Self Notes - Week2

Simple Linear Regression

Important Notes:

1. Assuming estimation is most important for this course in terms of examination point of view.
2. Difference between deterministic relationship and statistical relationship.
3. **Least squares regression -** we sum the square of the errors and with that we can minimise the errors.
4. For MCQs - we can get a question like give the assumptions or list the assumptions from the given plot (slide 26)
5. Hypothesis testing - https://www.section.io/engineering-education/hypothesis-testing-data-science/

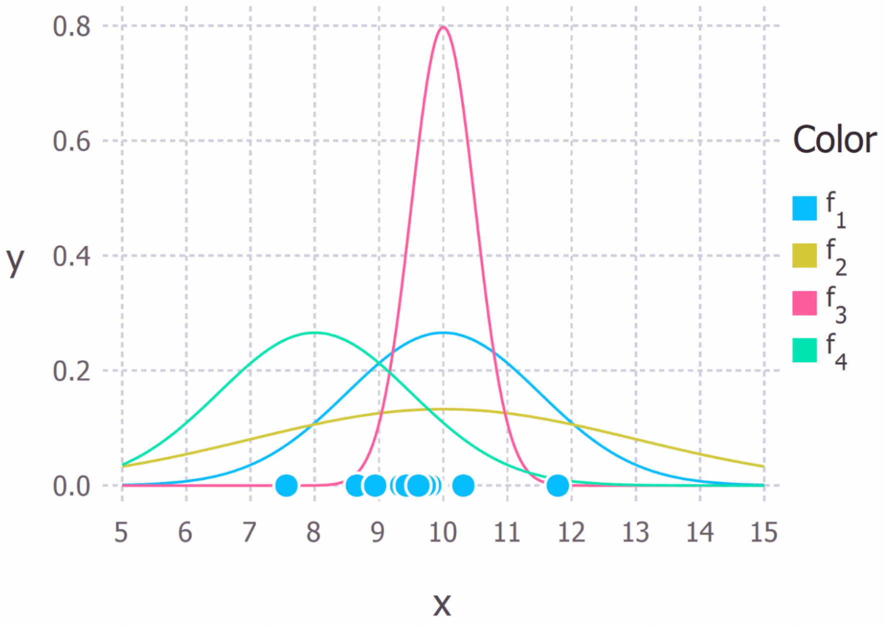
Extra Notes:

1. In Linear Models we have 1 target variable and 1 predictor variable
2. For estimating the population variable we generally use Alpha, Beta for linear models(equation of a line) or Beta0, Beta1 etc.
3. we generally use least square estimates method for linear regression.
4. ANOVA - Analysis of variance table to test the significance of the parameters.

5. The more the scatter, the less the accuracy of the model.
6. Model checking - testing the assumptions.
7. **Slope** and **intercept** are the two measures that describes the strength of linear model.
8. **Intercept(alpha)** is the value of “y” when “x” equals 0.
9. **Slope(beta)** is how much the value for “y” changes when we change the value for “x”.
10. if slope increases or if slope is positive then when x increases, y increases. If slope is negative then when x increases, y decreases.
11. sd and variance is a spread in a plot.
12. R code - lm(Sales~Type) to check the significance.
13. R code - confint(model) is confident interval
14. R code - anova(model) is used to determine the coefficient of determination(R^2) - Residuals are not captured from regression (can be say it as residuals = errors [need to check])
15. Residual mean square is the estimate of the variance of Y variable.

<u>Paragraph for more explanation or useful links:</u>

- Maximum Likelihood estimates is a method that determines the parameters of a model. The parameter values are found such that they maximise the likelihood or let’s say maximise the correctness of data points for a function. Good read: https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1



The 10 data points and possible Gaussian distributions from which the data were drawn. f1 is normally distributed with mean 10 and variance 2.25 (variance is equal to the square of the standard deviation), this is also denoted f1 ~ N (10, 2.25). f2 ~ N (10, 9), f3 ~ N (10, 0.25) and f4 ~ N (8, 2.25). The goal of maximum likelihood is to find the parameter values that give the distribution that maximise the probability of observing the data.

- Least square estimates is a method of estimating the values of alpha, beta such that it minimises the error. For this we can think of finding the **loss function** first for the error and then **using square of partial derivative of this loss function** we can find the least square estimates.
- **Deterministic relationship** is when all the points lie on the straight-line or the curve which we used as a function.
- **Statistical relationship** is when we get a line which is in between those scattered data points and for which we can get our

population parameters.

- **Normal Distribution** - https://www.analyticsvidhya.com/blog/2021/05/normal-distribution-an-ultimate-guide/