Week 2: Linear Regression Exercise

You will need to refer to the week 2 materials (data file, R script files and videos) to help with this exercise.

1. Locate the .csv file named “height\_data\_fake.csv” and save it to a dedicated working directory.

**Ans**. I saved in below directory

/Users/abhijitghosh/Documents/DataScience/DataAnalytics

1. Create an R script file, set the working directory and read in the data file named “height\_data\_fake.csv”

**Ans**. Below is the code. I saved the file named as week2.R

setwd("/Users/abhijitghosh/Documents/DataScience/DataAnalytics")

ageandheight <-read.csv("height\_data\_fake.csv")

1. Plot *age* against *height* and check that your results are the same as in video 1.

**Ans**.

hgt <- ageandheight$Height\_feet

age <- ageandheight$Age\_months

plot(age,hgt)

1. Fit a simple linear regression model with *height* as the outcome (dependent) variable and *age* as the predictor (independent) variable. Check your results against video 1.

**Ans**.

curve(0.14632\*x+0.84485,xlim=c(1,5),xlab="Age in months",ylab="Height in feet",ylim=(c(1:2)))

points(age,hgt)

We are able to draw a best fit line and it follows a simple linear regression. It matches with the result video 1. From figure we can interpret for young kids the sample data we take, it follows a simple linear regression and height is increasing as they grow older.

1. Fit a linear regression model with *height* as the outcome (dependent) variable, and *age*, *biological sex* and *height* at birth as the predictor (independent) variables.

**Ans**.

ageandheight <-read.csv("height\_data\_fake.csv")

hgt <- ageandheight$Height\_feet

bio <- ageandheight$Biological\_sex

age <- ageandheight$Age\_months

birthHeight <- ageandheight$Height\_birth\_feet

bio[which(bio=="M")] <- 0

bio[which(bio=="F")] <- 1

bio <- as.numeric(bio)

summary(lm(hgt~age+bio+birthHeight))

1. Interpret the output of the model in step 5. Make sure to discuss effect estimates and p values.

**Ans**. Below is summary output

**Call:**

**lm(formula = hgt ~ age + bio + birthHeight)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-0.12423 -0.05973 -0.01978 0.07995 0.13173**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 0.42622 0.11412 3.735 0.001991 \*\***

**age 0.12798 0.01596 8.017 8.39e-07 \*\*\***

**bio -0.10473 0.04591 -2.281 0.037558 \***

**birthHeight 0.63760 0.12789 4.986 0.000163 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.09068 on 15 degrees of freedom**

**Multiple R-squared: 0.9037, Adjusted R-squared: 0.8845**

**F-statistic: 46.93 on 3 and 15 DF, p-value: 7.379e-08**

**Age (0.12798):**

* For **each additional month of age**, the expected height **increases by 0.12798 units**.
* The effect is **statistically significant** (p-value = 8.39e-07 < 0.001), indicating strong evidence that age affects height.

**Biological Sex (-0.10473):**

* Since bio is treated as a **numeric variable (0 for Male, 1 for Female)**:
  + **Males (bio = 0) are the reference group**.
  + **Females (bio = 1) are, on average, 0.10473 units shorter than males, holding age and birth height constant**.
* The effect is **statistically significant (**p = 0.037558 < 0.05**)**, meaning there is moderate evidence that biological sex influences height.

**Birth Height (0.63760):**

* For **each additional unit of birth height**, the expected current height **increases by 0.63760 units**.
* This effect is **highly significant (**p = 0.000163 < 0.001**)**, indicating a strong relationship between birth height and current height.

**R Square (0.9037) and Adjusted R Square (0.8845)**

Suggest model is a very good fit since the value is close to 1.

**F-statistics**

Very small P value (7.379e-08) says overall model is **statistically significant**.

1. Is the model in step 5 a classification model or a regression model? Explain why.

**Ans**. This is a regression model, not a classification model, because it predicts a continuous numerical value (height) rather than a categorical label.

1. Is the model in step 5 a simple linear regression model or a multiple regression model? Explain why.

**Ans**. It’s a multiple linear regression because here we are considering multiple predictors (age, biological sex and birth height) in predicting models. Simple linear regression is when we are considering one predictor in model.

1. Is linear regression an example of supervised or unsupervised learning? Explain why.

**Ans**. Linear regression is a supervised learning because it learns from dataset that contains both X and Y labeled data. Here height is dependent variable and age, birth height and biological sex are independent variable and the model predicts height for given input.

Unsupervised learning is to find pattern/structure of the data learning from unlabeled data.

1. Which metric(s) did you use to evaluate the quality of your results? Explain why you used them.

**Ans**.

**R Square**: It says how well the independent variables explain the variation in the dependent variable (height).

**P Values: P-value tells us whether each predictor (age, bio, birth height) has a statistically significant impact on height.**

**F-Statistic:** This is to Tests whether at least one predictor is significantly related to the outcome. A **very low p-value confirms the overall model is statistically significant.**