Holsenbeck_S_7

Stephen Synchronicity

2017-10-26

Homework Outline

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. 2. Calculate the covariance between age and IQ. 3. Calculate their correlation. What does the number you get indicate? 4. Calculate the regression coefficients β_0 and β_1 and write out the equation of the best-fit line relating age and IQ. 5. Calculate the predicted $\hat{y}_i$ for each x_i. 6. Calculate

$$R^2$$

from the TSS/SSE equation. How does it relate to the correlation? What does the number you get indicate? 7. Calculate the standard error of β_1, and use that to test (using the t test) whether β_1 is significant. 8. Calculate the p-value for β_1 and interpret it. 9. Calculate the 95% CI for β_1 and interpret it. β_0. 10. Confirm your results by regressing IQ on Age using R. 11. Plot your points again using R, including the linear fit line with its standard error. 12. What are you final conclusions about the relationship between age and IQ?

Homework 7

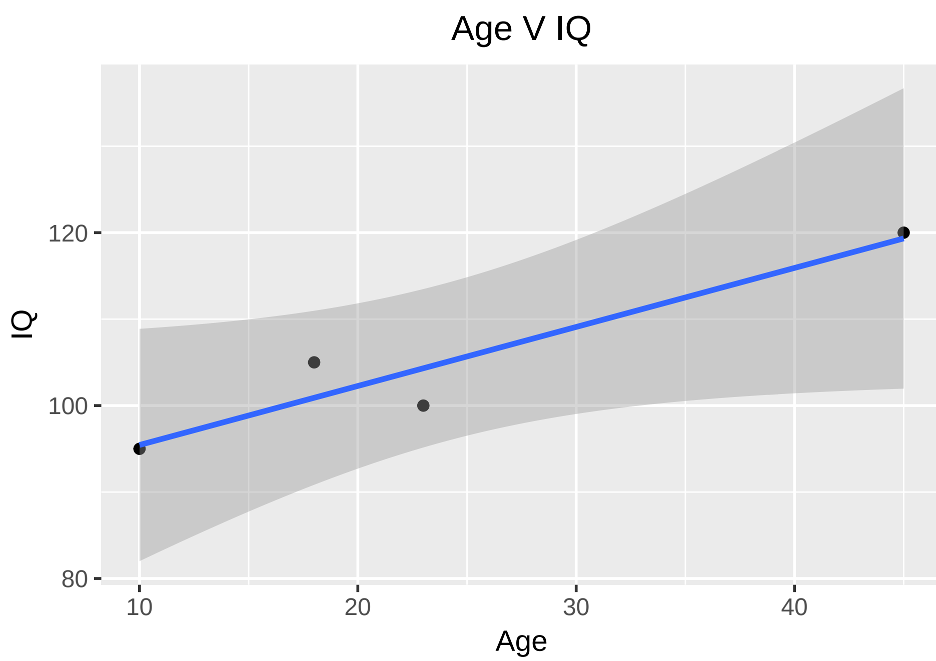
```
ageIQ <- tibble::tribble(~Age, ~IQ, 23L, 100L, 18L, 105L, 10L, 95L, 45L, 120L)
ageIQ.s <- cbind(ageIQ, Sums = rowSums(ageIQ))
ageIQ.s <- rbind(ageIQ.s, Sums = colSums(ageIQ.s))
summary(ageIQ)
```

```
##       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
```

1

Plot these four points using R.

```
ggplot(data = ageIQ, mapping = aes(x = Age, y = IQ)) + geom_point() + geom_smooth(method = lm) +
    ggtitle("Age V IQ") + theme(plot.title = element_text(hjust = 0.5))
```



2

Calculate the covariance between age and IQ.

A)

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

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

$$\text{Cov}(x, y) = 153.\bar{3}$$

```
ageIQ.cov <- ageIQ %>% mutate(S = (Age - mean(ageIQ$Age)) * (IQ - mean(ageIQ$IQ))) %>%
    colSums()
c(ageIQ.cov[3]/3, cov(ageIQ$Age, ageIQ$IQ))
```

```
##        S
## 153.3333 153.3333
```

3

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

A)

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

$$r = \frac{153.\bar{3}}{14.9\bar{8})(10.80123)}$$

$$r = .947$$

The closeness of the r value to 1 indicates a strong positive correlation between age & IQ.

```
(r <- 153.33333/(sd(ageIQ$Age) * sd(ageIQ$IQ)))
```

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

```
c(r, cor(ageIQ$Age, ageIQ$IQ))
```

```
## [1] 0.9470957 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.

A)

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

$$\frac{r}{\frac{s_x}{s_y}} = \beta_1$$

$$\beta_1 = \frac{0.9470957}{\frac{14.9\bar{8}}{10.80123}}$$

$$\beta_1 = 0.6824926$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_0 = 105 - 0.6824926 * 24$$

$$\beta_0 = 88.62018$$

Line of best fit:

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

$$\hat{y}_i = 88.62018 + 0.6824926 x_i$$

```r
(b1 <- r/(sd(ageIQ$Age)/sd(ageIQ$IQ)))
```

```
## [1] 0.6824926
```

```r
(b0 <- mean(ageIQ$IQ) - b1 * mean(ageIQ$Age))
```

```
## [1] 88.62018
```

5

Calculate the predicted $\hat{y}_i$ for each x_i.

A)

$$104.3175 = (88.62018) + (0.6824926)23$$

$$100.9050 = (88.62018) + (0.6824926)18$$

$$95.4451 = (88.62018) + (0.6824926)10$$

$$119.3323 = (88.62018) + (0.6824926)45$$

```r
(y.i <- ageIQ %>% mutate(`y^i` = b0 + b1 * Age))
```

```
## # A tibble: 4 x 3
##     Age    IQ    `y^i`
##   <int> <int>    <dbl>
## 1    23   100 104.3175
## 2    18   105 100.9050
## 3    10    95  95.4451
## 4    45   120 119.3323
```

6

Calculate R^2 from the TSS/SSE equation. How does it relate to the correlation? What does the number you get indicate?

A)

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

$$\text{Where } TSS = \sum_i (y_i - \bar{y})^2 \text{ and } SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$TSS = \sum_i (y_i - \bar{y})^2$$

$$TSS = \sum_i (100 - 105)^2 + (105 - 105)^2 + (95 - 105)^2 + (120 - 105)^2$$

$$TSS = 350$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$SSE = \sum_i (100 - 104.3175)^2 + (105 - 100.9050)^2 + (95 - 95.4451)^2 + (120 - 119.332)^2$$

$$SSE = 36.05341$$

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

$$R^2 = \frac{350 - 36.05341}{350}$$

$$R^2 = 0.8969903$$

The

$$R^2$$

value describes the proportion of the variation in the Y data that is explained by the values for X. The

$$R^2$$

value is also sometimes called the proportional reduction in error and is described as the proportional reduction of error in the variation in Y. A value of 0.897 indicates that 89.7% of the variation in Y is explained by X.

```
(TSS <- ageIQ %>% mutate(tss = ((IQ - mean(ageIQ$IQ))^2)))
```

```
## # A tibble: 4 x 3
##     Age    IQ   tss
##   <int> <int> <dbl>
## 1    23   100    25
## 2    18   105     0
## 3    10    95   100
## 4    45   120   225
```

```
sum(TSS$tss)
```

```
## [1] 350
```

```
(SSE <- y.i %>% mutate(sse = (IQ - `y^i`)^2))
```

```
## # A tibble: 4 x 4
##     Age    IQ    `y^i`        sse
##   <int> <int>    <dbl>      <dbl>
## 1    23   100 104.3175 18.6408704
```

```
## 2    18   105 100.9050 16.7686597
## 3    10    95  95.4451  0.1981176
## 4    45   120 119.3323  0.4457647
```

```r
sum(SSE$sse)
```

```
## [1] 36.05341
```

```r
(`r^2` <- (sum(TSS$tss) - sum(SSE$sse))/sum(TSS$tss))
```

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

7

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

A)

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

$$se_{\hat{y}} = \sqrt{\frac{SSE}{n-2}}$$

$$se_{\hat{y}} = \sqrt{\frac{36.05341}{4-2}}$$

$$se_{\hat{y}} = 4.245787$$

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

$$se_{\beta_0} = 4.245787\sqrt{\frac{2978}{4(674)}}$$

$$se_{\beta_0} = 4.462319$$

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

$$se_{\beta_1} = 4.245787\frac{1}{\sqrt{674}}$$

$$se_{\beta_1} = 0.1635416$$

T-Test:

$$H_0$$

:

$$\beta_1 = 0$$

$$H_a$$

:

$$\beta_1 \neq 0$$

$$t_{stat} = \frac{\beta_1 - \mu_0}{se_{\beta_1}}$$

$$t_{stat} = \frac{0.6824926 - 0}{0.1635416}$$

$$t_{stat} = 4.173205$$

$$df = n - k - 1$$

$$df = 4 - 1 - 1 = 2$$

$$t_{crit} = 4.302653$$

We fail to reject the null hypothesis at the 95% confidence level with a 2-tailed test. The effect of age on IQ is not statistically significant.

```
(sey <- sqrt(sum(SSE$sse)/2))
```

```
## [1] 4.245787
```

```
(se <- ageIQ %>% mutate(`xi^2` = Age^2, `xi-x^2` = (Age - mean(ageIQ$Age))^2))
```

```
## # A tibble: 4 x 4
##     Age    IQ `xi^2` `xi-x^2`
##   <int> <int>  <dbl>    <dbl>
## 1    23   100    529        1
## 2    18   105    324       36
## 3    10    95    100      196
## 4    45   120   2025      441
```

```
(seb0 <- sey * sqrt(sum(se$`xi^2`)/(4 * sum(se$`xi-x^2`))))
```

```
## [1] 4.462319
```

```
(seb1 <- sey * (1/sqrt(sum(se$`xi-x^2`))))
```

```
## [1] 0.1635416
```

```r
(ts <- (b1 - 0)/seb1)
```

```
## [1] 4.173205
```

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

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

8

Calculate the p-value for β_1 and interpret it.

A) The $p_{value} = 0.05290431$, which is slightly greater than $\alpha = 0.05$ indicating that the influence of age on IQ is not statistically significant.

```r
(pv <- 2 * pt(ts, 2, lower.tail = F))
```

```
## [1] 0.05290431
```

9

Calculate the 95% CI for β_1 and interpret it.

A)
$$\text{CI} = 0.6824926 \pm 4.173205 * 0.1635416 = [1.110223e^{-16}, 1.364985]$$

The 95% confidence interval for β_1 indicates that for each year of age we would expect that IQ would increase between $1.110223e^{-16}$ and 1.364985.

```r
(ci <- c(b1 + ts * seb1, b1 - ts * seb1))
```

```
## [1] 1.364985e+00 1.110223e-16
```

10

Confirm your results by regressing IQ on Age using R.

A)

```r
summary(lm(IQ ~ Age, data = ageIQ))
```

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

11

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

A) This was included in the initial plot.

12

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

A) $r = .947$ indicating there is a strong positive correlation between age and IQ. $R^2 = 0.897$ indicating that approximately 89.7% of the variation in IQ is explained by age. The t-test at the 95% confidence level indicates that the effect of age on IQ is not statistically significant despite the strength of the correlation.