Homework 9

Shivam Bajaj

2023-03-25

#### 1. Linear regression.

We went through a manual linear regression process in class. Follow the same steps using the data in the death\_by\_gender dataset available [here](https://jlucasmckay.bmi.emory.edu/global/bmi510/death_by_gender.csv). This dataset is based on CDC data, and is a sample of deaths recorded during a particular period labeled with gender The columns are age and gender.

Use age as the response variable and gender as the predictor variable. Treat F as the reference group. Fit the following model:

* Build the design matrix and create a matrix version of the response variable age. **(1 point)**
* Apply the normal equations to derive the OLS estimates of the s. **(1 point)**
* Calculate the residuals and the residual sum of squares . **(1 point)**
* Calculate the residual standard error . **(1 point)**
* Calculate , the matrix used to derive the standard errors of the s. **(1 point)**
* Calculate , the standard error of . **(1 point)**
* Calculate a t statistic for and compare it to the t-statistic from the function lm. **(1 point)**

url <- “<https://jlucasmckay.bmi.emory.edu/global/bmi510/death_by_gender.csv>” death\_by\_gender <- read\_csv(url)

X <- model.matrix(~ gender, data = death\_by\_gender)

age <- death\_by\_gender$age

betas <- solve(t(X) %*% X) %*% t(X) %\*% age

residuals <- age - X %\*% betas RSS <- sum(residuals^2)

n <- nrow(X) p <- ncol(X) s <- sqrt(RSS / (n - p))

C <- s^2 \* solve(t(X) %\*% X)

s\_beta0 <- sqrt(C[1, 1])

s\_beta1 <- sqrt(C[2, 2])

t\_stat\_beta0 <- betas[1] / s\_beta0 t\_stat\_beta1 <- betas[2] / s\_beta1

linear\_model <- lm(age ~ gender, data = death\_by\_gender) lm\_coefficients <- summary(linear\_model)$coefficients

cat(“Manual calculations:”) cat(“beta0 (Intercept):”, betas[1], “”) cat(“beta1 (genderM):”, betas[2], “”) cat(“Std. Error for beta0:”, s\_beta0, “”) cat(“Std. Error for beta1:”, s\_beta1, “”) cat(“t-statistic for beta0:”, t\_stat\_beta0, “”) cat(“t-statistic for beta1:”, t\_stat\_beta1, “”)

cat(“lm function output:”) cat(“beta0 (Intercept):”, lm\_coefficients[1, 1], “”) cat(“beta1 (genderM):”, lm\_coefficients[2, 1], “”) cat(“Std. Error for beta0:”, lm\_coefficients[1, 2], “”) cat(“Std. Error for beta1:”, lm\_coefficients[2, 2], “”) cat(“t-statistic for beta0:”, lm\_coefficients[1, 3], “”) cat(“t-statistic for beta1:”, lm\_coefficients[2, 3], “”)

```

#### 2. Centering and scaling data for regression.

In many cases, the parameters of linear models can be more interpretable by *centering* and *scaling* the independent and dependent variables prior to entry into regression models. *Centering* refers to removing the mean value of the variable, and *scaling* refers to scaling the variable to have some convenient range. The most common scaling method is to scale the variable so that it has unit variance (and also unit standard deviation, since ).

* Use apply to build a function that centers and scales all columns of an input matrix x. **(1 point)**
* Test your function on the first four columns of the iris dataset. **(1/2 point)** and compare your results to those of scale. **(1/2 point)**
* Consider the following model: .
  + What does represent? What does it represent if is centered and scaled prior to fitting the model? **(1/3 point)**
  + What does represent? What does it represent if is centered and scaled prior to fitting the model? **(1/3 point)**
  + What does represent if is centered and scaled to units of 5 years prior to fitting the model? **(1/3 point)**

What does represent? What does it represent if is centered and scaled prior to fitting the model? **(1/3 point)** represents expected height when Age is 0. If is centered and scaled, prior to fitting the model, represents expected height for the average Age.

What does represent? What does it represent if is centered and scaled prior to fitting the model? represents the change in height for a one-unit increase in . If is centered and scaled prior to fitting the model, represents the change in height for a one standard deviation increase in .

What does represent if is centered and scaled to units of 5 years prior to fitting the model? If is centered and scaled to units of 5 years prior to fitting the model, represents the change in height for a 5-year increase in