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. 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:

$$age_i = \beta_0 + \beta_1 \cdot \text{Gender}_i + \epsilon_i$$

- Build the design matrix $\mathbf{X}$ and create a matrix version of the response variable `age`. **(1 point)**
- Apply the normal equations to derive the OLS estimates of the βs. **(1 point)**
- Calculate the residuals and the residual sum of squares RSS. **(1 point)**
- Calculate the residual standard error s. **(1 point)**
- Calculate C, the matrix used to derive the standard errors of the βs. **(1 point)**
- Calculate $s_{\beta 1}$, the standard error of β_1. **(1 point)**
- Calculate a t statistic for β_1 and compare it to the t-statistic from the function `lm`. **(1 point)**

---

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 $\sigma = \sqrt{\sigma^2} = \sqrt{1^2} = 1$).

- 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: height $= \beta_0 + \beta_1$Age.
  - What does β_0 represent? What does it represent if Age is centered and scaled prior to fitting the model? **(1/3 point)**
  - What does β_1 represent? What does it represent if Age is centered and scaled prior to fitting the model? **(1/3 point)**
  - What does β_1 represent if Age is centered and scaled to units of 5 years prior to fitting the model? **(1/3 point)**

What does β_0 represent? What does it represent if Age is centered and scaled prior to fitting the model? **(1/3 point)** β_0 represents expected height when Age is 0. If Age is centered and scaled, prior to fitting the model, β_0 represents expected height for the average Age.

What does β_1 represent? What does it represent if Age is centered and scaled prior to fitting the model? β_1 represents the change in height for a one-unit increase in Age. If Age is centered and scaled prior to fitting the model, β_1 represents the change in height for a one standard deviation increase in Age.

What does β_1 represent if Age is centered and scaled to units of 5 years prior to fitting the model? If Age is centered and scaled to units of 5 years prior to fitting the model, β_1 represents the change in height for a 5-year increase in Age