Homework 7 Solution

Mohsen Nabian

7/09/2015

You collect the following data on four people sampled at random: Age IQ 23 100 18 105 10 95 45 120 Is there an effect of Age on IQ? Please perform all calculations by hand using the equations in the lessons unless otherwise specified. 1. Plot these four points using R.

```
require(ggplot2)
```

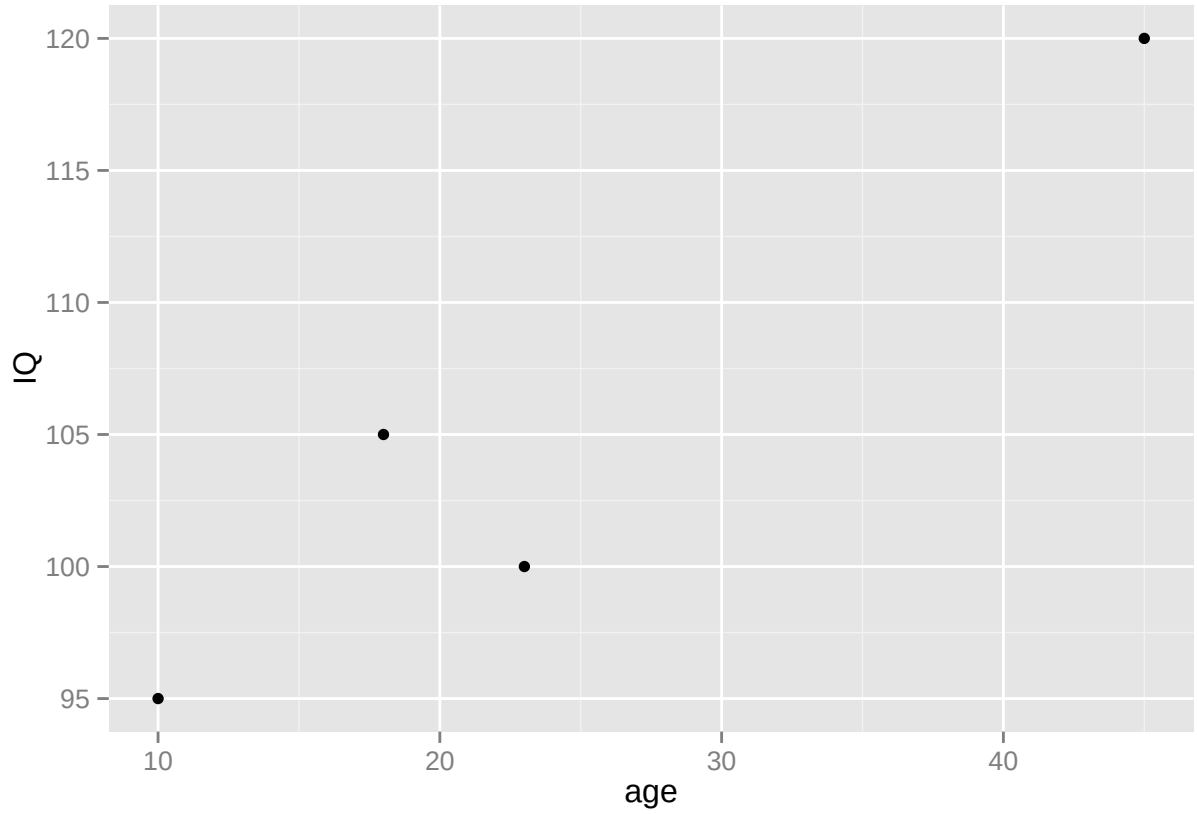
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```

```
age<-c(23,18,10,45)
IQ<-c(100,105,95,120)
df<-data.frame(age,IQ)
df
```

```
##   age  IQ
## 1  23 100
## 2  18 105
## 3  10  95
## 4  45 120
```

```
a<-ggplot(df, aes(x=age,y=IQ))
a+geom_point()
```



2. Calculate the covariance between age and IQ.

$$\bar{x} = (23 + 18 + 10 + 45)/4 = 24$$

```
sd_x<-sd(df$age)
print(sd_x)
```

```
## [1] 14.98888
```

$$s_x = 14.99$$

$$\bar{y} = (100 + 105 + 95 + 120)/4 = 105$$

```
sd_y<-sd(df$IQ)
print(sd_y)
```

```
## [1] 10.80123
```

$$s_y = 10.80$$

$$\text{Var}(x) = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

$$\text{Var}(x) = \frac{1}{4-1}((23-24)^2 + (18-24)^2 + (10-24)^2 + (45-24)^2) = 224.67$$

$$\text{Var}(y) = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$$

$$\text{Var}(y) = \frac{1}{4-1}((100-105)^2 + (105-105)^2 + (95-105)^2 + (120-105)^2) = 116.67$$

$$\text{Cov}(x,y) = \frac{1}{(n-1)}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Cov}(x,y) = \frac{1}{(4-1)} \times ((23-24)(100-105)+(18-24)(105-105)+(10-24)(95-105)+(45-24)(120-105)) = 153.33$$

3. Calculate their correlation. What does the number you get indicate?

$$r = \frac{\text{Cov}(x,y)}{s_x s_y}$$

$$r = \frac{153.33}{14.99 * 10.80} = 110.47$$

4. Calculate the regression coefficients B0 and B1 and write out the equation of the best-fit line relating age and IQ.

$$y = \beta_0 + \beta_1 x$$

$$\beta_1 = \frac{\text{Cov}(x,y)}{\text{Var}(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = 153.33/224.67 = 0.68$$

$$r = \frac{\text{Cov}(x,y)}{s_x s_y} = \beta_1 \frac{s_x}{s_y}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 105 - 0.68 \times 24 = 88.68$$

5. Calculate the predicted yˆi for each xi

$$y = \beta_0 + \beta_1 x$$

$$y = 88.68 + 0.68 \times x$$

$$x(1) = 23, \hat{Y(1)} = 88.68 + 0.68 \times 23 = 104.32$$

$$x(2) = 18, \hat{Y(2)} = 88.68 + 0.68 \times 18 = 100.92$$

$$x(3) = 10, \hat{Y(3)} = 88.68 + 0.68 \times 10 = 95.48$$

$$x(4) = 45, \hat{Y(4)} = 88.68 + 0.68 \times 45 = 119.28$$

6. Calculate Rˆ2 from the TSS/SSE equation. How does it relate to the correlation? What does the number you get indicate?

$$TSS = \sum_i (y_i - \bar{y})^2 = (104.32-105)^2 + (100.92-105)^2 + (95.48-105)^2 + (119.28-105)^2 = 311.66$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = (104.32-100)^2 + (100.92-105)^2 + (95.48-95)^2 + (119.28-120)^2 = 36.06$$

$$R^2 = \frac{TSS - SSE}{TSS} = 0.884$$

7. Calculate the standard error of B1, and use that to test (using the t test) whether B1 is significant.

$$se_{\hat{y}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

$$se_{\hat{y}} = \sqrt{\frac{(104.32-100)^2 + (100.92-105)^2 + (95.48-95)^2 + (119.28-120)^2}{4-2}} = 4.24$$

$$se_{\beta_0} = se_{\hat{y}} \sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}}$$

$$se_{\beta_1} = se_{\hat{y}} \frac{1}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$se_{\beta_1} = 4.24 \times \frac{1}{\sqrt{(23-24)^2 + (18-24)^2 + (10-24)^2 + (45-24)^2}} = 0.163$$

$$T_Statistics = \frac{B1 - 0}{se_{\beta_1}}$$

$$T_Statistics = \frac{0.68 - 0}{/} 0.163 = 4.17$$

$$dg_freedom = n - k - 1 = 4 - 1 - 1 = 2$$

We need to calculate the thresholds with 95% CI and two tailed:

```
thrsld<-qt(0.975,2)
print(thrsld)
```

```
## [1] 4.302653
```

So the threshold is 4.30. As a result we are not able to reject the null and might not say B1 is significant.

8. Calculate the p-value for B1 and interpret it.

```
p_value=pt(0.965,2)
print(p_value)
```

```
## [1] 0.7818206
```

Which means that assuming B1=0, p_value lies on 78.1 percent and does not exceed 95 percent.

9. Calculate the 95% CI for B1 and interpret it.

```
c_i<-qt(0.975,2)
print(c_i)
```

```
## [1] 4.302653
```

95% CI for T_statistics(B1) is 4.30. So:

$$\beta_1 = 4.30 \times 0.163 = 0.688$$

So assuming having same standard deviation, if B1 was calculated as 0.688 it could be proven to be sigificant B1 with 95 percent chance.

10. Confirm your results by regressing IQ on Age using R.

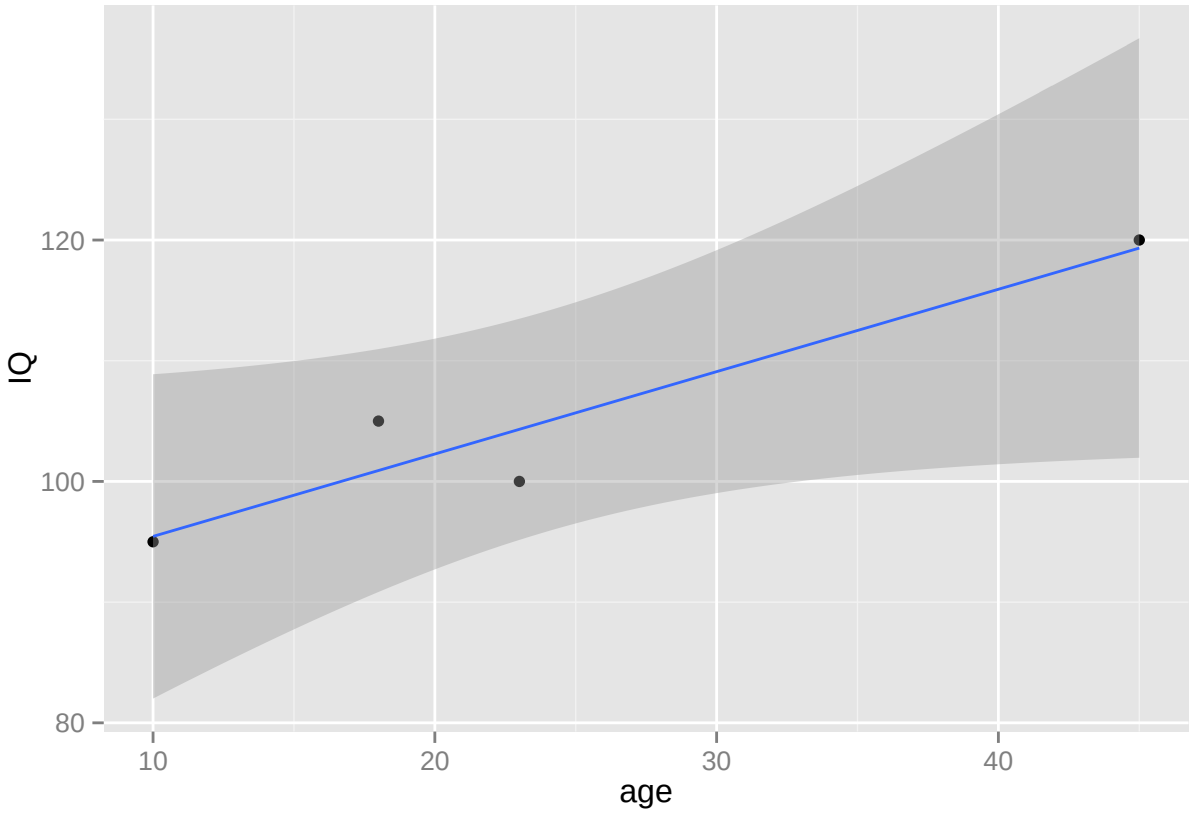
```
biv_model <- lm(df$IQ~df$age,data=df)
summary(biv_model)
```

```
##
## Call:
## lm(formula = df$IQ ~ df$age, data = df)
##
## Residuals:
##       1       2       3       4
## -4.3175  4.0950 -0.4451  0.6677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  88.6202     4.4623  19.860  0.00253 **
## df$age        0.6825     0.1635   4.173  0.05290 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.246 on 2 degrees of freedom
## Multiple R-squared:  0.897,  Adjusted R-squared:  0.8455
## F-statistic: 17.42 on 1 and 2 DF,  p-value: 0.0529
```

So the R code is onfirming our hand calculations.

11. Plot your points again using R, including the linear fit line with its standard error.

```
require(ggplot2)
ggplot(df, aes(x=df$age, y=df$IQ)) + geom_point()+geom_smooth(method=lm)+xlab("age")+ylab("IQ")
```



12. What are you final conclusions about the relationship between age and IQ?

We might say with the given data we can not prove a 95% chance of linear dependency, although it was close to 95%. Morover, R^2=0.88 is not very satisfactory and assuring for this important study. I would suggest to have much more data to make a more strong statement.