Tutorial 3

Zayd

24/01/2022

Swiss Fertility Data

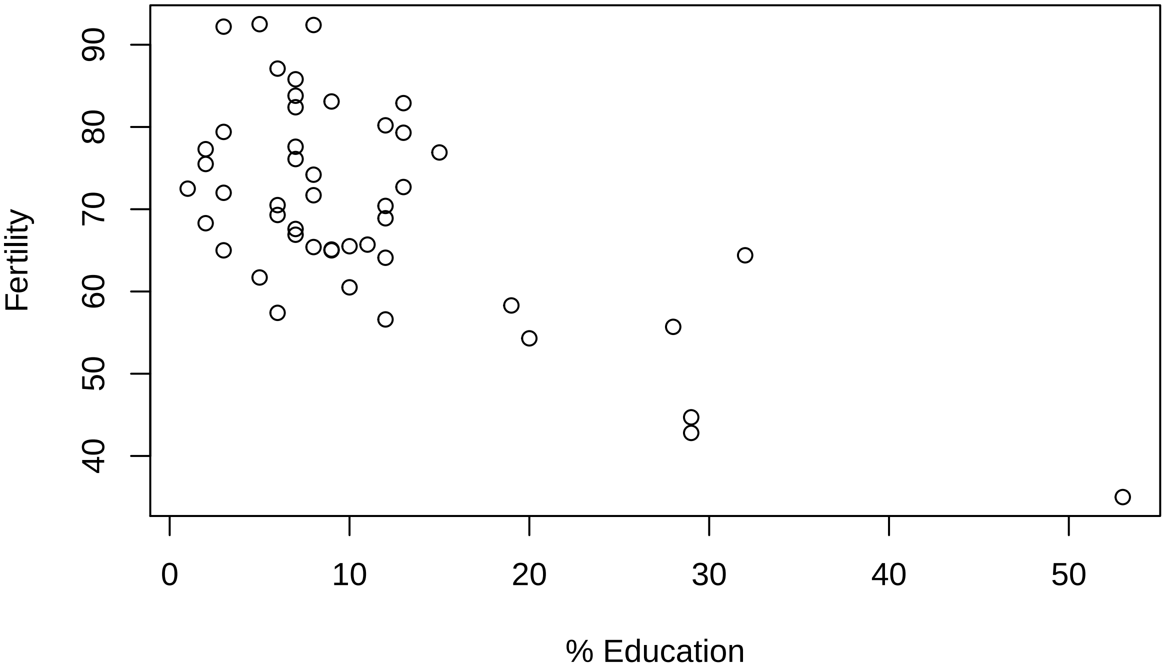
```
data_swiss = read.csv("/Users/zaydomar/math_204_w22/Tutorials/Tutorial 3/swiss.csv")

fertility = data_swiss$Fertility
education = data_swiss$Education
n = length(fertility)
```

Model Fitting

As always plot the data and do a quick visual exploration.

```
plot(education, fertility, xlab = "% Education", ylab = "Fertility")
```



```
##
## Call:
## lm(formula = fertility ~ education)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.036  -6.711  -1.011   9.526  19.689
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.6101     2.1041  37.836  < 2e-16 ***
## education    -0.8624     0.1448  -5.954 3.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.446 on 45 degrees of freedom
## Multiple R-squared:  0.4406, Adjusted R-squared:  0.4282
## F-statistic: 35.45 on 1 and 45 DF,  p-value: 3.659e-07
```

Hypothesis Test

To do the hypothesis test we first calculate the test-statistic. Which is given by,

$$T = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} = \frac{-0.8624}{0.1448} = -5.954.$$

This is a statistic that we have to be able to calculate by hand, however, notice, if we are "allowed" to use the summary function then we can also get the exact same summary statistic. The rejection region is given by, the following code or can be obtained from the t-table. Remember that in this case we have a two sided test so we need to find the upper and lower rejection region and then see whether or not the test-statistic lies in side the region or not.

```
RR_upper = qt(p = 0.975, df = n-2)
RR_lower = qt(p = 0.025, df = n-2)

# qnorm(p = 0.975)    # Normal dist upper region
# qnorm(p = 0.025)    # Normal dist lower region

# The rejection region is given by,
RR_lower
```

```
## [1] -2.014103
```

```
RR_upper
```

```
## [1] 2.014103
```

Our statistic clearly lies beyond the region and hence we reject the null hypothesis.

What if we failed to reject the null-hypothesis?

Confidence interval

We want to calculate the 95% confidence interval for $\hat{\beta}_1$. First we calculate this by hand using our formulas (we need to be able to do these by hand). We use the formula from the slides and a t-table or the computer to get the critical value.

$$-0.8624 \pm 2.014 \times 0.1448 = [-1.154, -0.571].$$

```
confint(fit,level = 0.99)
```

```
##                  0.5 %     99.5 %
## (Intercept) 73.950910 85.2692066
## education   -1.251922 -0.4727781
```

Correlation

The r and r^2 can be computed by hand or using the `summary()` function. Note that the function only provides the coefficient of determination, i.e r^2 and we can use this to obtain the value of r.

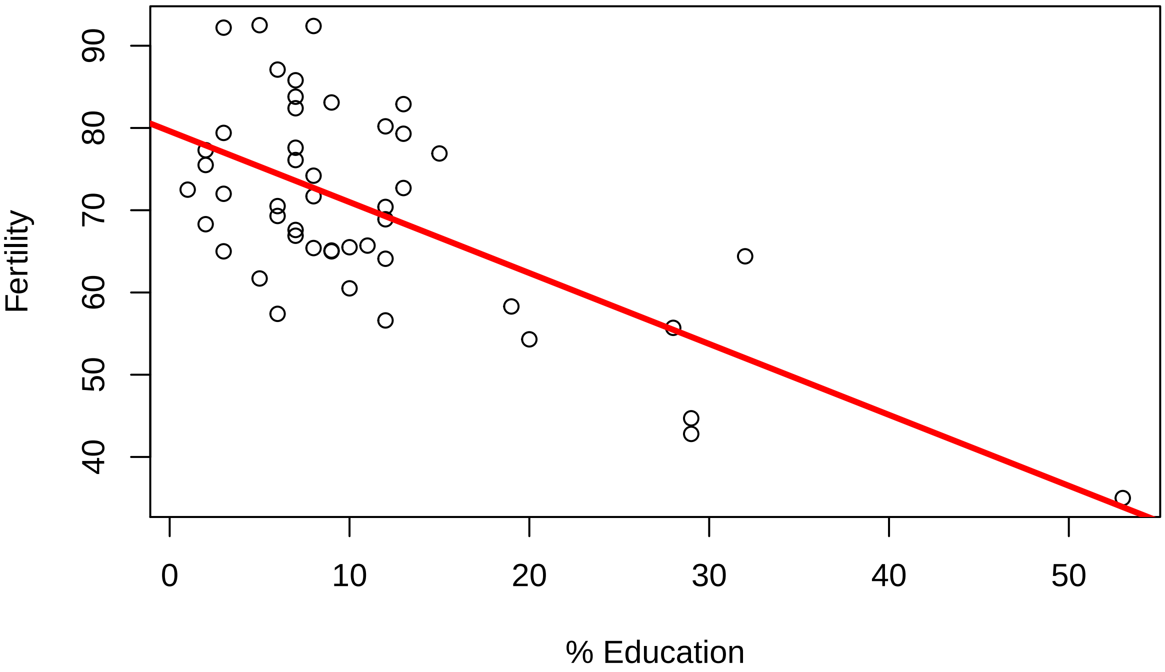
```
summary(fit)
```

```
## 
## Call:
## lm(formula = fertility ~ education)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -17.036  -6.711  -1.011   9.526  19.689 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  79.6101     2.1041  37.836  < 2e-16 ***
## education    -0.8624     0.1448  -5.954 3.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9.446 on 45 degrees of freedom
## Multiple R-squared:  0.4406, Adjusted R-squared:  0.4282 
## F-statistic: 35.45 on 1 and 45 DF,  p-value: 3.659e-07
```

$r^2 = 0.4406$, implying that about 44% of the variation in *Fertility* is explained by the *Education* variable. At the same time the correlation is given $r = \sqrt{r^2}$. But what is the correct sign of r? Do we have a positive correlation or a negative correlation? It is easy to find this out by looking at what is the sign in front of $\hat{\beta}_1$. In this case it is a negative sign and so we have negative correlation.

Fitted Line

```
plot(education, fertility, xlab = "% Education", ylab = "Fertility")
abline(a=fit$coefficients[1],b=fit$coefficients[2],col="red",lwd=3)
```



Confidence Interval and Prediction Inveral

For $Education = 15$ we have the estimate for the mean response is $\hat{y} = 79.61 - (0.8624)(15) = 66.674$. The confidence interval is given by,

$$\hat{y} \pm t_{\alpha/2}^{n-2} \times \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}.$$

The prediction interval is given by,

$$\hat{y} \pm t_{\alpha/2}^{n-2} \times \hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

```
mean_edu = mean(education)

SSxx = sum((education-mean_edu)^2)
simga_hat = summary(fit)$sigma
# predicted value

y_hat=predict(fit,newdata = data.frame(education=15))
y_hat
```

```
##       1
## 66.6748
```

```
# Confidence Interval
y_hat-2.014*simga_hat*sqrt(1/n+((15-mean_edu)^2)/SSxx)  # lower limit
```

```
##        1
## 63.66206
```

```
y_hat+2.014*simga_hat*sqrt(1/n+((15-mean_edu)^2)/SSxx)  # upper limit
```

```
##        1
## 69.68755
```

```
# Prediction Interval
y_hat-2.014*simga_hat*sqrt(1+1/n+((15-mean_edu)^2)/SSxx)  # lower limit
```

```
##        1
## 47.41343
```

```
y_hat+2.014*simga_hat*sqrt(1+1/n+((15-mean_edu)^2)/SSxx)  # upper limit
```

```
##        1
## 85.93618
```

```
## Or using the predict function we have that
predict(fit,newdata = data.frame(education=15),interval = "confidence")
```

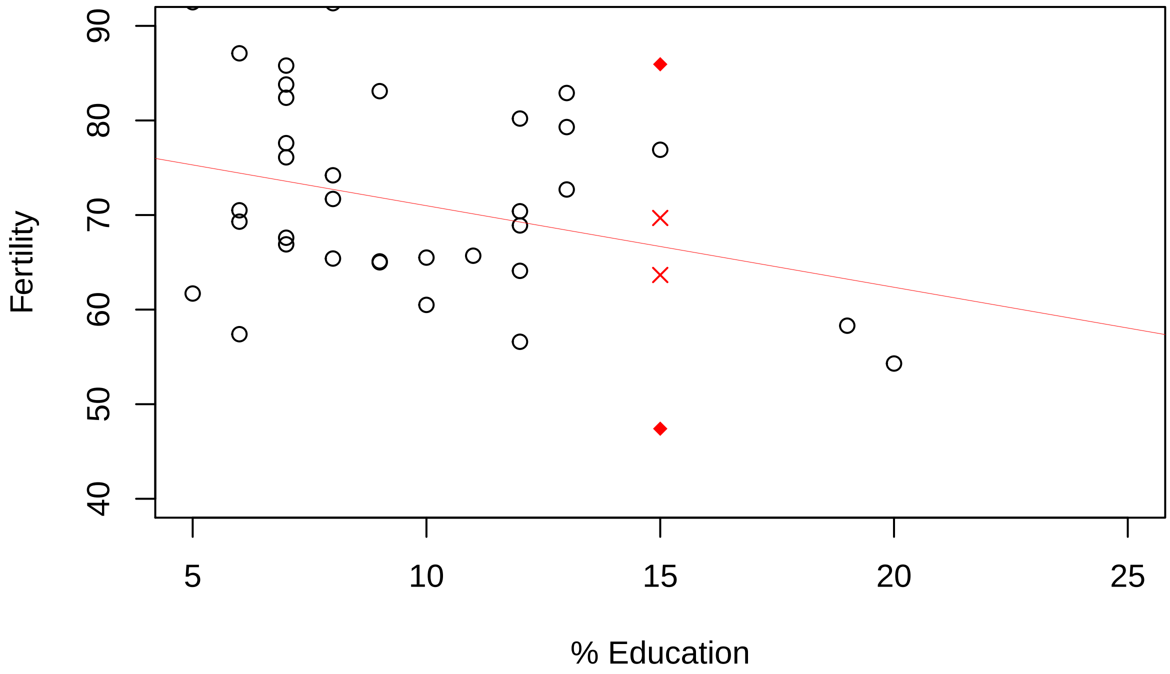
```
##       fit      lwr     upr
## 1 66.6748 63.66191 69.6877
```

```
predict(fit,newdata = data.frame(education=15),interval = "prediction")
```

```
##       fit      lwr      upr
## 1 66.6748 47.41244 85.93717
```

The CI and the PI look like this on the graph.

```
plot(education, fertility, xlab = "% Education", ylab = "Fertility", xlim = c(5,25),ylim=c(40,90))
abline(a=fit$coefficients[1],b=fit$coefficients[2],col="red",lwd=0.2)
points(c(15,15),c(63.66,69.69), pch = 4, col = "red")
points(c(15,15),c(47.41,85.94), pch = 18, col = "red", lwd = 0.2)
```



Subsetting the Data

This is an extra section. In this section I do some data subsetting using the snow-geese data set. I do th same thing using a few different functions so that you can choose any one you want to try, there is no one way to get the correct answer.

```
data_geese = read.csv("/Users/zaydomar/Dropbox/Zayd/Class_Notes/MATH 204 Principles of Statistics/Asn2/S

# See the following three ways to subset the data


## OPTION1: Using the dplyr library
data_subset1 = filter(data_geese,Diet != "Chow")

## OPTION2: Using Base R
data_subset2 = data_geese[data_geese$Diet != "Chow", ]

## OPTION3: Using Base R
data_subset3 = subset(data_geese, Diet != "Chow")
```

Test the datasets and you will see that they will all give you the same result. Remember, to use the `filter()` from *dplyr*, you **NEED** to have installed and attached the *dplyr* library. If you are not able to install this, you can use the other two functions provided.