Rozdział 6 - Regresja liniowa i wieloraka - ćwiczenia

## Regresja dla danych ilościowych

1. Wczytanie i przygotowanie danych oraz zamiana na system metyczny

rm(list=ls()) #usuwanie danych  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

Dane<-trees  
Dane<-Dane %>% transmute('Średnica'=Girth\*2.54,  
 'Wysokość'=Height\*2.54\*12/100,'Objętość'=Volume\*(2.54\*12/100)^3)

1. Wyznaczanie modelu regresji liniowej i jego optymalizacja

model<-lm(data = Dane,formula = Objętość~.) #regresja z uwzględnieniem wszystkich kolumn  
summary(model) #parametry modelu regresji

##   
## Call:  
## lm(formula = Objętość ~ ., data = Dane)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.181411 -0.075021 -0.008143 0.062306 0.240260   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.642028 0.244607 -6.713 2.75e-07 \*\*\*  
## Średnica 0.052488 0.002946 17.816 < 2e-16 \*\*\*  
## Wysokość 0.031517 0.012091 2.607 0.0145 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1099 on 28 degrees of freedom  
## Multiple R-squared: 0.948, Adjusted R-squared: 0.9442   
## F-statistic: 255 on 2 and 28 DF, p-value: < 2.2e-16

anova(model) #analiza dokładności modelu (ANOVA)

## Analysis of Variance Table  
##   
## Response: Objętość  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Średnica 1 6.0794 6.0794 503.1503 < 2e-16 \*\*\*  
## Wysokość 1 0.0821 0.0821 6.7943 0.01449 \*   
## Residuals 28 0.3383 0.0121   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

print(RMSE<-sqrt(mean((Dane$Objętość-model$fitted.values)^2))) #średni bład kwadratow

## [1] 0.1044672

step(model,direction = "both") #metoda krokowa optymalizajci modelu

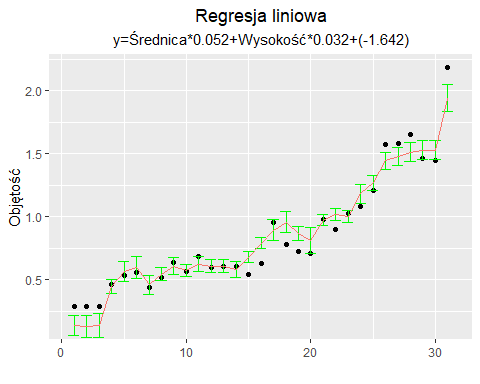
## Start: AIC=-134.05  
## Objętość ~ Średnica + Wysokość  
##   
## Df Sum of Sq RSS AIC  
## <none> 0.3383 -134.051  
## - Wysokość 1 0.0821 0.4204 -129.316  
## - Średnica 1 3.8352 4.1735 -58.162

##   
## Call:  
## lm(formula = Objętość ~ Średnica + Wysokość, data = Dane)  
##   
## Coefficients:  
## (Intercept) Średnica Wysokość   
## -1.64203 0.05249 0.03152

Tutaj warto zinterpretować współczynniki modelu oraz te, które określają jego dokładność. Prawdopodobieństwo ***p-value*** testu ***ANOVA*** jest mniejsze od 0,05, czyli odrzucamy hipotezę o nieistotności wszystkich parametrów modelu. Jest co najmniej jeden czynnik, który tłumaczy zmiany zmiennej objaśnianej, czyli dalsza analiz ma sens. ***Adjusted R-squared*** - skorygowany współczynnik determinacji jest na bardzo wysokim poziomie (prawie 95% zmienności danych da się wyjaśnić za pomocą modelu). Z testu T prawdopodobieństwo ***p-value*** dla każdego wyznaczonego parametru modelu jest mniejsze od 0,05, czyli każdy parametr jest istotny statystycznie, choć wymiar wysokość jest obarczona większym błędem. Przyrost o 1 cm średnicy to zwiększenie objętości średnio o 0,052 m3, a zmiana wysokości o 1 m to przyrost objętości o 0,032 m3.

3.Wizualizacja modelu regresji

library(ggplot2)  
rownanie<-paste("y=Średnica\*",round(model$coefficients[[2]],3),"+Wysokość\*",  
 round(model$coefficients[[3]],3),"+(" ,round(model$coefficients[[1]],3),")",sep = "")  
#punkty na wykresie to Dane oryginalne,czerwona linia to Dane z modelu  
#słupki błedów to 95% przedział istotności  
ggplot(Dane,aes(y=Objętość,x=1:length(Dane[,1])))+geom\_point()+  
 xlab("")+geom\_errorbar(ymin=predict(model,Dane,interval="confidence")[,2],  
 ymax=predict(model,Dane,interval="confidence")[,3],col="green")+  
 geom\_line(aes(y=model$fitted.values,color="red"))+  
 ggtitle(label="Regresja liniowa",subtitle=rownanie)+  
 theme(plot.title = element\_text(hjust = 0.5),plot.subtitle = element\_text(hjust = 0.5))+  
 theme(legend.position = "")

 4. Prognoza dla nowych danych

predict(model,newdata = (data.frame(Wysokość=50,Średnica=25)),interval = "confidence")

## fit lwr upr  
## 1 1.246053 0.5516679 1.940438

## Analiza statystyczna opisowa pliku *03 - analiza płac.csv*

1. Wczytanie i transformacja pliku csv *03 - analiza płac.csv*.

rm(list=ls())  
Dane<-read.table("./Dane/03 - analiza płac.csv",header = T,  
 dec=",",sep = ";",encoding = "UTF-8")  
colnames(Dane)[1]<-"Płeć"  
colnames(Dane)[5]<-"Płaca"  
Dane<-Dane[complete.cases(Dane),] #Wyrzucenie z danych niepełnych obserwacji NA

1. Wyznaczenie modelu

model<-lm(data=Dane,formula=Płaca~.)  
summary(model)

##   
## Call:  
## lm(formula = Płaca ~ ., data = Dane)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1056.90 -182.98 -48.38 128.48 2076.38   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.560 60.158 -0.142 0.886867   
## PłećMężczyzna 171.502 19.929 8.606 < 2e-16 \*\*\*  
## WykształcenieŚrednie ogólne 311.986 49.283 6.331 3.44e-10 \*\*\*  
## WykształcenieŚrednie zawodowe 313.785 40.895 7.673 3.44e-14 \*\*\*  
## WykształcenieWyższe 565.200 42.922 13.168 < 2e-16 \*\*\*  
## WykształcenieZasadnicze 119.346 44.003 2.712 0.006779 \*\*   
## Wiek 12.386 1.149 10.779 < 2e-16 \*\*\*  
## Staż.pracy -4.600 1.182 -3.893 0.000105 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 316.1 on 1210 degrees of freedom  
## Multiple R-squared: 0.3034, Adjusted R-squared: 0.2993   
## F-statistic: 75.27 on 7 and 1210 DF, p-value: < 2.2e-16

Prawdopodobieństwo ***p-value*** testu ***ANOVA*** jest mniejsze od 0,05, czyli dalsza analiza ma sens. ***Adjusted R-squared*** - skorygowany współczynnik determinacji jest na bardzo niskim poziomie (prawie 30%), to znaczy, że nie jest to dokładny model. Z testu T prawdopodobieństwo ***p-value*** dla każdego wyznaczonego parametru modelu jest mniejsze od 0,05, czyli każdy parametr jest istotny statystycznie. W prezentowanym modelu dla czynnika płeć, odniesieniem są kobiety. Parametr ***PłećMężczyzna*** można zinterpretować jako kwotę płacy (171,5) o jaką mężczyźni średnio dostają więcej. Dla wykształcenia poziom edukacji w szkole podstawowej to czynnik referencyjny, czyli np. dla kategorii ***WykształcenieWyższe*** osoby zarabiają średnio o 565,2 więcej.

1. Zmiana czynnika odniesienia

contrasts(Dane$Płeć) #wyświetlenie wartośći całkowitych przyporzadkowanych do czynników

## Mężczyzna  
## Kobieta 0  
## Mężczyzna 1

Dane$Płeć<-relevel(Dane$Płeć,ref="Mężczyzna") #zmiana czynnika względem,   
Dane$Wykształcenie<-relevel(Dane$Wykształcenie,ref="Wyższe") #którego bedzie zbudowany model   
  
model<-lm(data=Dane,formula=Płaca~.)  
summary(model)

##   
## Call:  
## lm(formula = Płaca ~ ., data = Dane)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1056.90 -182.98 -48.38 128.48 2076.38   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 728.142 47.300 15.394 < 2e-16 \*\*\*  
## PłećKobieta -171.502 19.929 -8.606 < 2e-16 \*\*\*  
## WykształceniePodstawowe -565.200 42.922 -13.168 < 2e-16 \*\*\*  
## WykształcenieŚrednie ogólne -253.214 36.983 -6.847 1.2e-11 \*\*\*  
## WykształcenieŚrednie zawodowe -251.415 23.948 -10.498 < 2e-16 \*\*\*  
## WykształcenieZasadnicze -445.854 30.535 -14.602 < 2e-16 \*\*\*  
## Wiek 12.386 1.149 10.779 < 2e-16 \*\*\*  
## Staż.pracy -4.600 1.182 -3.893 0.000105 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 316.1 on 1210 degrees of freedom  
## Multiple R-squared: 0.3034, Adjusted R-squared: 0.2993   
## F-statistic: 75.27 on 7 and 1210 DF, p-value: < 2.2e-16

anova(model)

## Analysis of Variance Table  
##   
## Response: Płaca  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Płeć 1 5771328 5771328 57.747 5.949e-14 \*\*\*  
## Wykształcenie 4 34824367 8706092 87.112 < 2.2e-16 \*\*\*  
## Wiek 1 10550494 10550494 105.566 < 2.2e-16 \*\*\*  
## Staż.pracy 1 1514335 1514335 15.152 0.0001046 \*\*\*  
## Residuals 1210 120929691 99942   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

print(RMSE<-sqrt(mean((Dane$Płaca-model$fitted.values)^2)))

## [1] 315.096

1. Prognoza dla nowych danych

nowe\_Dane<-data.frame(Płeć=c("Kobieta","Mężczyzna"),Wykształcenie=c("Wyższe","Wyższe"),  
 Wiek=c(65,65),Staż.pracy=c(30,30))  
print(prognoza<-predict(model,newdata = nowe\_Dane,interval = "confidence"))

## fit lwr upr  
## 1 1223.731 1156.305 1291.156  
## 2 1395.233 1325.736 1464.729

diff(prognoza) #różnica między wierszami

## fit lwr upr  
## 2 171.5023 169.4315 173.5732

summary(model)$coefficients #czyli współczynnik PłećKobieta to różnica

## Estimate Std. Error t value  
## (Intercept) 728.141605 47.300101 15.394081  
## PłećKobieta -171.502346 19.929373 -8.605506  
## WykształceniePodstawowe -565.199683 42.921643 -13.168174  
## WykształcenieŚrednie ogólne -253.214039 36.982554 -6.846851  
## WykształcenieŚrednie zawodowe -251.414602 23.948058 -10.498330  
## WykształcenieZasadnicze -445.853993 30.534709 -14.601547  
## Wiek 12.385889 1.149048 10.779257  
## Staż.pracy -4.599716 1.181664 -3.892577  
## Pr(>|t|)  
## (Intercept) 5.718134e-49  
## PłećKobieta 2.324940e-17  
## WykształceniePodstawowe 4.173815e-37  
## WykształcenieŚrednie ogólne 1.197660e-11  
## WykształcenieŚrednie zawodowe 9.793864e-25  
## WykształcenieZasadnicze 1.345810e-44  
## Wiek 6.218935e-26  
## Staż.pracy 1.045974e-04

## Ćwiczenia

1. Sprawdzić, jak i czy skala nasilenia choroby jest zależna od pozostałych czynników (plik “05 - choroba.csv”)
2. Zbadać, od czego zależy ndciśnienie i w jakim stopniu te czynniki wpływają na na podwyższenie ciśnienia,
3. Sprawdzić, jak nakłady środków na reklamę (radio, tv, prasa) wpływa na sprzedaż (plik “07 - Reklama.csv”)