Name: Dishant Naik
CWID: 10454341

HW 3

Q1 - Show that the following relationship on the simple linear regression class notebook is true:

$$\frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

Answer:

Let's expand for RHS

$$\frac{\sum_{i\,=\,1}^{n}(x_i\ -\ \bar{X})(y_i\ -\ \bar{Y})}{\sum_{i\,=\,1}^{n}(x_i\ -\ \bar{X})^2}$$

$$= \frac{\sum_{i\,=1}^{n} x_i y_i\ -\ \bar{Y}\ \sum_{i\,=\,1}^{n} x_i\ -\ \bar{X}\ \sum_{i\,=\,1}^{n} y_i\ +\ \bar{X}\,\bar{Y}\ \sum_{i\,=\,1}^{n} 1}{\sum_{i\,=\,1}^{n}(x_i^2\ -\ 2(x_i)(\bar{X})\ +\ \bar{X}^2)}$$

$$= \frac{\sum_{i\,=\,1}^{n} x_i y_i\ -\ \bar{Y}\,n\bar{X}\ -\ \bar{X}\,n\bar{Y}\ +\ \bar{X}\ \bar{Y}\,n}{\sum_{i\,=\,1}^{n} x_i^2\ -\ 2\bar{X}\ \sum_{i\,=\,1}^{n} x_i\ +\ \bar{X}^2\ \sum_{i\,=\,1}^{n} 1}$$

$$= \frac{\sum_{i\,=\,1}^{n} x_i y_i\ -\ 2\,n\,\bar{X}\,\bar{Y}\ +\ n\,\bar{X}\,\bar{Y}}{\sum_{i\,=\,1}^{n} x_i^2\ -\ 2n\,\bar{X}^2\ +\ n\,\bar{X}^2}$$

$$= \frac{\sum_{i\,=\,1}^{n} x_i y_i\ -\ n\bar{X}\,\bar{Y}}{\sum_{i\,=\,1}^{n} x_i^2\ -\ n\,\bar{X}^2}$$

$$= LHS$$

Hence proved.

Q2 - We showed that in a simple linear regression, the MLE estimates for $\beta 1$ and $\beta 0$ are the same as the OLS estimates.

Answer:

Log likelihood function for the simple linear reaeration can be written as bellow,

$$LL\ =\ -\frac{n}{2}\log(2\pi)\ -\frac{n}{2}\log\sigma^2\ -\frac{\mathrm{n}}{2\,\sigma^2}\sum_{\mathrm{i}\,=\,1}^{\mathrm{n}}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i\right)^2$$

Now differentiate LL with respect to σ^2,

$$\frac{2\partial\ \mathrm{LL}}{2\partial\ \sigma^2}\ =\ -\ \frac{\mathrm{n}}{2\,\sigma^2} + \frac{1}{2\,\sigma^4}\sum_{\mathrm{i}\,=\,1}^{\mathrm{n}}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i\right)^2$$

Now let's do MLE for σ^2. Which can be done by setting derivation to 0.

$$-\ \frac{\mathrm{n}}{2\,\sigma^2} + \frac{1}{2\,\sigma^4}\sum_{\mathrm{i}\,=\,1}^{\mathrm{n}}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i\right)^2\ =\ 0$$

$$\frac{\mathrm{n}}{2\,\sigma^2}\ =\ \frac{1}{2\,\sigma^4}\sum_{\mathrm{i}\,=\,1}^{\mathrm{n}}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i\right)^2$$

$$\frac{\mathrm{n}\ 2\ \sigma^4}{2\ \sigma^2} = \sum_{\mathrm{i}\,=\,1}^{\mathrm{n}} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2$$

$$\mathrm{n}\ \sigma^2 = \sum_{\mathrm{i}\,=\,1}^{\mathrm{n}} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2$$

$$\widehat{\sigma^2} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}\,=\,1}^{\mathrm{n}} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2$$

Hence proved.

Q3 - In a simple linear regression, show that the OLS regression line always passes through the mean (average) of both x and y.

Answer:

We know that

$$\mathrm{Y_i} = \widehat{\beta_0} + \widehat{\beta_1}\,\mathrm{X_i} + Recdual$$

And this is true for all the values of I (1, 2, 3, ….. ,n)
So, lets write individual formula

$$\begin{aligned}
\mathrm{Y_1} &= \widehat{\beta_0} + \widehat{\beta_1}\mathrm{X_1} + Recdual \\
\mathrm{Y_2} &= \widehat{\beta_0} + \widehat{\beta_1}\,\mathrm{X_2} + Recdual \\
\mathrm{Y_3} &= \widehat{\beta_0} + \widehat{\beta_1}\,\mathrm{X_3} + Recdual \\
. \quad & \quad . \qquad . \qquad\quad . \\
. \quad & \quad . \qquad . \qquad\quad . \\
. \quad & \quad . \qquad . \qquad\quad . \\
. \quad & \quad . \qquad . \qquad\quad . \\
\mathrm{Y_n} &= \widehat{\beta_0} + \widehat{\beta_1}\,\mathrm{X_n} + Recdual
\end{aligned}$$

Now sum up all the equation

$\mathrm{Y_1} + \mathrm{Y_2} + \mathrm{Y_3} + \ldots\ldots.. + \mathrm{Y_n} = (\widehat{\beta_0} + \widehat{\beta_1}\mathrm{X_1} + Recdual) + (\widehat{\beta_0} + \widehat{\beta_1}\,\mathrm{X_2} + Recdual) + \ldots\ldots\ldots.. + (\widehat{\beta_0} + \widehat{\beta_1}\,\mathrm{X_n} + Recdual)$

Which is equivalent to

$$\sum_{i\,=\,1}^{n} \mathrm{Y_i} = n\widehat{\beta_0} + \widehat{\beta_1} \sum_{i\,=\,1}^{n} \mathrm{X_i} + 0 \qquad As\ sum\ of\ Recdual\ is\ 0$$

Dividing equation by n

$$\left.\sum\nolimits_{\boldsymbol{i}\,=\,\mathbf{1}}^{\boldsymbol{n}} \mathrm{Y_i}\middle/\boldsymbol{n}\right. = \left.\boldsymbol{n}\,\widehat{\beta_0}\middle/\boldsymbol{n}\right. + \left.\widehat{\beta_1}\ \sum\nolimits_{i\,=\,1}^{n} \mathrm{X_i}\middle/\boldsymbol{n}\right.$$

And here $\left.\sum\nolimits_{\boldsymbol{i}\,=\,\mathbf{1}}^{\boldsymbol{n}} \mathrm{Y_i}\middle/\boldsymbol{N}\right.$ and $\left.\sum\nolimits_{i\,=\,1}^{n} \mathrm{X_i}\middle/\boldsymbol{n}\right.$ is just a mean of X and Y ($\bar{X}$ and $\bar{Y}$)

Hence proved that OLS regression line always passes through the mean (average) of both x and y

Q4 -

In the following estimated multiple regression model, we study the amount of sleep (in minutes) an average person gets every night. Each observation is a day, and if the day is a holiday, *Holiday* = 1; if the day is Monday, *Monday = 1.* In other words, all independent variables are dummy variables.

$$\widehat{sleep} = 458.5 + 8.6Holiday - 4.9Monday - 7.7Tuesday - 7.4Wednesday - 4.7\ Thursday$$
$$+ 23.4Friday + 30.6Saturday$$

a. Explain why Sunday is not included in the model as an independent variable.

b. Explain in plain English the meaning of the intercept 458.5. Does this number make sense?

c. According to the model, how much sleep does an average person get on a day that is both a Saturday and a Holiday?

Answer:

(a) Generally, in most of the workplaces, schools, collages etc. Sunday is considered as a holiday and in our model, we already have holiday as one independent variable. Hence there is no need to consider Sunday as a separate independent variable. Or one other view can be that when value of all the dummy variable is then it automatically means that it is Sunday.

(b) Here intercept is 458.5 which represents the avg munities of sleep an average person get no matter what the day is. And if we convert this intercept into hours then it is 7.6 hours. According to many scientific studies an average person takes a sleep of 7 to 8 hours of sleep. So, this intercept dose makes sense.

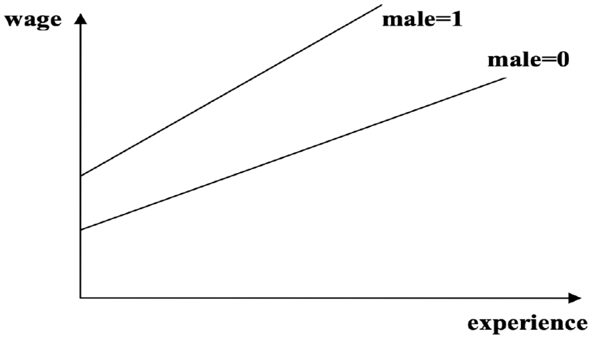
(c) $\widehat{sleep} = 458.5 + 8.6(1) + 30.6(1) \;=\; 497.7\ munites$

Q5 –

We estimate the following regression model to predict wage of employees in a firm. Exp is years of experience, *male* is a dummy variable that equals 1 if the employee is male.

$$\widehat{wage} = \beta_0 + \beta_1 exp + \beta_2 male + \beta_3 exp \times male$$

If the true relationship between the variables are as the following plot suggests, what do you expect the signs (positive or negative) of β_1, β_2, and β_3 to be? Explain.



Answer:

$\beta 0$ will be positive no matter male is 0 or 1 as it is an intercept.

$\beta 1$ will be positive as we can clearly infer from the graph that as exp increases wage of a person increases.

$\beta 2$ will be positive as we can clearly observe from the graph that if person is male his wage rises.

$\beta 3$ will also be positive as graph clearly tells us that if person is male with more experience than his wage will be higher.