A7.Zunqiu.Wang

Zunqiu Wang

10/19/2021

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

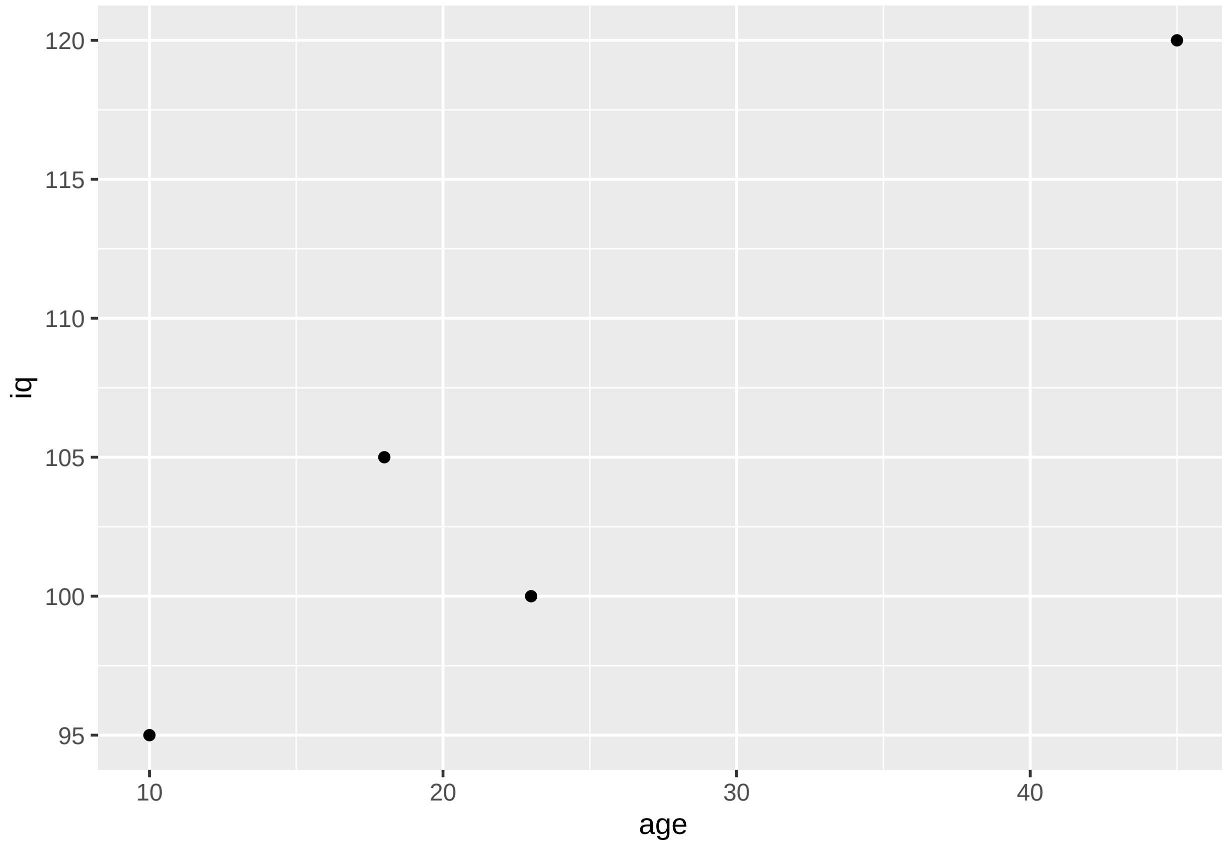
```r
library(ggplot2)
```

Q1

```r
age <- c(23,18,10,45)
iq <- c(100, 105, 95, 120)
ageiq.df <- data.frame(Age=age, IQ=iq)
summary(ageiq.df)
```

```
##       Age              IQ
##  Min.   :10.0   Min.   : 95.00
##  1st Qu.:16.0   1st Qu.: 98.75
##  Median :20.5   Median :102.50
##  Mean   :24.0   Mean   :105.00
##  3rd Qu.:28.5   3rd Qu.:108.75
##  Max.   :45.0   Max.   :120.00
```

```r
ageiq.df %>% ggplot(aes(x=age, y=iq)) + geom_point()
```



Q2

```r
# get mean
mean(ageiq.df$Age)
```

```
## [1] 24
```

```r
mean(ageiq.df$IQ)
```

```
## [1] 105
```

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

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

$$Cov(x,y) = 153.3$$

Q3

```r
#get sd
sd(ageiq.df$Age)
```

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

```r
sd(ageiq.df$IQ)
```

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

$$r = \frac{Cov(x,y)}{s_x s_y}$$
$$r = \frac{153.3}{(14.98888)(10.80123)}$$
$$r = 0.947$$

strong positive correlation between age and IQ

Q4

```r
var(ageiq.df$Age)
```

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

$r = \frac{Cov(x,y)}{s_x s_y} = \beta_1 \frac{s_x}{s_y}$
$\beta_1 = \frac{r}{\frac{s_x}{s_y}} = \frac{0.947}{\frac{14.98888}{10.80123}} = 0.6824926$
$\beta_0 = \bar{y} - \beta_1 \bar{x} = 105 - 0.6824926 * 24 = 88.62018$
LINE OF BEST FIT:
$\hat{y}_i = \beta_0 + \beta_1 x_i$
$\hat{y}_i = 88.62018 + 0.6824926 x_i$

Q5
104.3175=(88.62018)+(0.6824926)*23*
*100.9050=(88.62018)+(0.6824926)*18
95.4451=(88.62018)+(0.6824926)*10*
*119.3323=(88.62018)+(0.6824926)*45

Q6
$TSS = \sum_i (y_i - \bar{y})^2$
$SSE = \sum_i (y_i - \hat{y}_i)^2$
$TSS = \sum_i (100 - 105)^2 + (105 - 105)^2 + (95 - 105)^2 + (120 - 105)^2 = 350$
$SSE = \sum_i (100 - 104.3175)^2 + (105 - 100.9050)^2 + (95 - 95.4451)^2 + (120 - 119.332)^2$
$SSE = 36.05341$ $R^2 = \frac{TSS - SSE}{TSS}$
$R^2 = \frac{350 - 36.05341}{350}$
$R^2 = 0.8969903$
R^2 describes the proportion in a similar manner as correlation but with a distionction that it only ranges 0 and 1. It means 89.6% of variation in IQ can be explained by Age. It indicates how well line of best fits predicts data. The R^2 also termed as proportional reduction in error by 89.7% and expained 89.7% of variation with the model.

Q7
$se_{\hat{y}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{36.05341/(4-2)} = 4.245787$
$se_{\beta_1} = se_{\hat{y}} \frac{1}{\sqrt{\sum (x_i - \bar{x})^2}} = 4.245787 * \frac{1}{\sqrt{(23-24)^2 + (18-24)^2 + (10-24)^2 + (45-24)^2}} = 4.245787 * \frac{1}{\sqrt{674}} = 0.1635416$

$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$
$t_{stat} = \frac{\beta_1 - \mu_0}{se_{\beta_1}} = \frac{0.6824926 - 0}{0.1635416} = 4.173205$

$n = 4, k = 1$
$df = n - k - 1 = 4 - 1 - 1 = 2$

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

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

$t_{stat} < t_{crit}$ So we fail to reject null hypothesis at 95% confidence level with 2 tailed test. Though r coefficient and R^2 suggests linear regression matches data but the correlation between age and IQ is not significant and may be due to chance.

Q8

```r
2 * pt(4.173205, 2, lower.tail = F)
```

```
## [1] 0.0529043
```

$p_{value} = 0.0529043$ Since p val is bigger than t crit, it indicates that with sample size 100 then it will probably result in about 5.2% of sample mean less than or equal to the one in current data. Thus, we fail to reject the null hypothesis.

Q9
$CI = 0.6824926 \pm 4.302653 * 0.1635416 = [-0.021, 1.3858]$

Q10

```r
ageiq.lm <- lm(ageiq.df$IQ ~ ageiq.df$Age)
summary(ageiq.lm)
```

```
##
## Call:
## lm(formula = ageiq.df$IQ ~ ageiq.df$Age)
##
## 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 **
## ageiq.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
```

```r
predict(ageiq.lm)
```

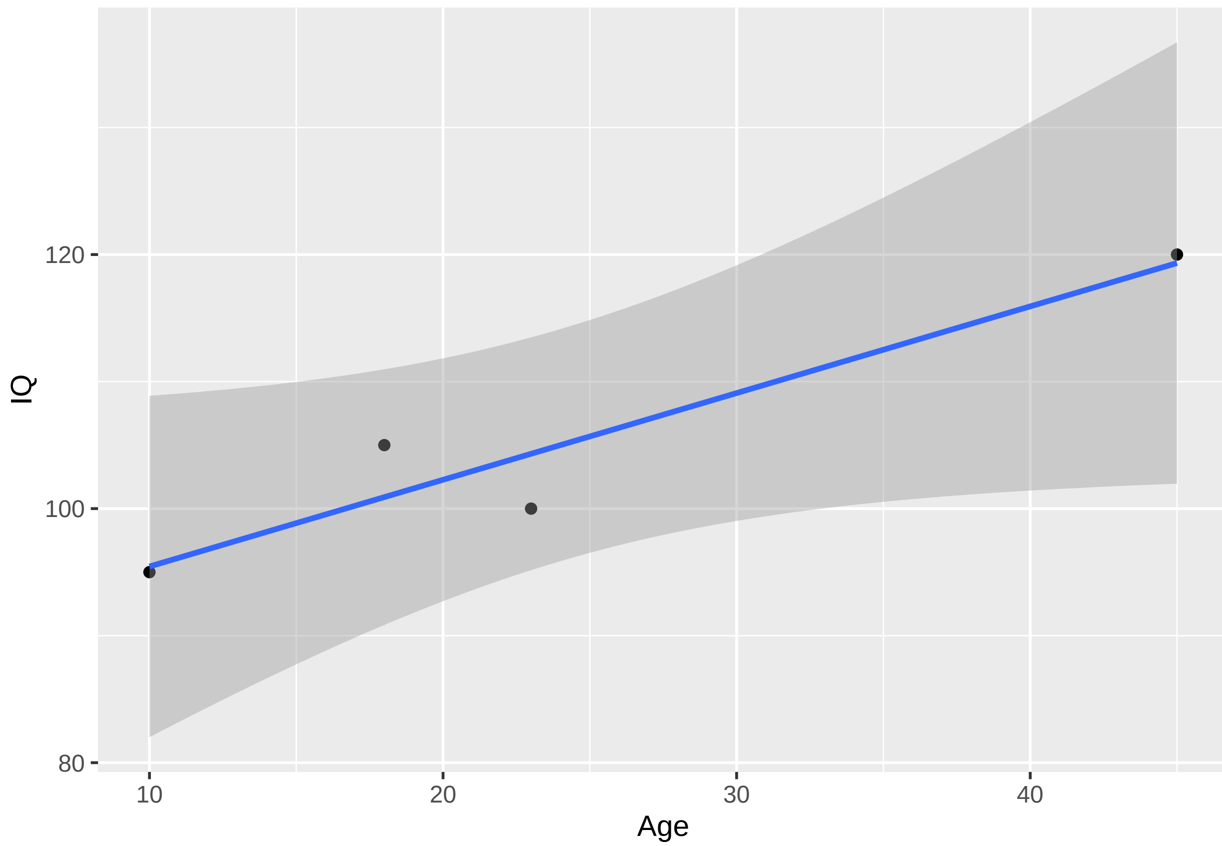
```
##        1        2        3        4
## 104.3175 100.9050  95.4451 119.3323
```

it confirms the above calculations by hand

Q11

```
ggplot(ageiq.df, aes(x=Age, y=IQ)) + geom_point() + geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Q12
r=0.947 suggests that there is a strong positive correlation between age and iq with R^2= 0.897 indicating about 89.7% of error is reduced and is explained by the line of best fit. However, t test tells another story where at 95% confidence level p val implies that it is not significant.