DASC-5301-ASSIGNMENT

NIVAS

2024-02-21

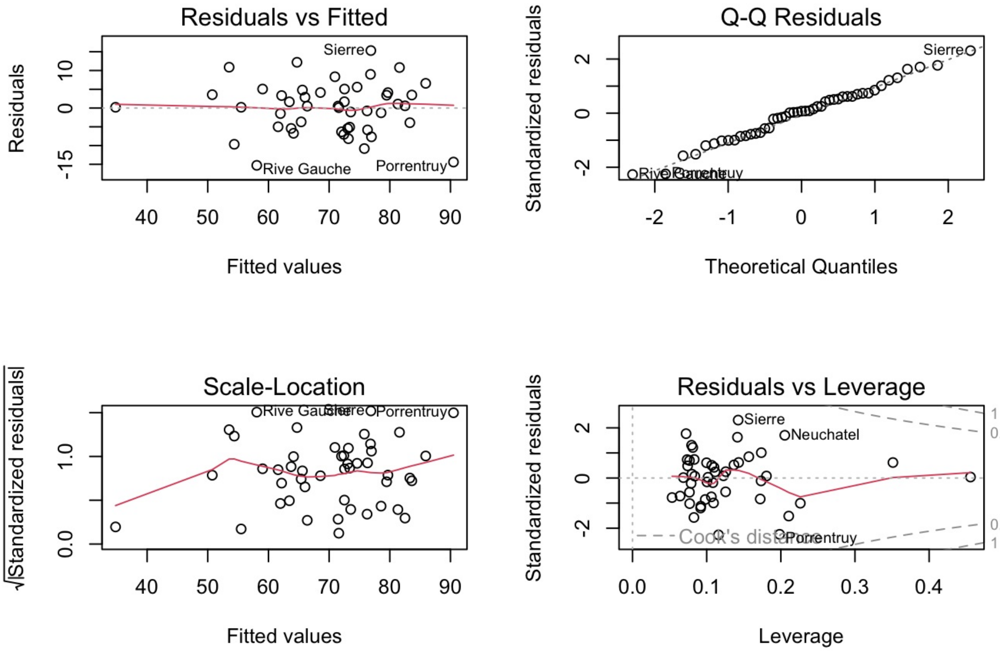
```
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1
## ✔ purrr     1.0.2     ✔ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
mydata<-datasets::swiss
mydata
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
## Broye             83.8        70.2          16         7    92.85
## Glane             92.4        67.8          14         8    97.16
## Gruyere           82.4        53.3          12         7    97.67
## Sarine            82.9        45.2          16        13    91.38
## Veveyse           87.1        64.5          14         6    98.61
## Aigle             64.1        62.0          21        12     8.52
## Aubonne           66.9        67.5          14         7     2.27
## Avenches          68.9        60.7          19        12     4.43
## Cossonay          61.7        69.3          22         5     2.82
## Echallens         68.3        72.6          18         2    24.20
## Grandson          71.7        34.0          17         8     3.30
## Lausanne          55.7        19.4          26        28    12.11
## La Vallee         54.3        15.2          31        20     2.15
## Lavaux            65.1        73.0          19         9     2.84
## Morges            65.5        59.8          22        10     5.23
## Moudon            65.0        55.1          14         3     4.52
## Nyone             56.6        50.9          22        12    15.14
## Orbe              57.4        54.1          20         6     4.20
## Oron              72.5        71.2          12         1     2.40
## Payerne           74.2        58.1          14         8     5.23
## Paysd'enhaut      72.0        63.5           6         3     2.56
## Rolle             60.5        60.8          16        10     7.72
## Vevey             58.3        26.8          25        19    18.46
## Yverdon           65.4        49.5          15         8     6.10
## Conthey           75.5        85.9           3         2    99.71
## Entremont         69.3        84.9           7         6    99.68
## Herens            77.3        89.7           5         2   100.00
## Martigwy          70.5        78.2          12         6    98.96
## Monthey           79.4        64.9           7         3    98.22
## St Maurice        65.0        75.9           9         9    99.06
## Sierre            92.2        84.6           3         3    99.46
## Sion              79.3        63.1          13        13    96.83
## Boudry            70.4        38.4          26        12     5.62
## La Chauxdfnd      65.7         7.7          29        11    13.79
## Le Locle          72.7        16.7          22        13    11.22
## Neuchatel         64.4        17.6          35        32    16.92
## Val de Ruz        77.6        37.6          15         7     4.97
## ValdeTravers      67.6        18.7          25         7     8.65
## V. De Geneve      35.0         1.2          37        53    42.34
## Rive Droite       44.7        46.6          16        29    50.43
## Rive Gauche       42.8        27.7          22        29    58.33
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
## Broye                    23.6
## Glane                    24.9
## Gruyere                  21.0
## Sarine                   24.4
## Veveyse                  24.5
## Aigle                    16.5
## Aubonne                  19.1
## Avenches                 22.7
## Cossonay                 18.7
## Echallens                21.2
## Grandson                 20.0
## Lausanne                 20.2
## La Vallee                10.8
## Lavaux                   20.0
## Morges                   18.0
## Moudon                   22.4
## Nyone                    16.7
## Orbe                     15.3
## Oron                     21.0
## Payerne                  23.8
## Paysd'enhaut             18.0
## Rolle                    16.3
## Vevey                    20.9
## Yverdon                  22.5
## Conthey                  15.1
## Entremont                19.8
## Herens                   18.3
## Martigwy                 19.4
## Monthey                  20.2
## St Maurice               17.8
## Sierre                   16.3
## Sion                     18.1
## Boudry                   20.3
## La Chauxdfnd             20.5
## Le Locle                 18.9
## Neuchatel                23.0
## Val de Ruz               20.0
## ValdeTravers             19.5
## V. De Geneve             18.0
## Rive Droite              18.2
## Rive Gauche              19.3
```

```
model1 <- lm(Fertility~., data = mydata)
par(mfrow = c(2, 2))
plot(model1)
```



```
#How do you interpret the intercept of the model?
coefficients <- coef(model1)
coefficients
```

```
##      (Intercept)      Agriculture      Examination        Education
##       66.9151817       -0.1721140       -0.2580082       -0.8709401
##         Catholic Infant.Mortality
##        0.1041153        1.0770481
```

```
intercept <- coefficients["(Intercept)"]
intercept
```

```
## (Intercept)
##    66.91518
```

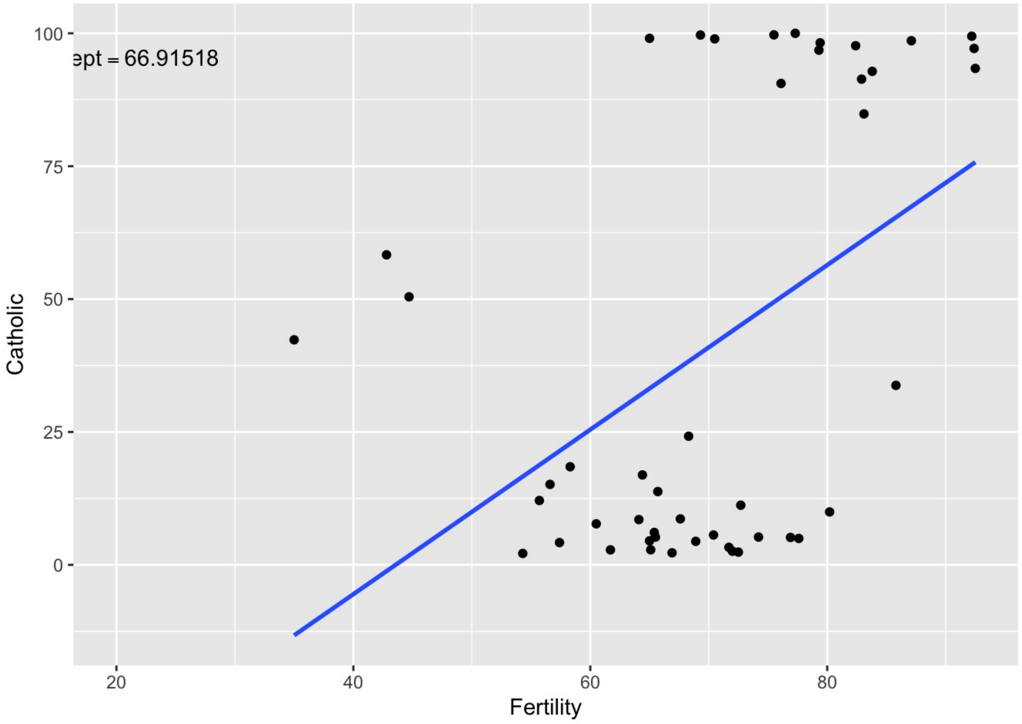
```
summary(model1) ; summary(model1)$coefficients[,1]
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

```
##      (Intercept)      Agriculture      Examination        Education
##       66.9151817       -0.1721140       -0.2580082       -0.8709401
##         Catholic Infant.Mortality
##        0.1041153        1.0770481
```

```
#Plot intercept
ggplot(mydata,aes(Fertility,Catholic))+geom_point()+stat_smooth(method="lm",se=F)+
  annotate("text",x=20,y=95,label=(paste0("Intercept==",coef(model1)[1])),parse=TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
lm(formula = Fertility~. , data = mydata)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = mydata)
##
## Coefficients:
##      (Intercept)       Agriculture       Examination         Education
##          66.9152           -0.1721           -0.2580           -0.8709
##         Catholic  Infant.Mortality
##           0.1041            1.0770
```

```
# How do you interpret each variable’s coefficient in the model (5 interpretations)?
for (coefficient_name in names(coefficients)) {
  cat(coefficient_name, coefficients[coefficient_name],"\n")
}
```

```
## (Intercept) 66.91518
## Agriculture -0.172114
## Examination -0.2580082
## Education -0.8709401
## Catholic 0.1041153
## Infant.Mortality 1.077048
```

```
#summary of the model
summary_model <- summary(model1)
summary_model
```

```
## 
## Call:
## lm(formula = Fertility ~ ., data = mydata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -15.2743  -5.2617   0.5032   4.1198  15.3213 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *  
## Examination      -0.25801    0.25388  -1.016  0.31546    
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 ** 
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671 
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

```
#p_value
p_values <- summary_model$coefficients[, 4]
p_values
```

```
##      (Intercept)      Agriculture      Examination        Education 
##     1.906051e-07     1.872715e-02     3.154617e-01     2.430605e-05 
##         Catholic Infant.Mortality 
##     5.190079e-03     7.335715e-03 
```

```
variable_names <- rownames(summary_model$coefficients)
variable_names
```

```
## [1] "(Intercept)"      "Agriculture"      "Examination"      "Education"       
## [5] "Catholic"         "Infant.Mortality"
```

```
significant_variables <- c()

for (i in seq_along(variable_names)) {
  if (p_values[i] < 0.05) {
    significant_variables <- c(significant_variables, variable_names[i],"=", p_values[i], "\n")
  }
}

if (length(significant_variables) > 0) {
  cat(significant_variables, "\n")
} else {
  cat("No significant variables present in the model.\n")
}
```

```
## (Intercept) = 1.90605128792694e-07 
##  Agriculture = 0.0187271543851753 
##  Education = 2.43060459073792e-05 
##  Catholic = 0.00519007854516597 
##  Infant.Mortality = 0.00733571532060151 
## 
```

```
#RSquare - Co-efficicient of determination:
r_squared <- summary_model$r.squared
r_squared
```

```
## [1] 0.706735
```

```
#F-Stastics
summary_model$fstatistic
```

```
##    value    numdf    dendf 
## 19.76106  5.00000 41.00000 
```

```
f_statistic <- summary_model$fstatistic[1]
p_value <- summary_model$fstatistic[2]
cat("F-statistic:", f_statistic, "\n")
```

```
## F-statistic: 19.76106
```

```
cat("p-value:", p_value, "\n")
```

```
## p-value: 5
```

```
if (p_value < 0.05) {
  cat("\nThe F-statistic is statistically significant.\n")
} else {
  cat("\nThe F-statistic is not statistically significant \n")
}
```

```
## 
## The F-statistic is not statistically significant
```

```
#assumption holding
plot(model1, which = 1)
qqnorm(resid(model1))
qqline(resid(model1))

#Confident Interval

confident_interval_95 <- confint(model1, "Catholic", level = 0.95)
confident_interval_99 <- confint(model1, "Catholic", level = 0.99)
print(confident_interval_95)
```

```
##               2.5 %  97.5 %
## Catholic 0.03291065 0.17532
```

```
print(confident_interval_99)
```

```
##                0.5 %    99.5 %
## Catholic 0.008877479 0.1993532
```

```
#model2
model2 <- lm(Fertility ~ Catholic + Education, data = mydata)

# Summary of the model
summary(model2)
```

```
## 
## Call:
## lm(formula = Fertility ~ Catholic + Education, data = mydata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -15.042  -6.578  -1.431   6.122  14.322 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 74.23369    2.35197  31.562  < 2e-16 ***
## Catholic     0.11092    0.02981   3.721  0.00056 ***
## Education   -0.78833    0.12929  -6.097 2.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.331 on 44 degrees of freedom
## Multiple R-squared:  0.5745, Adjusted R-squared:  0.5552 
## F-statistic:  29.7 on 2 and 44 DF,  p-value: 6.849e-09
```

```
#adjusted r square

adjusted_r_square_model1 <- summary(model1)$adj.r.squared
adjusted_r_square_model2 <- summary(model2)$adj.r.squared
cat("Adjusted R-squared for model1:", adjusted_r_square_model1, "\n")
```

```
## Adjusted R-squared for model1: 0.670971
```

```
cat("Adjusted R-squared for model2:", adjusted_r_square_model2, "\n")
```

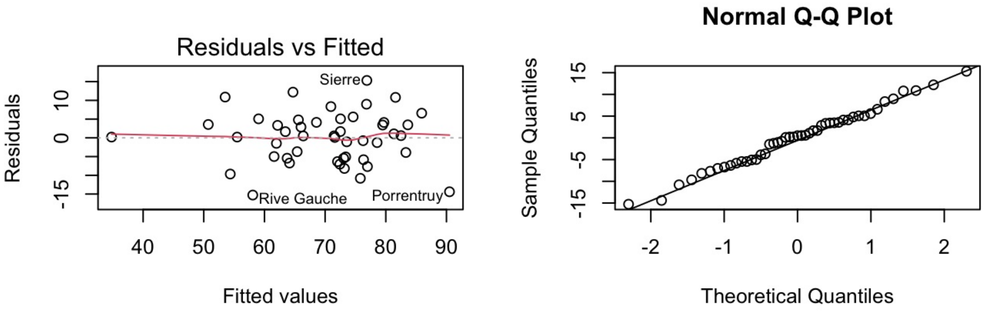
```
## Adjusted R-squared for model2: 0.5551665
```

```
if (adjusted_r_square_model1 > adjusted_r_square_model2) {
  cat("Model 1 is higher.\n")

} else if (adjusted_r_square_model1 < adjusted_r_square_model2) {
  cat("Model 2 is higher.\n")

} else {
  cat("Both models are same.\n")
}
```

```
## Model 1 is higher.
```

```
# Anova
anova_result <- anova(model2, model1)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ Catholic + Education
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     44 3054.2
## 2     41 2105.0  3    949.13 6.1621 0.001478 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residuals vs Fitted
Residuals
Fitted values
Sierre
Rive Gauche
Porrentruy
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles