assignment2

Sidra Effendi

20/02/2021

```
#knitr::opts_knit$set(root.dir = normalizePath("/Users/sidraeffendi/Desktop/PP639/"))
library(haven)
getwd()
```

```
## [1] "/Users/sidraeffendi/Desktop/PP639"
```

```
pset2 <- read_dta("/Users/sidraeffendi/Desktop/PP639/pset2.dta")
head(pset2)
```

```
## # A tibble: 6 x 11
##   province city  county urban yrsed  farmer drate_esoph water_grade airpollution
##   <chr>    <chr> <chr>  <dbl> <dbl>   <dbl>       <dbl>       <dbl>        <dbl>
## 1 Yunnan   Dali~ Xiang~     0  2.96 0.983         0.689        1           0.180
## 2 Guizhou  Qian~ Dushan     0  3.01 0.876         0.306        1           0.400
## 3 Guangxi  Guil~ Xiang~     1  6.53 0.00223       6.53         1           0.545
## 4 Guangxi  Hech~ Luoch~     0  3.00 0.713        12.0          1.33        0.411
## 5 Hubei    Xian~ Xiang~     0  1.16 0.932        20.4          1.5         0.371
## 6 Hubei    Xian~ Guche~     0  2.93 0.972        12.2          1.5         0.371
## # ... with 2 more variables: clean_river <dbl>, dirty_river <dbl>
```

```
pset2
```

```
## # A tibble: 145 x 11
##    province city  county urban yrsed  farmer drate_esoph water_grade
##    <chr>    <chr> <chr>  <dbl> <dbl>   <dbl>       <dbl>       <dbl>
##  1 Yunnan   Dali~ Xiang~     0 2.96  0.983         0.689        1
##  2 Guizhou  Qian~ Dushan     0 3.01  0.876         0.306        1
##  3 Guangxi  Guil~ Xiang~     1 6.53  0.00223       6.53         1
##  4 Guangxi  Hech~ Luoch~     0 3.00  0.713        12.0          1.33
##  5 Hubei    Xian~ Xiang~     0 1.16  0.932        20.4          1.5
##  6 Hubei    Xian~ Guche~     0 2.93  0.972        12.2          1.5
##  7 Yunnan   Yuxi  Tongh~     0 3.24  0.690         0.614        1.67
##  8 Xicang   Lasa  Mozhu~     0 0.384 0.887         0.939        2
##  9 Liaoning Dand~ Fengc~     0 2.94  0.780         1.81         2
## 10 Hebei    Chen~ Fengn~     0 3.42  0.955         2.28         2
## # ... with 135 more rows, and 3 more variables: airpollution <dbl>,
## #   clean_river <dbl>, dirty_river <dbl>
```

```
getwd()
```

```
## [1] "/Users/sidraeffendi/Desktop/PP639"
```

Q.1

(a)

```
# measures of spread
summary(pset2$drate_esoph)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.801   7.836  13.157  16.108  91.978
```

The mean of death rate from esophageal cancer is 13.157 and the median is 7.836. The minimum death rate from esophageal cancer is 0 while as the max is 91.98.

```
# get the percentile information
quantile(pset2$drate_esoph)
```

```
##        0%       25%       50%       75%      100%
##  0.000000  3.800585  7.836037 16.108120 91.978317
```

The 25th percentile for death rate from esophageal cancer is 3.8, the 50th paercentile is 7.836 and the 75th percentile is 16.108

(b)

```
# apply mask for clean and dirty rivers
clean_rivers <- pset2[pset2$clean_river == 1,]
dirty_rivers <- pset2[pset2$dirty_river == 1,]
#head(clean_rivers)
#head(dirty_rivers)
```

```
# Select numeric columns for clean rivers
clean_rivers.numcols <- clean_rivers[, sapply(clean_rivers, is.numeric)]
# find the mean
colMeans(clean_rivers.numcols)
```

```
##        urban        yrsed       farmer  drate_esoph  water_grade airpollution
##    0.3255814    4.4043368    0.5440947   10.3238502    3.1631449    0.4640643
##  clean_river  dirty_river
##    1.0000000    0.0000000
```

```r
# Select numeric columns dirty rivers
dirty_rivers.numcols <- dirty_rivers[, sapply(dirty_rivers, is.numeric)]
# find the mean
colMeans(dirty_rivers.numcols)
```

```
##        urban        yrsed       farmer  drate_esoph  water_grade airpollution
##    0.3728814    4.2410474    0.5678236   17.2868135    4.4071044    0.4916754
##  clean_river  dirty_river
##    0.0000000    1.0000000
```

When the rivers are clean we see the water quality is

(c)

```r
# apply mask for urban and non-urban areas
urban <- pset2[pset2$urban == 1,]
non_urban <- pset2[pset2$urban == 0,]
```

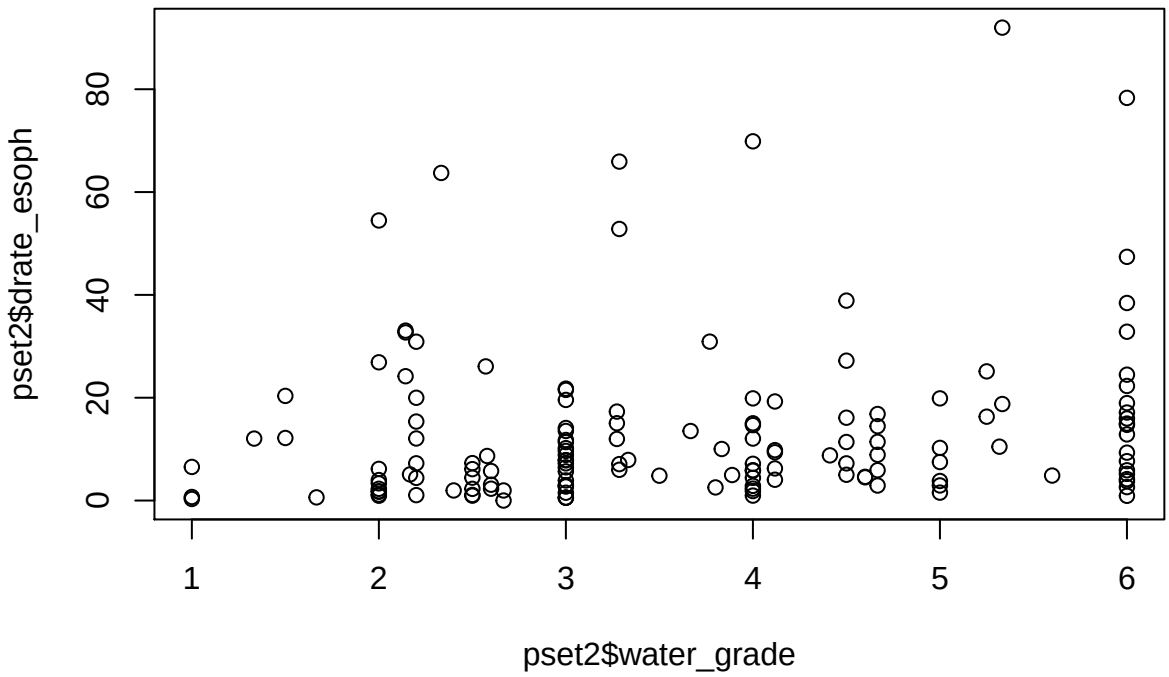
```r
# Select numeric columns for urban areas
urban.numcols <- urban[, sapply(urban, is.numeric)]
# find the mean
colMeans(urban.numcols)
```

```
##        urban        yrsed       farmer  drate_esoph  water_grade airpollution
##   1.00000000   6.06150380   0.01676372   8.04595296   3.77922112   0.49251063
##  clean_river  dirty_river
##   0.56000000   0.44000000
```

```r
# Select numeric columns for non-urban areas
non_urban.numcols <- non_urban[, sapply(non_urban, is.numeric)]
# find the mean
colMeans(non_urban.numcols)
```

```
##        urban        yrsed       farmer  drate_esoph  water_grade airpollution
##    0.0000000    3.4307323    0.8363742   15.8471101    3.6114586    0.4662405
##  clean_river  dirty_river
##    0.6105263    0.3894737
```

(d)



(e) Yes,there appears to be a relationship between esophageal cancer and water quality. The higher the water grade, the worse is the water quality and higher the death rate due to esophageal cancer. For water grade above 5 we see the death rate going 80 and above, while as for water less than 2 the death rate from esophageal cancer is at most 20 (approximately).

Q.2 Regression analysis I

(a)

```
# regress death rate from esophageal cancer against water grade.
linearMod <- lm(pset2$water_grade ~ pset2$drate_esoph, data=pset2)
summary(linearMod)
```

```
##
## Call:
## lm(formula = pset2$water_grade ~ pset2$drate_esoph, data = pset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5702 -1.0064 -0.2927  1.0234  2.5138
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.472421   0.148444  23.392   <2e-16 ***
## pset2$drate_esoph 0.014964   0.007279   2.056   0.0416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.366 on 143 degrees of freedom
## Multiple R-squared:  0.0287, Adjusted R-squared:  0.02191
## F-statistic: 4.226 on 1 and 143 DF,  p-value: 0.04163
```

$$\hat{drate\_esoph}_i = 0.015 \ + \ 3.472 \ * \ water\_grade_i$$

b(i)

$$\hat{\beta_1} = 3.472$$

On average, for a unit increase in the value of water grade, there is an expected increase of 3.472 in the death rate from esophageal cancer.

b(ii)

$$SE(\hat{\beta_1}) = 0.148$$

For the different samples of a population we get different values of

$$\hat{\beta_1}$$

. So, the standard error of the slope represents how much far apart do the the slope values lie from the mean. This is supposed to be a normal distribution with the mean at the center.

b(iii)

$$T\_stat \ for \ the \ test \ of \ null \ hypothesis \ that \ \hat{\beta_1} = 0 \ is \ 23.392$$

With 95% confidence, since the test statistic is much greater than the critical value of 1.96, we reject the null hypothesis that

$$\hat{\beta_1} = 0$$

.

b(iv)

95% confidence interval =

$$[mean(\hat{\beta_1}) - 1.96 * SE(\hat{\beta_1}), mean(\hat{\beta_1}) + 1.96 * SE(\hat{\beta_1})]$$

=

```
res <- t.test(pset2$drate_esoph, pset2$water_grade, data = pset2)
res
```

```
##
##  Welch Two Sample t-test
##
## data:  pset2$drate_esoph and pset2$water_grade
## t = 7.2792, df = 146.25, p-value = 1.904e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.911812 12.063684
## sample estimates:
## mean of x mean of y
## 13.157056  3.669308
```

We are confident that the actual value of the slope,

$$\beta_1$$

lies between 6.912 and 12.064.

b(v)

$$\hat{\beta_0} = 0.015$$

On average, the expected value of death rates due to esophageal cancer when the water-grade value is 0, is 0.015 ## b(vi) The no.of observations =

b(vii)

$$R^2 \ = \ 0.0287$$

Water grade is able to explain 2.87% of the variance in the death rates due to esophageal cancer.

(c)

For a site with water grade = 3

$$\hat{drate\_esoph_i} = 0.015 \ + \ 3.472 \ * \ 3$$

$= 0.015 + =$

For a site with water grade = 5

$$\hat{drate\_esoph_i} = 0.015 \ + \ 3.472 \ * \ 5$$

$= 0.015 + =$

Regression analysis II

(a)

```
# regress death rate from esophageal cancer against water grade.
linearMod <- lm(pset2$water_grade ~ pset2$drate_esoph, data=pset2)
summary(linearMod)
```

```
##
## Call:
## lm(formula = pset2$water_grade ~ pset2$drate_esoph, data = pset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5702 -1.0064 -0.2927  1.0234  2.5138
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.472421   0.148444  23.392   <2e-16 ***
## pset2$drate_esoph 0.014964   0.007279   2.056   0.0416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.366 on 143 degrees of freedom
## Multiple R-squared:  0.0287, Adjusted R-squared:  0.02191
## F-statistic: 4.226 on 1 and 143 DF,  p-value: 0.04163
```

(b)

THe omitted category is...

c(i)

$$\hat{\beta_1} = 3.472$$

On average, for a unit increase in the value of water grade, there is an expected increase of 3.472 in the death rate from esophageal cancer.

c(ii)

$$SE(\hat{\beta_1}) = 0.148$$

For the different samples of a population we get different values of

$$\hat{\beta_1}$$

. So, the standard error of the slope represents how much far apart do the the slope values lie from the mean. This is supposed to be a normal distribution with the mean at the center.

c(iii)

$$T\_stat\ for\ the\ test\ of\ null\ hypothesis\ that\ \hat{\beta_1} = 0\ is\ 23.392$$

With 95% confidence, since the test statistic is much greater than the critical value of 1.96, we reject the null hypothesis that

$$\hat{\beta_1} = 0$$

.

c(iv)

95% confidence interval =

$$[mean(\hat{\beta}_1) - 1.96 * SE(\hat{\beta}_1), mean(\hat{\beta}_1) + 1.96 * SE(\hat{\beta}_1)]$$

=

```
res <- t.test(pset2$drate_esoph, pset2$water_grade, data = pset2)
res
```

```
##
##  Welch Two Sample t-test
##
## data:  pset2$drate_esoph and pset2$water_grade
## t = 7.2792, df = 146.25, p-value = 1.904e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.911812 12.063684
## sample estimates:
## mean of x mean of y
## 13.157056  3.669308
```

We are confident that the actual value of the slope,

$$\beta_1$$

lies between 6.912 and 12.064.

c(v)

$$\hat{\beta_0} = 0.015$$

On average, the expected value of death rates due to esophageal cancer when the water-grade value is 0, is 0.015 # c(vi) The no.of observations =

c(vii)

$$R^2 \ = \ 0.0287$$

Water grade is able to explain 2.87% of the variance in the death rates due to esophageal cancer.

(d)

For a site with water grade = 3

$$\hat{drate\_esoph_i} = 0.015 \ + \ 3.472 \ * \ 3$$

= 0.015 + =

For a site with water grade = 5

$$\hat{drate\_esoph_i} = 0.015 \ + \ 3.472 \ * \ 5$$

= 0.015 + =

Q.4

Q.5

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

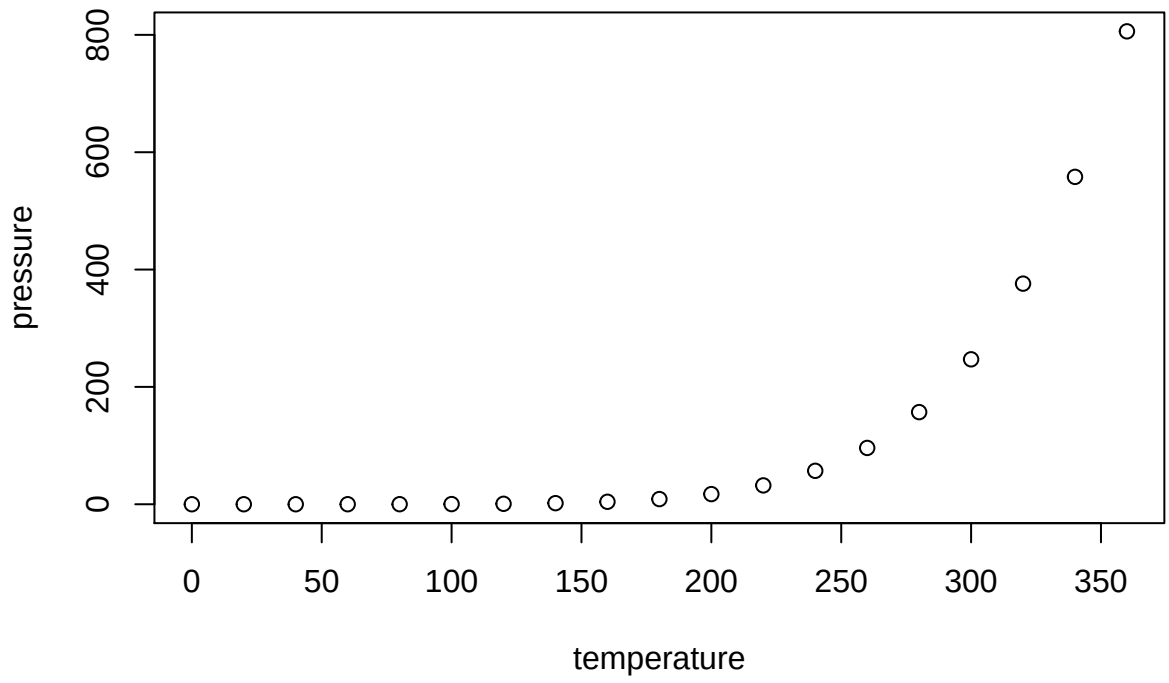
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.