HOMEWORK - VI

Melih Gündüz

2024-04-18

```
install.packages("DALEX")
install.packages("gridExtra")
install.packages("ggplot2")
library(DALEX)
library(gridExtra)
library(ggplot2)
```

1) DATASET

```
data(titanic)
str("titanic")
```

```
 chr "titanic"
```

```
titanic$age[is.na(titanic$age)] <- mean(titanic$age, na.rm = TRUE)
```

1.1) Regression

```
newdata <- titanic
newdata$survived <- ifelse(newdata$survived == "yes", 1, 0)
model <- lm(survived ~ age, data = newdata)
summary(model)
```

```
Call:
lm(formula = survived ~ age, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3767 -0.3302 -0.3103  0.6607  0.7387

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3773028  0.0268100  14.073   <2e-16 ***
age         -0.0018118  0.0008181  -2.215   0.0269 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.467 on 2205 degrees of freedom
Multiple R-squared:  0.00222,   Adjusted R-squared:  0.001767
F-statistic: 4.905 on 1 and 2205 DF,  p-value: 0.02688
```

The analysis, the age variable has a negative impact on survival, meaning that as age increases, the probability of survival decreases. This suggests that individuals in the younger age group have a higher probability of survival. The model has low explanatory power and limited accuracy. Therefore, it can be seen that the age variable alone is not sufficient to explain its impact on survival and that other independent variables are required.

1.2) Histogram

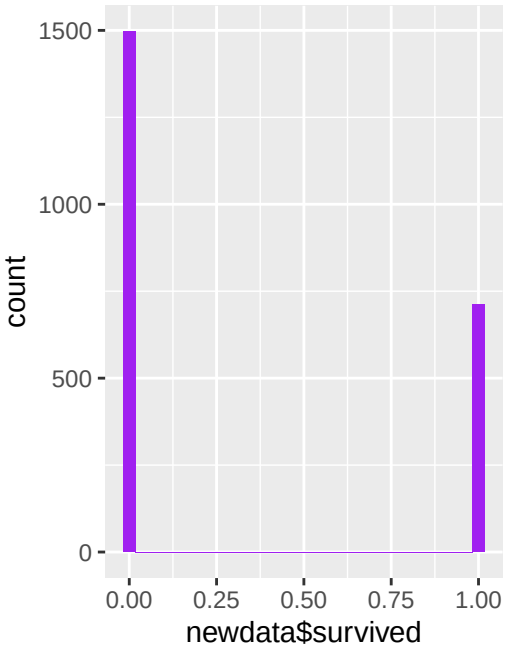
```
hist1 <- ggplot(newdata, aes(x=newdata$survived))+
  geom_histogram(fill="purple")
  hist2 <- ggplot(newdata, aes(x=newdata$age))+
  geom_histogram(fill="blue")
grid.arrange(hist1, hist2, ncol = 2)
```

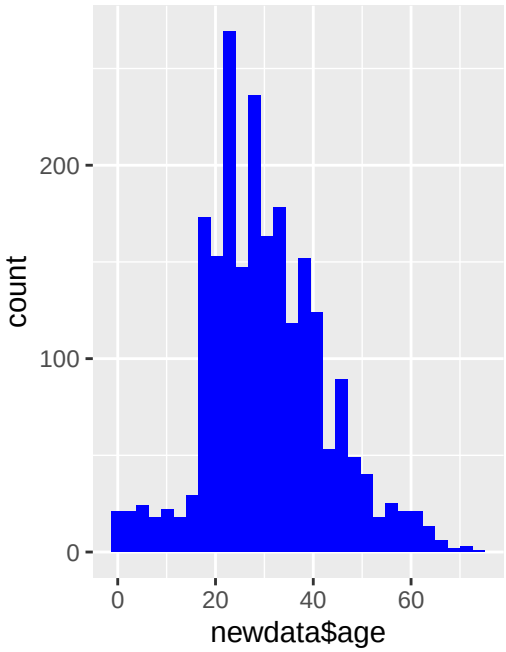
```
Warning: Use of `newdata$survived` is discouraged.
i Use `survived` instead.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning: Use of `newdata$age` is discouraged.
i Use `age` instead.
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





1.3) ANOVA

```r
anova <- aov(survived ~ age, data = newdata)
summary(anova)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
age            1    1.1  1.0698   4.905 0.0269 *
Residuals   2205  480.9  0.2181
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA results, the age variable is statistically significant in relation to survival, considering a critical value of 0.05. This indicates that the age variable explains a significant variance in the model, suggesting its importance as a determinant factor for survival.

1.4) ANCOVA

```r
ancova <- lm(survived ~ age + fare, data = newdata)
summary(ancova)
```

```
Call:
lm(formula = survived ~ age + fare, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0114 -0.3055 -0.2673  0.5964  0.8178

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3612452  0.0262054  13.785  < 2e-16 ***
age         -0.0029347  0.0008043  -3.649  0.00027 ***
fare         0.0026840  0.0002254  11.906  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4537 on 2178 degrees of freedom
  (26 observations deleted due to missingness)
Multiple R-squared:  0.06318,   Adjusted R-squared:  0.06232
F-statistic: 73.44 on 2 and 2178 DF,  p-value: < 2.2e-16
```

The age and fare variables are significant for survival. As age increases, the probability of survival decreases, while an increase in fare is associated with an increase in survival probability. However, the model has a low overall explanatory power, and other factors also influence survival.

2) DATASET

```r
set.seed(123)
x = 1:300
y = rnorm(100, mean=15, sd=2)
z = 2*x + y/3
data = data.frame(x, y, z)
summary(data)
```

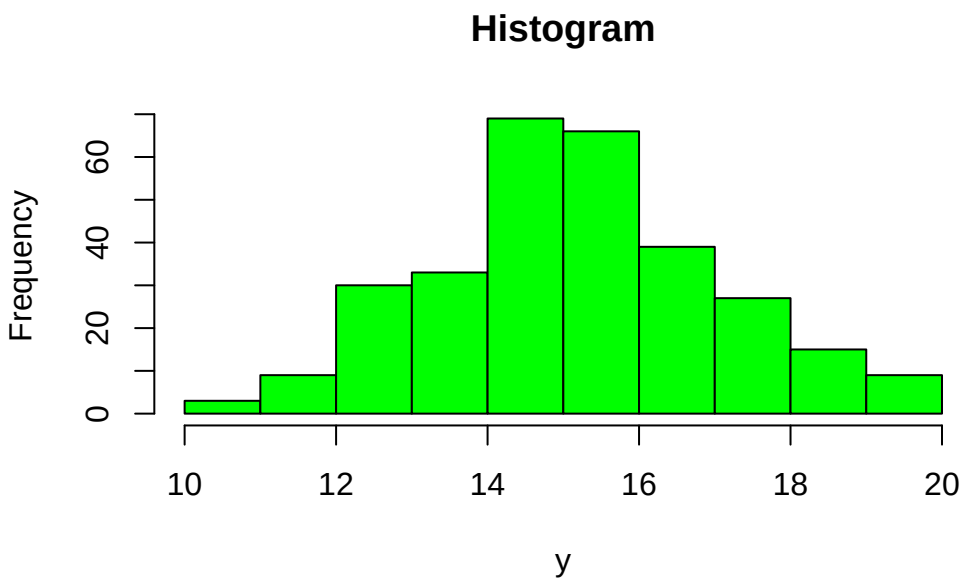
```
       x                y               z          
 Min.   :  1.00   Min.   :10.38   Min.   :  6.626  
 1st Qu.: 75.75   1st Qu.:14.01   1st Qu.:156.898  
 Median :150.50   Median :15.12   Median :306.057  
 Mean   :150.50   Mean   :15.18   Mean   :306.060  
 3rd Qu.:225.25   3rd Qu.:16.38   3rd Qu.:454.906  
 Max.   :300.00   Max.   :19.37   Max.   :604.316  
```

```
str(data)
```

```
'data.frame':   300 obs. of  3 variables:
 $ x: int  1 2 3 4 5 6 7 8 9 10 ...
 $ y: num  13.9 14.5 18.1 15.1 15.3 ...
 $ z: num  6.63 8.85 12.04 13.05 15.09 ...
```

2.1) Histogram

```
hist(data$y, main = "Histogram", xlab="y", col = "green")
```

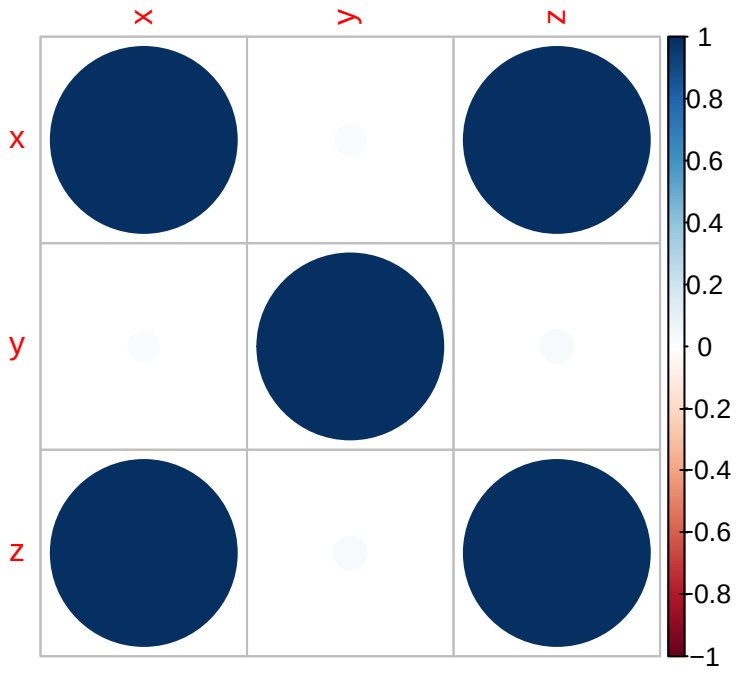


2.2) Corplot

```
corplot <- cor(data)
print(corplot)
```

```
           x          y          z
x 1.00000000 0.02660159 0.99999390
y 0.02660159 1.00000000 0.03009445
z 0.99999390 0.03009445 1.00000000
```

```
install.packages("corrplot")
library(corrplot)
corrplot(corplot, method = "circle")
```



2.3) Regression Model

```
model2 <- lm(z ~ y + x, data = data)
summary(model2)
```

```
Call:
lm(formula = z ~ y + x, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-1.010e-13 -2.745e-14 -1.387e-14  6.050e-15  1.534e-12

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 1.567e-13  5.922e-14 2.646e+00  0.00857 **
y           3.333e-01  3.813e-15 8.741e+13  < 2e-16 ***
x           2.000e+00  7.999e-17 2.500e+16  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.199e-13 on 297 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 3.129e+32 on 2 and 297 DF,  p-value: < 2.2e-16
```

The model results seem to show a perfect fit, but this is often unrealistic as it indicates overfitting to the data. The coefficients are also very large, suggesting that the model is fitting in an statistically insignificant manner. Therefore, it's important to adjust the model more balanced.

2.4) ANOVA

```r
anova2 <- aov(model2)
summary(anova2)
```

```
             Df  Sum Sq Mean Sq   F value Pr(>F)
y             1    8153    8153 5.668e+29 <2e-16 ***
x             1 8993531 8993531 6.252e+32 <2e-16 ***
Residuals   297       0       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This table shows that the variables "Y" and "X" have highly significant effects on the response. The small p-values indicate statistical significance, and the high "F value" suggests that the

model effectively explains the variability in the data. These results indicate a strong fit of the model and a good explanatory power for the data.

2.5) ANCOVA

```
ancova2 <- aov(model2)
summary(ancova2)
```

```
             Df  Sum Sq Mean Sq   F value Pr(>F)
y             1    8153    8153 5.668e+29 <2e-16 ***
x             1 8993531 8993531 6.252e+32 <2e-16 ***
Residuals   297       0       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results show that both predictors “x” and “y” have a highly significant impact on the response variable, with “x” having a stronger influence. The model fits the data well, as indicated by the low residual error.