HW-07_Gupta_S

Sumit Gupta

October 22, 2017

You collect the following data on four people sampled at random:

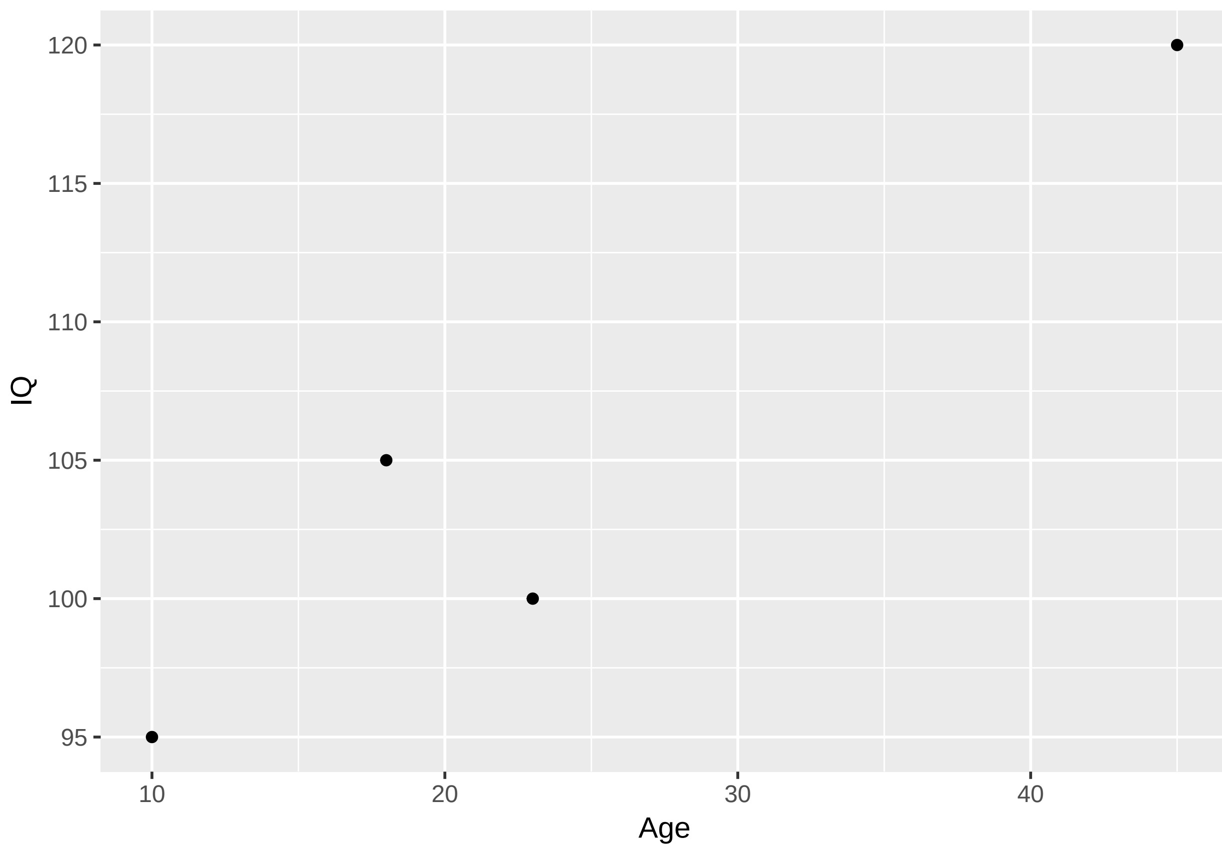
Age IQ 23 100 18 105 10 95 45 120

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.

```r
set.seed(1)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```r
Age <- c(23,18, 10, 45)
IQ<- c(100,105,95,120)
ggplot(data.frame(x=Age, y=IQ), aes(x=Age, y=IQ)) + geom_point()
```



2. Calculate the covariance between age and IQ.

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

$\text{Cov}(Age, IQ) = \frac{1}{(4-1)}[(23-24)(100-105)+(18-24)(105-105)+(10-24)(95-105)+(45-24)(120-105)] = 153.3333$

Verifying using R:

```
cov(Age, IQ)
```

```
## [1] 153.3333
```

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

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

$\sigma_{\text{age}} = 14.988$ $\sigma_{\text{IQ}} = 10.801$

$r = 153.333/(14.988 * 10.801) = 0.9470$

Verifying using R:

```
cor(Age,IQ)
```

```
## [1] 0.9470957
```

4. Calculate the regression coefficients β_0 and β_1 and write out the equation of the best-fit line relating age and IQ.

$y = \beta_0 + \beta_1 x$

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

$Var(x) = \frac{1}{(n-1)} \sum (x_i - \bar{x})^2 = 224.6667$

```
var(Age)
```

```
## [1] 224.6667
```

Thus, $\beta_1 = (153.3333/224.6667) = 0.6824$

$\beta_0 = \bar{y} - \beta_1 \bar{x} = (105 - 16.3776) = 88.6224$

Best fit line equation can be given as:

$y = \beta_0 + \beta_1 x$ That is, $y = 18.6224 + 0.6824x$

5. Calculate the predicted ˆyi for each xi.

$y = 88.6224 + 0.6824(23) = 104.3176$ $y = 88.6224 + 0.6824(18) = 100.9056$ $y = 88.6224 + 0.6824(10) = 95.4464$ $y = 88.6224 + 0.6824(45) = 119.3704$

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

$TSS = \sum_i (y_i - \bar{y})^2 = (100 - 105)^2 + (105 - 105)^2 + (95 - 105)^2 + (120 - 105)^2 = 350$

So, now we compute SSE using the predicted y values.

$SSE = \sum_i (y_i - \hat{y}_i)^2 = (100 - 104.3175)^2 + (105 - 100.905)^2 + (95 - 95.4451)^2 + (120 - 119.3323)^2 = 36.053$

Now, $R^2 = \frac{TSS - SSE}{TSS} = (350 - 36.053)/350 = 0.897$

R-squared is the square of the correlation. It ranges from values (0,1) unlike correlation which ranges between (-1,+1) . The R-squared value of 0.897 indicates that 89.7% of the variation is captured by the model which is good.

7. Calculate the standard error of β_1, and use that to test (using the t test) whether β_1 1 is significant.

$se_{\beta_1} = se_{\hat{y}} \frac{1}{\sqrt{\sum (x_i - \bar{x})^2}} = 0.1635$ Also, we have $\beta_1 = 0.6825$

So, our t-statistic = (0.6824-0)/0.1635 = 4.1737 Degrees of freedom = n-k-1 = 4-1-1 = 2

```r
qt(0.975,2)
```

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

Since, t-statistic (4.17)$<$ 4.30 we conclude that the effect of Age on IQ (β_1) is not statistically significant.

8. Calculate the p-value for β_1 and interpret it.

```r
2*pt(4.173,2, lower.tail = F)
```

```
## [1] 0.05290909
```

9. Calculate the 95% CI for β_1 and interpret it.

$CI_{0.95}(\beta_1) = 0.6824 \pm 4.30265 * 0.1635 = 0.6824 \pm 0.7034 = [-0.021, 1.3858]$

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

```r
dat <- data.frame(Age,IQ)
biv_model <- lm(IQ~Age,data=dat)
summary(biv_model)
```

```
##
## Call:
## lm(formula = IQ ~ Age, data = dat)
##
## 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 **
## 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
```

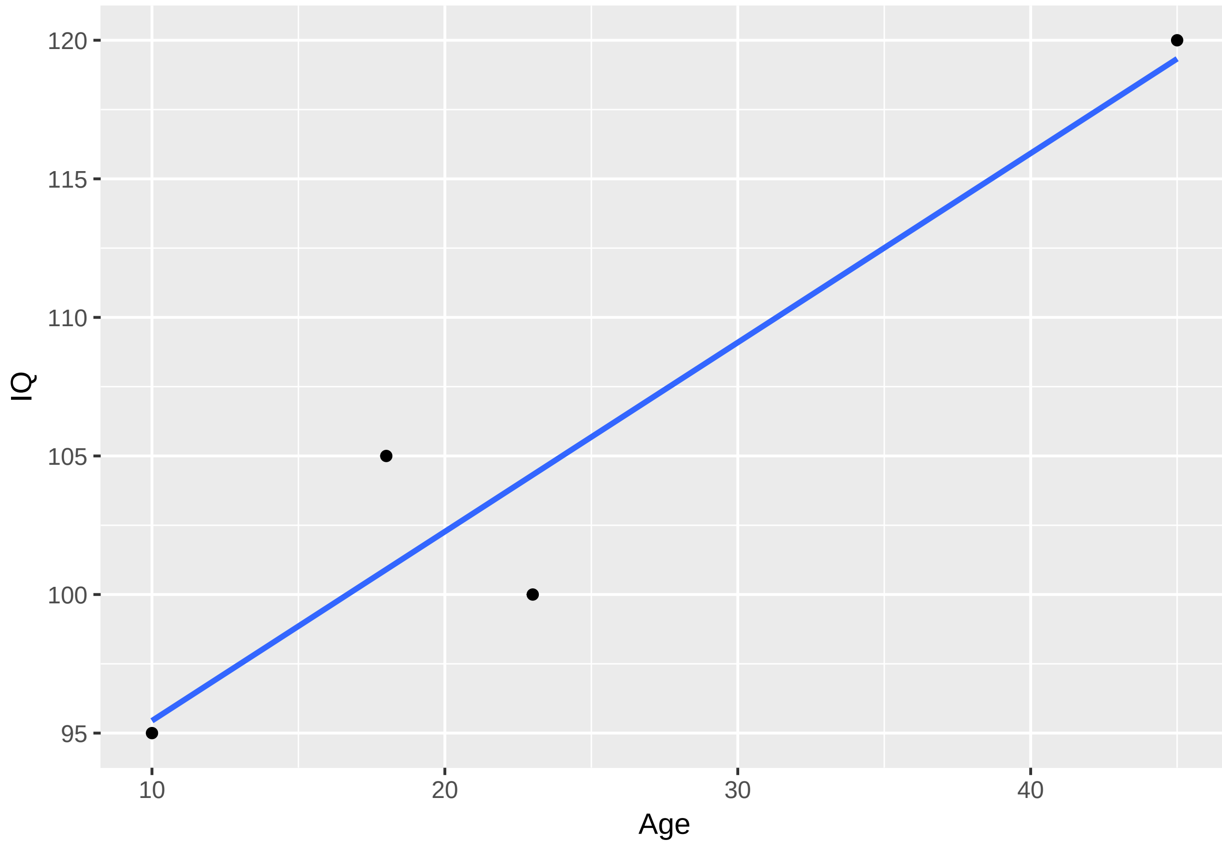
```r
predict(biv_model)
```

```
##        1        2        3        4
## 104.3175 100.9050  95.4451 119.3323
```

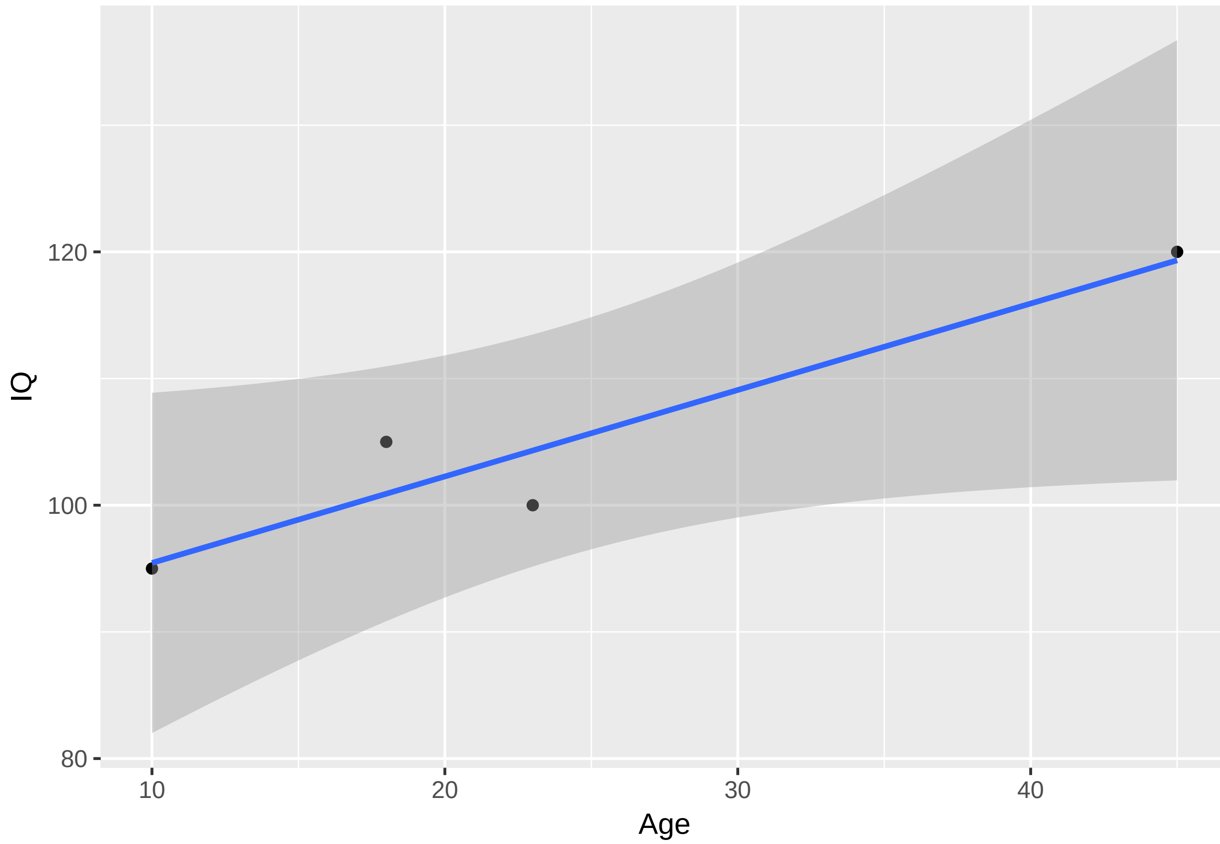
Thus, we confirm that the R-squared, beta values are matching with our computations

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

```r
# Plotting regressions lines without the standard error
ggplot(dat, aes(x=Age, y=IQ)) + geom_point() + geom_smooth(method=lm, se=FALSE)
```




```
# Plotting regression lines with standard error
ggplot(dat, aes(x=Age, y=IQ)) + geom_point() + geom_smooth(method=lm)
```



The shaded area is the 95% CI in the line. The grey area combines the uncertainty in the intercept and slope.

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

a. The effect of Age on IQ is not statistically significant as witnessed earlier.
b. Also, from the plot, we fail to conclude that IQ increases/decreases with Age.
c. The adjusted R squared value of 0.8455 suggest that the model is fairly stable and hence we can confidently support the claims made earlier.