Exercise 2

———-

2023-03-19

2.1

Generell:

GPA (aka dependent variable): The GPA varies depending on the ACT.

ACT (aka explanatory variable): If you change the ACT score the GPA score should also change.

Least-square estimators: $\widehat{\beta}_1$, $\widehat{\beta}_2$ They are used to create the least-square regression line which is a 'mean' line between all values.

The $\widehat{\beta}_1$ is calculated by $\bar{y}$-$\widehat{\beta}_2\bar{x}$. Which equals the distance on the y-axis.

The $\widehat{\beta}_2$ is calculated by $\frac{S_{xy}}{S_{xx}}$ which we proved the last time is equal to $\frac{cov(x_i,y_i)}{var(x_i)}$ and is the slope of the least-square regression line.

In this case we have the following data:

```
df <- data.frame(GPA=c(2.8,3.4,3.0,3.5,3.6,3.0,2.7,3.7), ACT=c(21,24,26,27,29,25,25,30))

avg_gpa_m <- sum(df$GPA) / length(df$GPA)
avg_gpa <- mean(df$GPA)
avg_act_m <- sum(df$ACT) / length(df$ACT)
avg_act <- mean(df$ACT)

nominator <- sum((df$GPA - avg_gpa) * (df$ACT - avg_act))
numerator <- sum((df$ACT - avg_act)*(df$ACT - avg_act))

beta2 <- nominator / numerator
beta1 <- avg_gpa - (beta2 * avg_act)


model <- lm(GPA ~ ACT, data=df)

summary(model)
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42308 -0.14863  0.06703  0.10742  0.37912
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.56813    0.92842   0.612   0.5630
## ACT          0.10220    0.03569   2.863   0.0287 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2692 on 6 degrees of freedom
## Multiple R-squared:  0.5774, Adjusted R-squared:  0.507
## F-statistic: 8.199 on 1 and 6 DF,  p-value: 0.02868
```

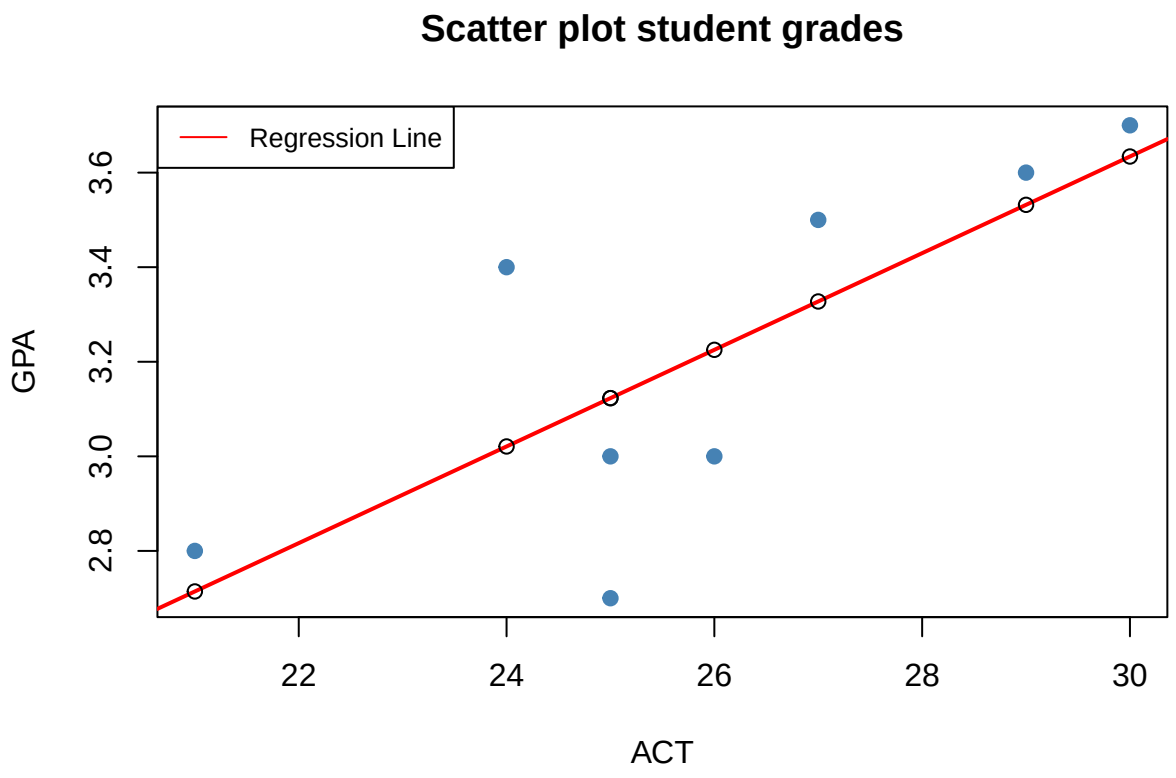
Solution a:

$\widehat{\beta}_1 = 0.568131868131867$

$\widehat{\beta}_2 = 0.102197802197802$

Solution b:

```
plot(df$ACT, df$GPA, main="Scatter plot student grades",
     xlab = "ACT",ylab = "GPA", pch = 19, col = "steelblue")
abline(model, col='red', lwd=2)
df$GPA_hat <- beta2 * df$ACT + beta1
points(df$ACT, df$GPA_hat)
legend("topleft", legend=c("Regression Line"), col=c("red"), lty=1:1, cex=0.8)
```



The better someone is at the ACT the better they are at the GPA.

For a student who gets 0 points at the ACT, the expected score on the GPA is 0.102197802197802

Solution c:

For each additional point on the ACT, the score on the GPA is expected to increase by 0.568131868131867

2.1

Solution a:

```
n <- length(df$GPA)
```

n = 8 -> since we have 8 students

Generel:

Residuals: are the difference between the line and the actual points $\widehat{u}_i = y_i - \widehat{y}_i$

Solution b:

```
df$Residuals <- df$GPA - df$GPA_hat
sumOfRes <- sum(df$Residuals)
```

sum of Residuals = 4.44089209850063e-16 ~ 0 -> close to 0

Solution c:

```
val20 <- (beta2 * 20) + beta1
```

val20 = 2.61208791208791

2.3

```
wagedata = read.table("wagedata.csv")
model2 <- lm(wage ~ educ, data = wagedata)
summary(model2)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = wagedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3396 -2.1501 -0.9674  1.1921 16.6085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.90485    0.68497  -1.321    0.187
## educ         0.54136    0.05325  10.167   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.378 on 524 degrees of freedom
## Multiple R-squared:  0.1648, Adjusted R-squared:  0.1632
## F-statistic: 103.4 on 1 and 524 DF,  p-value: < 2.2e-16
```

$\widehat{\beta}_1$ = -0.90485 -> intercept parameter

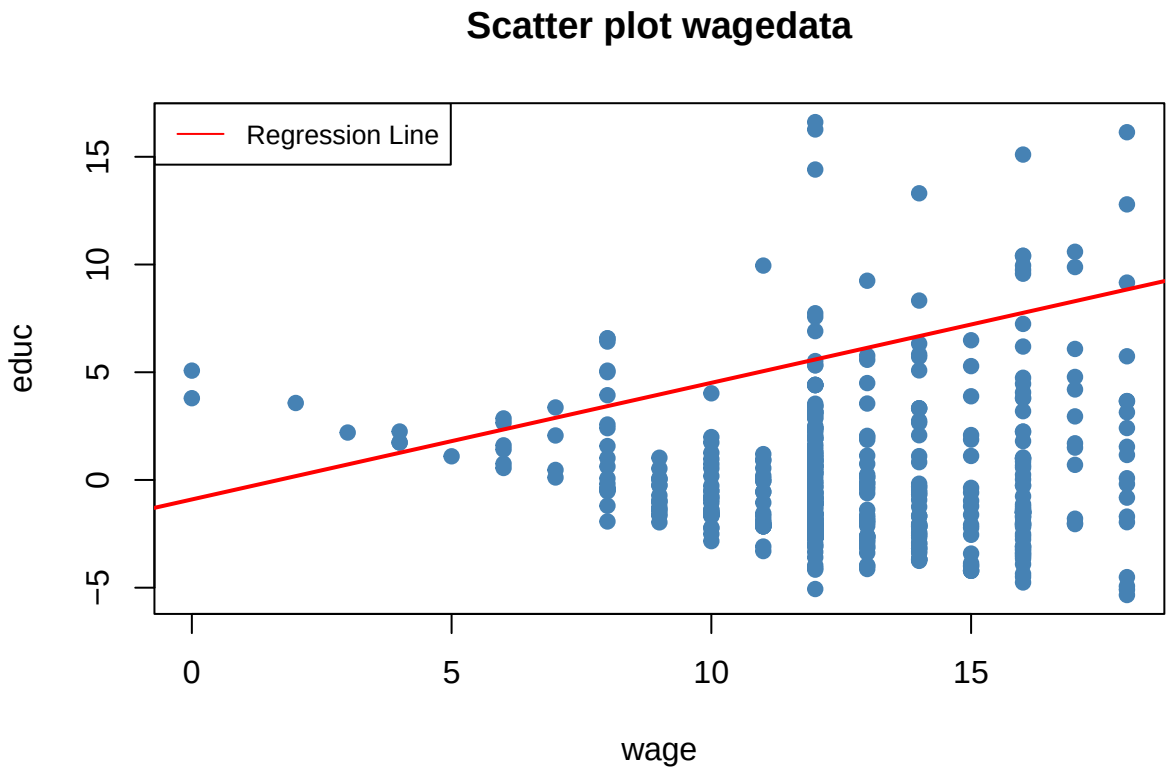
$\widehat{\beta}_2 = 0.54136$ -> slope parameter

Solution a:

Since the slope parameter is positive: The higher someones education value is, the higher is their wage.

Solution b:

```
wagedata$y_hat <- model2[["coefficients"]][["educ"]] * wagedata$educ +
  model2[["coefficients"]][["(Intercept)"]]
wagedata$u_hat <- wagedata$wage - wagedata$y_hat
plot(wagedata$educ, wagedata$u_hat, main="Scatter plot wagedata",
     xlab = "wage", ylab = "educ", pch = 19, col = "steelblue")
legend("topleft", legend=c("Regression Line"), col=c("red"), lty=1:1, cex=0.8)
abline(model2, col='red', lwd=2)
```



2.4

You can find it under Multiple R-squared: 0.1648

The R-squared measure is between 0 and 1 where 0 means none of the variance is explained by the predictor variable and 1 means 100% of the variance is explained by the predictor variable.

$S_(yy) = \frac{1}{n} \sum_{i=1}^{n} y_i(y_i - \bar{y})$