BIOSTAT 650 SEC 002 HW4

Jaehoon Kim

2024-09-30

```r
mySASData = read.sas7bdat("C:/Users/aquil/Desktop/STAT 650/completedata.sas7bdat")

df <- mySASData
reg <- lm(Depression ~ Fatalism+Age+Sex+R_E, df)
summary(reg)
```

```
##
## Call:
## lm(formula = Depression ~ Fatalism + Age + Sex + R_E, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.055 -3.591 -1.208  2.039 22.049
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.54281    1.39969   4.674 3.63e-06 ***
## Fatalism     0.25471    0.03877   6.570 1.08e-10 ***
## Age         -0.08839    0.01755  -5.037 6.24e-07 ***
## Sex          0.51520    0.41052   1.255    0.210
## R_E          0.29460    0.42771   0.689    0.491
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.027 on 607 degrees of freedom
## Multiple R-squared:  0.1065, Adjusted R-squared:  0.1006
## F-statistic: 18.08 on 4 and 607 DF,  p-value: 4.829e-14
```

```r
#define new observation
avg_fatal <- mean(df$Fatalism)
newdata = data.frame(Fatalism=avg_fatal, Age=70, Sex=1, R_E=1)

#use model to predict points value
predict(reg, newdata, se.fit=TRUE)
```

```
## $fit
##        1
## 5.611173
##
## $se.fit
## [1] 0.3575402
##
## $df
```

```
## [1] 607
##
## $residual.scale
## [1] 5.027098
```

```r
#estimated sigma squared
summary(reg)$sigma
```

```
## [1] 5.027098
```

```r
X = df[c("Fatalism", "Age", "Sex", "R_E")]
X = cbind(1,X)
X_i = cbind(a=1, newdata)


VarYhat <-
  (summary(reg)$sigma)^2*as.matrix(X_i)%*%solve(t(as.matrix(X))%*%as.matrix(X))%*%t(as.matrix(X_i))

VarYhat
```

```
##          [,1]
## [1,] 0.127835
```

```r
sqrt(VarYhat)
```

```
##           [,1]
## [1,] 0.3575402
```

Problem 1.

Adjusting for demographic variables age, sex, and race/ethnicity, we estimate on average there is a 0.25 unit points increase in depression for every unit point increase in fatalism. This association is statistically significant with a p-value 1.08e-10 at a significance level $\alpha = 0.05$.

Problem 2.

The adjusted R^2 of 0.1006 indicates that approximately 10.06% of the variation of the depression variable is explained by the multiple linear regression model containing fatalism, age, sex, race/ethnicity variables. The model shows a slight improvement as a predictor for the depression score compared to the simple model with only the fatalism variable.

Problem 3.

The fitted model prediction estimates that on average 70 year old, female, Mexican Americans have a depression score of approximately 5.611. The variance of the estimate is

$$\text{Var}(\hat{Y}_i) = \text{Var}(\mathbf{X}_i^T \hat{\boldsymbol{\beta}}) = \mathbf{X}_i^T \text{Var}(\hat{\boldsymbol{\beta}})\mathbf{X}_i = \hat{\sigma}^2 \mathbf{X}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_i = 0.127835$$

Thus, the standard error of the estimate is approximately $\sqrt{\text{Var}(\hat{Y}_i)} = 0.3575$.

Problem 4.

(a)

All the matrices $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ are symmetrical and idempotent. We prove idempotency as such:

$$\mathbf{A}_1^2 = \mathbf{A}_1\mathbf{A}_1 = \frac{1}{9}\begin{bmatrix}1 & 1 & 1\\1 & 1 & 1\\1 & 1 & 1\end{bmatrix}\begin{bmatrix}1 & 1 & 1\\1 & 1 & 1\\1 & 1 & 1\end{bmatrix}$$
$$= \frac{1}{9}\begin{bmatrix}3 & 3 & 3\\3 & 3 & 3\\3 & 3 & 3\end{bmatrix} = \frac{1}{3}\begin{bmatrix}1 & 1 & 1\\1 & 1 & 1\\1 & 1 & 1\end{bmatrix}$$

$$\mathbf{A}_2^2 = \mathbf{A}_2\mathbf{A}_2 = \frac{1}{4}\begin{bmatrix}1 & -1 & 0\\-1 & 1 & 0\\0 & 0 & 0\end{bmatrix}\begin{bmatrix}1 & -1 & 0\\-1 & 1 & 0\\0 & 0 & 0\end{bmatrix}$$
$$= \frac{1}{4}\begin{bmatrix}2 & -2 & 0\\-2 & 2 & 0\\0 & 0 & 0\end{bmatrix} = \frac{1}{2}\begin{bmatrix}1 & -1 & 0\\-1 & 1 & 0\\0 & 0 & 0\end{bmatrix}$$

$$\mathbf{A}_3^2 = \mathbf{A}_3\mathbf{A}_3 = \frac{1}{36}\begin{bmatrix}1 & 1 & -2\\1 & 1 & -2\\-2 & -2 & 4\end{bmatrix}\begin{bmatrix}1 & 1 & -2\\1 & 1 & -2\\-2 & -2 & 4\end{bmatrix}$$
$$= \frac{1}{36}\begin{bmatrix}6 & 6 & -4\\6 & 6 & -4\\-4 & -4 & 0\end{bmatrix} = \frac{1}{6}\begin{bmatrix}1 & 1 & -2\\1 & 1 & -2\\-2 & -2 & 4\end{bmatrix}$$

$\mathbf{A}_1$ has rank 1, $\mathbf{A}_2$ has rank 1, $\mathbf{A}_3$ has rank 1. Accordingly, given the quadratic form $Q_i = \mathbf{Y}^T\mathbf{A}_i\mathbf{Y}$ where $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ for $i = 1, 2, 3$,

$$\frac{Q_i}{\sigma^2} = \frac{\mathbf{Y}^T\mathbf{A}_i\mathbf{Y}}{\sigma^2} \sim \chi_{1,\lambda}$$

with non-centrality parameter $\lambda = \frac{\mu^T\mathbf{A}_i\mu}{\sigma^2}, \forall i = 1, 2, 3$

(b)

$$\mathbf{A}_1\mathbf{A}_2 = \frac{1}{12}\begin{bmatrix}2-2 & -2+2 & 0\\2-2 & -2+2 & 0\\2-2 & -2+2 & 0\end{bmatrix} = \mathbf{0}_{3\times 3}$$

$$\mathbf{A}_1\mathbf{A}_3 = \frac{1}{18}\begin{bmatrix}2-2 & -2+2 & -4+4\\2-2 & 2-2 & -4+4\\2-2 & -2+2 & -4+4\end{bmatrix} = \mathbf{0}_{3\times 3}$$

$$\mathbf{A}_2\mathbf{A}_3 = \frac{1}{12}\begin{bmatrix}1-1 & 1-1 & -2+2\\2-2 & 2-2 & -4+4\\2-2 & 2-2 & -4+4\end{bmatrix} = \mathbf{0}_{3\times 3}$$

Thus, the $Q_i = \mathbf{Y}^T\mathbf{A}_i\mathbf{Y}$ are pairwise independent.