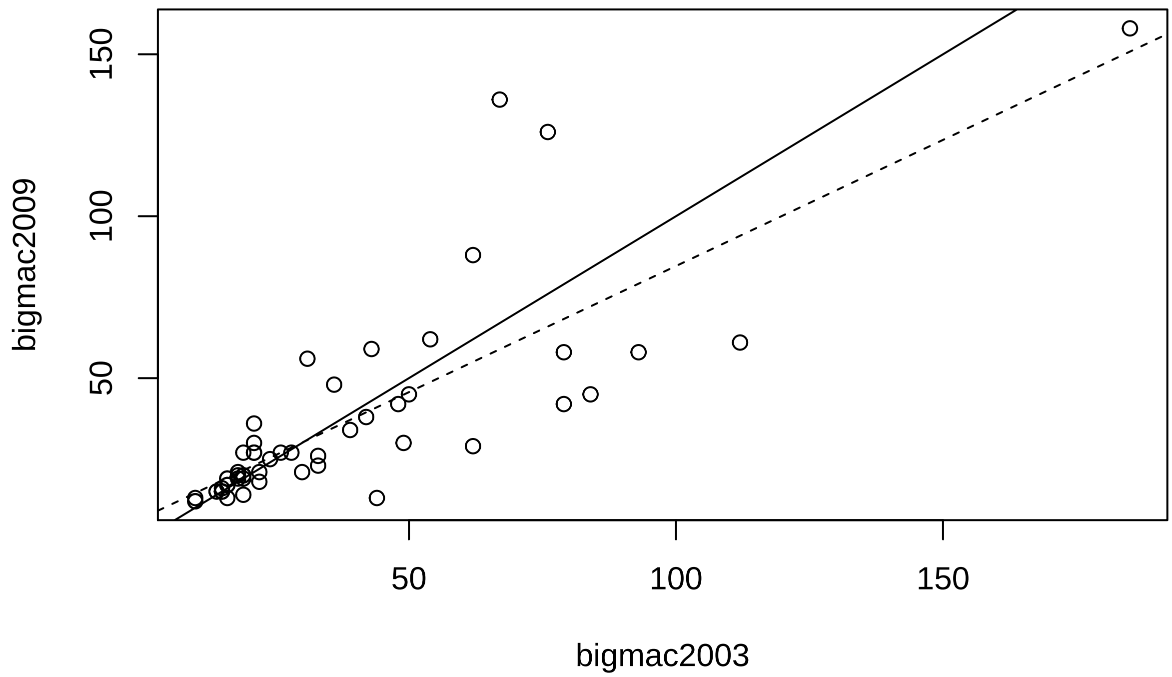
If beta1 $>$ 0 then there is possibility of exponential growth If beta1 $=$ 1 then there would be linear growth. If beta1 $<$ 1 then there would be linear growth but slower as compared to when beta1 $=$ 1

beta0 signify the intercept which is dependent on gamma0. if gamma0 is greater than 1, then log(gamma0) $>$ 0 which would mean the fitted curve will shift up. On the other hand, if gamma0 $<$ 1 then log(gamma0) $<$ 0 which would mean the fitted curve will shift down.

2.4

- 2.4.1

```
with(UBSprices, plot(bigmac2009~bigmac2003))
ols <- lm(bigmac2009~bigmac2003, data = UBSprices)
abline(ols, lty = 2)
abline(0,1)
```



The most unusual cases:

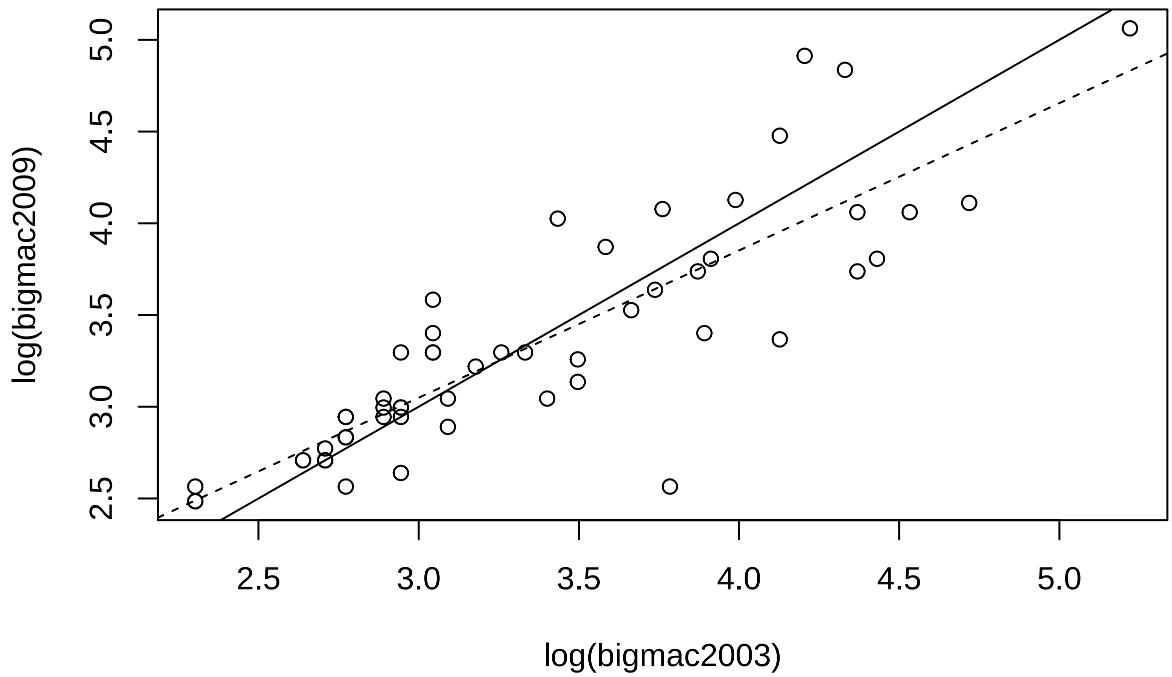
a) Countries such as Caracus and Jakarta that appear to be outliers. These points can significantly affect the simple linear regression line.
b) The countries Nairobi and Mumbai appear to be leverage points with high price in 2003 and low price in 2009. These points can also significantly affect the simple linear regression line.

- 2.4.2

Because of the unusual cases mentioned in 2.4.1, the skewness, outliers and high leverage points can significantly affect the evaluation of model parameters for simple linear regression. Thus a simple linear regression model won't be a good fit for this data.

- 2.4.3

```
with(UBSprices, plot(log(bigmac2009)~log(bigmac2003)))
ols <- lm(log(bigmac2009)~log(bigmac2003), data = UBSprices)
abline(ols, lty = 2)
abline(0,1)
```



The graph can be more sensibly summarized with linear regression because:

a) The distribution of data doesn't appear to be skewed as compared to plot in 2.4.1. Also the variability appears to be constant.
b) The data point - Mumbai doesn't appear to be seperated from the rest of data points. Also, we can see the data point Nairobi appears near to the line with slope 1 and intercept 0. Thus there is plausibility that they can no longer be considered as a leverage point.
c) Though there are some outliers(Caracus and Jakarta) but they appear to be less seperated from the rest of data points as compared to plot in 2.4.1