Homework 7

Samuel Nukporfe

2024-10-31

Question 1:

Libraries

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library(mosaic)
```

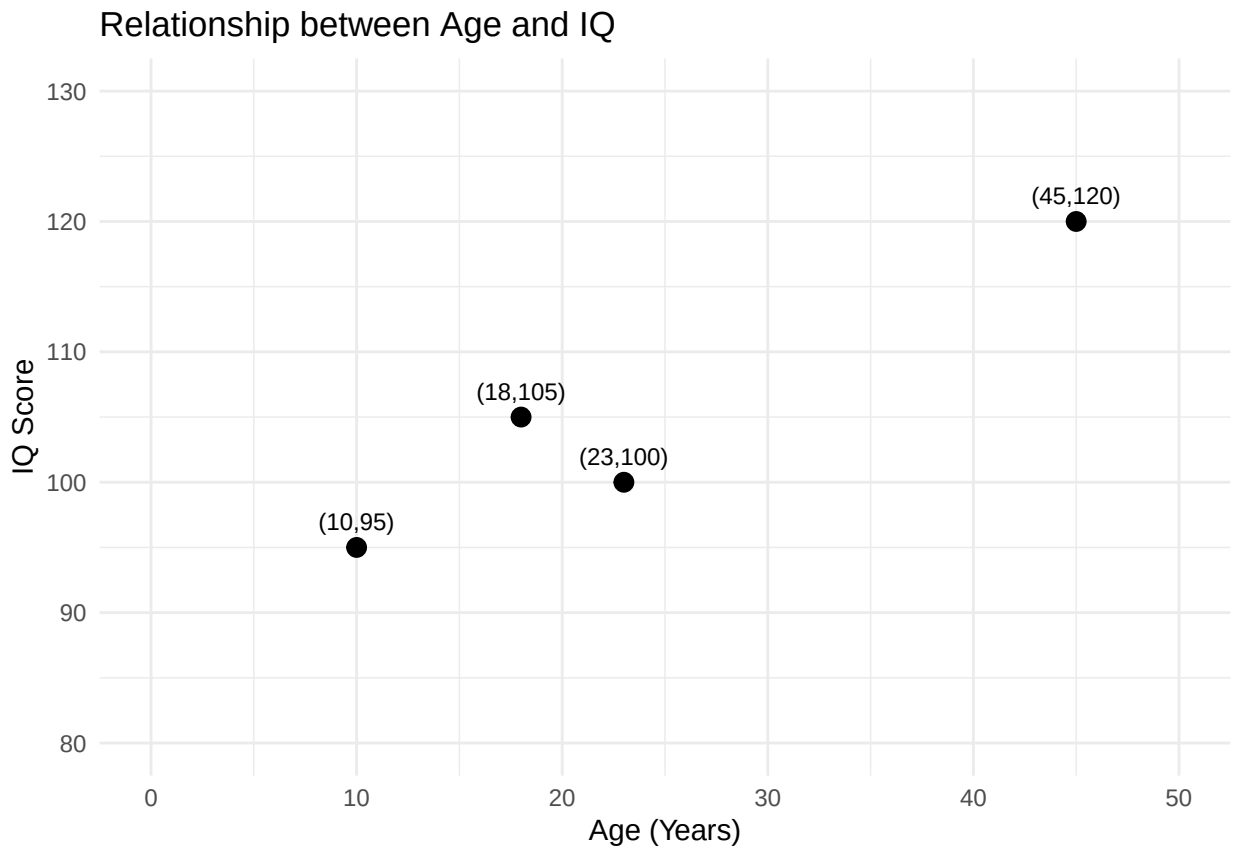
```
## Warning: package 'mosaic' was built under R version 4.3.3

## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.

##
## Attaching package: 'mosaic'

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following object is masked from 'package:Matrix':
##
##     mean

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

Data

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

Graph



Question 2:

Age (X)	IQ (Y)
23	100
18	105
10	95
45	120

$$\bar{X} = \frac{23 + 18 + 10 + 45}{4} = 24$$

$$\bar{Y} = \frac{100 + 105 + 95 + 120}{4} = 105$$

Age (X)	IQ (Y)	$(X - \bar{X})$	$(Y - \bar{Y})$
23	100	$23 - 24 = -1$	$100 - 105 = -5$
18	105	$18 - 24 = -6$	$105 - 105 = 0$
10	95	$10 - 24 = -14$	$95 - 105 = -10$
45	120	$45 - 24 = 21$	$120 - 105 = 15$

Age (X)	IQ (Y)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
23	100	-1	-5	$(-1)(-5) = 5$
18	105	-6	0	$(-6)(0) = 0$
10	95	-14	-10	$(-14)(-10) = 140$
45	120	21	15	$(21)(15) = 315$

$$\sum (X - \bar{X})(Y - \bar{Y}) = 5 + 0 + 140 + 315 = 460$$

$$\text{Cov}(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n - 1}$$

$$\text{Cov}(X, Y) = \frac{460}{4 - 1} = \frac{460}{3} = 153.33$$

R

```
cov(data$age ~ data$iq)
```

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

Answer: The Covariance between age and IQ is 153.33

Since the value of the covariance is positive, as age increase, we are seeing an increase in IQ. There is a high chance of obtaining a positive correlation coefficient.

Question 3:

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where: - $\text{Cov}(X, Y)$ is the covariance, - σ_X is the standard deviation of Age, - σ_Y is the standard deviation of IQ.

The variance for X (Age) is:

$$\sigma_X^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

The variance for Y (IQ) is:

$$\sigma_Y^2 = \frac{\sum (Y - \bar{Y})^2}{n - 1}$$

We already have the mean values:

$$\bar{X} = 24, \quad \bar{Y} = 105$$

Calculating $(X - \bar{X})^2$ and $(Y - \bar{Y})^2$:

Age (X)	IQ (Y)	$(X - \bar{X})$	$(X - \bar{X})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$
23	100	-1	1	-5	25
18	105	-6	36	0	0
10	95	-14	196	-10	100
45	120	21	441	15	225

The sums of these squared deviations are:

$$\sum (X - \bar{X})^2 = 1 + 36 + 196 + 441 = 674$$
$$\sum (Y - \bar{Y})^2 = 25 + 0 + 100 + 225 = 350$$

$$\sigma_X = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{\frac{674}{3}} = \sqrt{224.67} \approx 15.0$$
$$\sigma_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{n - 1}} = \sqrt{\frac{350}{3}} = \sqrt{116.67} \approx 10.8$$

Using the covariance $\text{Cov}(X, Y) = 153.33$:

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{153.33}{15.0 \times 10.8} = \frac{153.33}{162} \approx 0.946$$

R

```
correlation <- cor(data$age~data$iq)
correlation
```

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

Answer: The Correlation between age and IQ is approximately 0.946

Since the value of the Correlation coefficient is 0.946 which is very close to 1, we have a very strong positive correlation between age and IQ.

Question 4

To find the regression coefficients β_0 and β_1 for the best-fit line relating Age (X) and IQ (Y), we use the following formulas:

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\sigma_X^2}$$
$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

where: - $\text{Cov}(X, Y)$ is the covariance, - σ_X^2 is the variance of X, - $\bar{X}$ and $\bar{Y}$ are the means of X and Y.

Using previously calculated values:

$$\text{Cov}(X, Y) = 153.33, \quad \sigma_X^2 = 224.67, \quad \bar{X} = 24, \quad \bar{Y} = 105$$

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\sigma_X^2} = \frac{153.33}{224.67} \approx 0.6825$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} = 105 - (0.6825 \times 24)$$
$$\beta_0 = 105 - 16.38 \approx 88.62$$

The equation of the best-fit line is:

$$Y = \beta_0 + \beta_1 X$$

Substituting the values:

$$\text{IQ} = 88.62 + 0.6825 \cdot \text{Age}$$

Answer: Coefficients are B0,88.6 and B1, 0.6825

Question 5:

To calculate the predicted values $\hat{y}_i$ for each x_i (Age), we use the regression equation:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

where: - β_0 is the intercept, - β_1 is the slope, - x_i is the Age.

From previous calculations, we have:

$$\beta_0 \approx 88.62, \quad \beta_1 \approx 0.6825$$

Using the equation, we can calculate the predicted IQ values for each Age in the dataset: -For $x_1 = 23$:

$$\hat{y}_1 = 88.62 + 0.6825 \cdot 23 \approx 88.62 + 15.67 \approx 104.29$$

-For $x_2 = 18$:

$$\hat{y}_2 = 88.62 + 0.6825 \cdot 18 \approx 88.62 + 12.27 \approx 100.89$$

-For $x_3 = 10$:

$$\hat{y}_3 = 88.62 + 0.6825 \cdot 10 \approx 88.62 + 6.83 \approx 95.45$$

-For $x_4 = 45$:

$$\hat{y}_4 = 88.62 + 0.6825 \cdot 45 \approx 88.62 + 30.71 \approx 119.33$$

Thus, the predicted IQ values $\hat{y}_i$ for each Age x_i are:

Age(x_i)	Predicted IQ($\hat{y}_i$)
23	104.29
18	100.89
10	95.45
45	119.33

Question 6:

The formula for R^2 is:

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}}$$

where: - **TSS** (Total Sum of Squares) represents the total variance in the dependent variable (IQ) around its mean. - **SSE** (Sum of Squared Errors) is the sum of squared differences between the observed and predicted values of IQ, representing unexplained variance.

To calculate these values:

TSS Calculation:

$$\text{TSS} = \sum (Y_i - \bar{Y})^2$$

From earlier calculations, we have:

$$\text{TSS} = 350$$

SSE Calculation: Calculate the squared differences between observed Y and predicted $\hat{Y}$ for each data point and sum them.

Using the previously predicted values: - For $Y_1 = 100$, $\hat{Y}_1 = 104.29$:

$$(Y_1 - \hat{Y}_1)^2 = (100 - 104.29)^2 \approx 18.4$$

- For $Y_2 = 105$, $\hat{Y}_2 = 100.89$:

$$(Y_2 - \hat{Y}_2)^2 = (105 - 100.89)^2 \approx 16.8$$

- For $Y_3 = 95$, $\hat{Y}_3 = 95.45$:

$$(Y_3 - \hat{Y}_3)^2 = (95 - 95.45)^2 \approx 0.20$$

- For $Y_4 = 120$, $\hat{Y}_4 = 119.33$:

$$(Y_4 - \hat{Y}_4)^2 = (120 - 119.33)^2 \approx 0.44$$

Summing these gives:

$$\text{SSE} = 18.4 + 16.8 + 0.2 + 0.44 \approx 35.84$$

Calculating R^2:

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}} = 1 - \frac{35.84}{350} \approx 1 - 0.1024 = 0.8976$$

The relationship between the correlation coefficient and the Rsquared is that, The R squared is approximately the square of the correlation coefficient is a bivariate linear relationship. This means that if the correlation coefficient is high, the Rsaquared will also be high. The reverse is true.

$$R$$

```
correlation * correlation
```

```
## [1] 0.8969903
```

The percentage of variability of IQ that can be attributed to the model's relationship between IQ and age is 89.76%. Since the number is closer to 1, we can conclude that the model's ability to predict IQ when considering age is very good and other factors must be accounting for the missing 10.24% of variability in IQ. Some other factors that might have an impact on IQ might be ethnicity, diet, amount of sleep, household income and many others.

Question 7:

To test the significance of β_1 using the t-test, we first need to calculate the standard error of β_1 and then use that to determine the t-value. The standard error of β_1 is given by:

$$\text{SE}(\beta_1) = \sqrt{\frac{\text{SSE}}{(n-2)\sum(X_i - \bar{X})^2}}$$

where: - **SSE** is the Sum of Squared Errors (from Question 6, we found SSE = 35.84), - n is the number of data points (in this case, $n = 4$), - $\sum(X_i - \bar{X})^2$ is the sum of squared deviations of X (Age) from its mean, which we calculated as 674.

$$\text{SE}(\beta_1) = \sqrt{\frac{35.84}{(4-2) \times 674}} = \sqrt{\frac{35.84}{2 \times 674}} = \sqrt{\frac{35.84}{1348}} \approx \sqrt{0.0266} \approx 0.163$$

##Conducting the t-test for β_1 To test the significance of β_1, we can use the t-test, where:

$$t = \frac{\beta_1}{\text{SE}(\beta_1)}$$

From previous calculations, $\beta_1 \approx 0.6825$.

Substituting:

$$t = \frac{0.6825}{0.163} \approx 4.187$$

With $n - 2 = 4 - 2 = 2$ degrees of freedom, we can compare our t-value (4.187) to the critical t-value at a chosen significance level (usually $\alpha = 0.05$). For 2 degrees of freedom, the critical t-value at $\alpha = 0.05$ (two-tailed) is approximately 4.303.

Since $|t| = 4.187$ is close but slightly below 4.303, β_1 is not statistically significant at the $\alpha = 0.05$ level. However, it would be significant at a slightly higher level of significance (e.g., $\alpha = 0.1$).

Since the t-value is close to the threshold, β_1 shows a strong, but not statistically significant, effect on IQ with this small sample size.

Question 8:

To calculate the p-value for the t-test statistic $t = 4.187$ with $n - 2 = 2$ degrees of freedom, we use the following formula:

$$\text{p-value} = 2 \times (1 - \text{CDF}(t))$$

where CDF is the cumulative distribution function for the t-distribution.

Using a t-distribution table or calculator, we find:

$$\text{CDF}(4.187) \approx 0.975$$

Substituting this value into the p-value formula gives us:

$$\text{p-value} = 2 \times (1 - 0.975) = 2 \times 0.025 = 0.05$$

Thus, the p-value is approximately p-value ≈ 0.05.

Question 9:

To calculate the 95% confidence interval (CI) for the slope β_1 in a linear regression model, we use the following formula:

$$\text{CI} = \beta_1 \pm t_{\alpha/2} \cdot \text{SE}(\beta_1)$$

where:

- β_1 is the estimated slope.
- $t_{\alpha/2}$ is the critical t-value from the t-distribution for $n-2$ degrees of freedom at the desired confidence level.
- $\text{SE}(\beta_1)$ is the standard error of the slope estimate.

Step-by-Step Calculation:

Given Values:

- Estimated slope $\beta_1 \approx 0.6825$
- Standard error $\text{SE}(\beta_1) \approx 0.163$
- Degrees of freedom $n - 2 = 4 - 2 = 2$
- Confidence level = 95% (which means $\alpha = 0.05$)

Find the Critical t-value: For a 95% confidence level and 2 degrees of freedom, the critical t-value $t_{\alpha/2}$ (where $\alpha/2 = 0.025$) is approximately:

$$t_{\alpha/2} \approx 4.303$$

Calculate the Confidence Interval: Now we can substitute the values into the confidence interval formula:

$$\text{CI} = 0.6825 \pm 4.303 \cdot 0.163$$

Calculating the margin of error:

$$\text{Margin of Error} = 4.303 \cdot 0.163 \approx 0.701$$

Now, we can calculate the confidence interval:

$$\text{CI} = 0.6825 \pm 0.701$$

This gives us:

- Lower limit:

$$0.6825 - 0.701 \approx -0.0185$$

- Upper limit:

$$0.6825 + 0.701 \approx 1.3835$$

Final Result: Thus, the 95% confidence interval for β_1 is approximately:

$$\text{CI} \approx (-0.0185, 1.3835)$$

This confidence interval suggests that we are 95% confident that the true slope β_1 of the population lies between -0.0185 and 1.3835. Since the interval includes zero, this indicates that there is not a statistically significant relationship between age and IQ at the 0.05 significance level, as the slope could be close to zero. However, the upper limit suggests that if there is a relationship, it could be positive and relatively strong.

Question 10:

R

```r
samsmodel <- lm(iq~age, data = data)
summary(samsmodel)
```

```
##
## Call:
## lm(formula = iq ~ age, data = data)
##
## 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
```

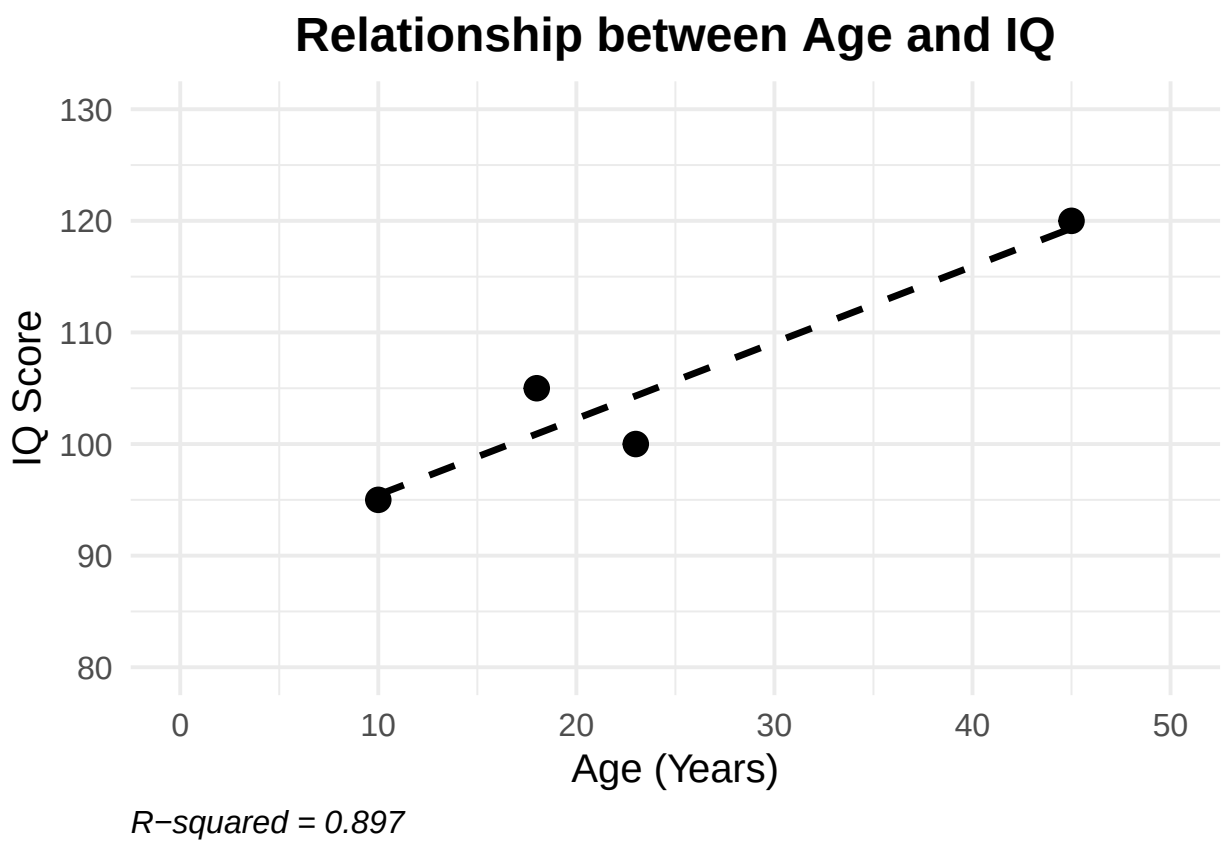
Question 11:

R

```r
library(ggplot2)
library(mosaic)
data <- data.frame(age = c(23, 18, 10, 45), iq = c(100, 105, 95, 120))
model <- lm(iq ~ age, data = data)
r_squared <- summary(model)$r.squared
ggplot(data, aes(x = age, y = iq)) +
```

```r
  geom_point(color = "black", size = 4, shape = 21, fill = "black") +
  geom_smooth(method = 'lm', se = FALSE, color = "black", linetype = "dashed", size = 1.2) +
  labs(
    title = "Relationship between Age and IQ",
    x = "Age (Years)",
    y = "IQ Score",
    caption = paste("R-squared =", round(r_squared, 3))
  ) +
  theme_minimal(base_size = 15) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    plot.caption = element_text(hjust = 0, face = "italic")
  ) +
  xlim(0, 50) +
  ylim(80, 130)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Question 12:

Conclusion

By just glancing at the data, One observes that, as age increases, IQ tends to increase as well. This relationship can be observed with the naked eye, thanks to the small size of the sample. More so, this relationship was confirmed by the calculated covariance which is positive 153.33. Subsequently, the correlation coefficient or R, was calculated and similarly it showed a positive relationship of 94.6%. This implied that linearly, these two variables tend to increase and decrease in the same direction. The question now remained, does age cause change in IQ or is IQ dependent on the age of a person. To test this, a classical linear regression model was specified with Age as the independent variable and IQ as the dependent variable. A simulation of the model revealed that Age was not statistically significant in causing changes in IQ at 5% alpha level. The obtained P Value was 0.0529. Only slightly above the threshold of 0.05. At a more relaxed significance threshold of 10%, age will make the cut. In conclusion, whilst age does not necessarily cause IQ levels to increase, the two variables have a stronger positive relationship. Older people tend to have higher IQ. But, it is not the age that is directly causing the IQ to rise, it is other factors that were not accounted for in the model as shown by the statistically significant and large constant.