0906 Linear Regression Model

Jongrak

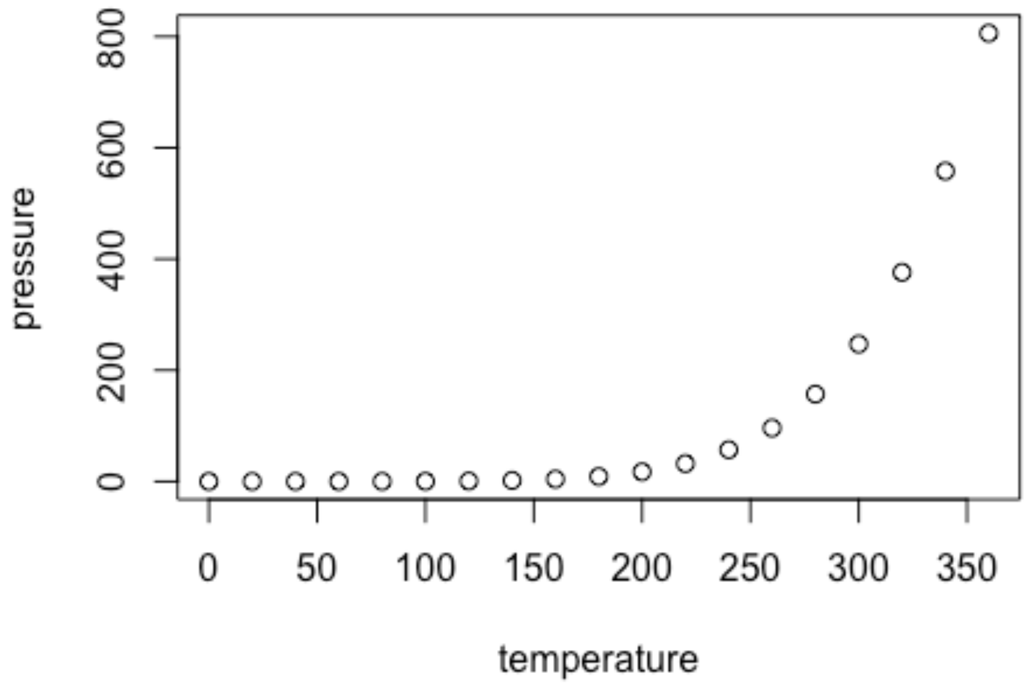
2022-12-31

```
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.

```
### Review

#plot(cars)

# linearModel <- lm(cars$dist ~ cars$speed)
#summary(linearModel)
#abline(linearModel)

# Outlier
#brokencars <- cars
#brokencars[51,] <- c(15, 160)
#brokencars[52,] <- c(25, 500)

#plot(brokencars)
#linearModel2 <- lm(brokencars$dist ~ brokencars$speed)


#summary(linearModel2)
#abline(linearModel2, col = "red")


# (Intercept) is beta 0 / cars$speed is beta 1
# 3.9324가 유효? t 분포의 확률값을 확인
# degress of freedom?
# Residual? Mean doesn't matter, Median -? > 대칭이 아니다. outlier가 있는가?

data <- read.table("P060.txt", header = TRUE) # txt는 헤더를 따로 설정해줘야
한다.

head(data)

##    Y X1 X2 X3 X4 X5 X6
## 1 43 51 30 39 61 92 45
## 2 63 64 51 54 63 73 47
## 3 71 70 68 69 76 86 48
## 4 61 63 45 47 54 84 35
## 5 81 78 56 66 71 83 47
## 6 43 55 49 44 54 49 34

linearModel <- lm(Y~X1+X2+X3+X4+X5+X6, data = data)
summary(linearModel)

##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6, data = data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -10.9418  -4.3555   0.3158   5.5425  11.5990 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 10.78708   11.58926   0.931 0.361634    
## X1           0.61319    0.16098   3.809 0.000903 ***
## X2          -0.07305    0.13572  -0.538 0.595594    
## X3           0.32033    0.16852   1.901 0.069925 .  
## X4           0.08173    0.22148   0.369 0.715480    
## X5           0.03838    0.14700   0.261 0.796334    
## X6          -0.21706    0.17821  -1.218 0.235577    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.068 on 23 degrees of freedom
## Multiple R-squared:  0.7326, Adjusted R-squared:  0.6628 
## F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

```
linearModel2 <- lm(Y~X1+X3, data = data)
summary(linearModel2)
```

```
## 
## Call:
## lm(formula = Y ~ X1 + X3, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.5568  -5.7331   0.6701   6.5341  10.3610 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   9.8709     7.0612   1.398    0.174    
## X1            0.6435     0.1185   5.432 9.57e-06 ***
## X3            0.2112     0.1344   1.571    0.128    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.817 on 27 degrees of freedom
## Multiple R-squared:  0.708,  Adjusted R-squared:  0.6864 
## F-statistic: 32.74 on 2 and 27 DF,  p-value: 6.058e-08
```

```
# 두 모델을 비교했을 때 Adjusted R-squared가 더 높은 2 모델이 더 좋다고 할 수 있다.
anova(linearModel, linearModel2)
```

```
## Analysis of Variance Table
## 
## Model 1: Y ~ X1 + X2 + X3 + X4 + X5 + X6
## Model 2: Y ~ X1 + X3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 1149.0
## 2     27 1254.7 -4   -105.65 0.5287 0.7158
```

```
#Pr(>F) is 기무가설이 맞을 확률. 베타2 베타 4 베타 5 베타 6 가 0일 확률. 이것이
높을수록 두번째 모델이 좋다.

# 지금까지는 연속형. 연속형이 아닌 것이라면? 성별, 직업, 학력, ? > Dummy Variable
사용

salarydata <- read.table("P130.txt", header = TRUE)
salaryModel0 <- lm(S~X+E+M, data = salarydata)
summary(salaryModel0)
```

```
##
## Call:
## lm(formula = S ~ X + E + M, data = salarydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2387.1  -657.6  -116.6   482.4  2922.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6963.48     665.69  10.460 2.88e-13 ***
## X             570.09      38.56  14.785  < 2e-16 ***
## E            1578.75     262.32   6.018 3.74e-07 ***
## M            6688.13     398.28  16.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1313 on 42 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9225
## F-statistic: 179.6 on 3 and 42 DF,  p-value: < 2.2e-16
```

```
#factor로 만들면 R에서 자동으로 dummy variable을 만들어준다.
salarydata$E <- factor(salarydata$E)
salarydata$M <- factor(salarydata$M)

salaryModel1 <- lm(S~X+E+M, data = salarydata)
summary(salaryModel1)
```

```
##
## Call:
## lm(formula = S ~ X + E + M, data = salarydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1884.60  -653.60    22.23   844.85  1716.47
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  8035.60     386.69  20.781  < 2e-16 ***
## X             546.18      30.52  17.896  < 2e-16 ***
## E2           3144.04     361.97   8.686 7.73e-11 ***
## E3           2996.21     411.75   7.277 6.72e-09 ***
## M1           6883.53     313.92  21.928  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1027 on 41 degrees of freedom
## Multiple R-squared:  0.9568, Adjusted R-squared:  0.9525 
## F-statistic: 226.8 on 4 and 41 DF,  p-value: < 2.2e-16
```

```
cig <- read.table("P088.txt", header = TRUE)
head(cig)
```

```
##   State  Age   HS Income Black Female Price Sales
## 1    AL 27.0 41.3   2948  26.2   51.7  42.7  89.8
## 2    AK 22.9 66.7   4644   3.0   45.7  41.8 121.3
## 3    AZ 26.3 58.1   3665   3.0   50.8  38.5 115.2
## 4    AR 29.1 39.9   2878  18.3   51.5  38.8 100.3
## 5    CA 28.1 62.6   4493   7.0   50.8  39.7 123.0
## 6    CO 26.2 63.9   3855   3.0   50.7  31.1 124.8
```

```
linearcig0 <- lm(Sales ~ Age+HS+Income+Black+Female+Price, data = cig)
summary(linearcig0)
```

```
## 
## Call:
## lm(formula = Sales ~ Age + HS + Income + Black + Female + Price, 
##     data = cig)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -48.398 -12.388  -5.367   6.270 133.213 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 103.34485  245.60719   0.421  0.67597   
## Age           4.52045    3.21977   1.404  0.16735   
## HS           -0.06159    0.81468  -0.076  0.94008   
## Income        0.01895    0.01022   1.855  0.07036 . 
## Black         0.35754    0.48722   0.734  0.46695   
## Female       -1.05286    5.56101  -0.189  0.85071   
## Price        -3.25492    1.03141  -3.156  0.00289 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 28.17 on 44 degrees of freedom
## Multiple R-squared:  0.3208, Adjusted R-squared:  0.2282
## F-statistic: 3.464 on 6 and 44 DF,  p-value: 0.006857
```

```
linearcig1 <- lm(Sales ~ Age+Income+Black+Price, data = cig)
anova(linearcig0, linearcig1)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Age + HS + Income + Black + Female + Price
## Model 2: Sales ~ Age + Income + Black + Price
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     44 34926
## 2     46 34960 -2   -33.799 0.0213 0.9789
```

```
linearCigFinal <- lm(Sales ~ Age+Income+Black+Price, data = cig)
summary(linearCigFinal)
```

```
##
## Call:
## lm(formula = Sales ~ Age + Income + Black + Price, data = cig)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.784 -11.810  -5.380   5.758 132.789
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 55.329580  62.395293   0.887   0.3798
## Age          4.191538   2.195535   1.909   0.0625 .
## Income       0.018892   0.006882   2.745   0.0086 **
## Black        0.334162   0.312098   1.071   0.2899
## Price       -3.239941   0.998778  -3.244   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.57 on 46 degrees of freedom
## Multiple R-squared:  0.3202, Adjusted R-squared:  0.2611
## F-statistic: 5.416 on 4 and 46 DF,  p-value: 0.001168
```