Second class essay Raul Adell

This essay may contain ideas given in class complemented by the book given on the references of the course and some notes of some statistics courses I took. Book section: 3. Linear Regression

Linear regression is a statistical modeling technique used to establish a relationship between a dependent variable and one or more independent variables. We have discussed in class the explanation of linear regression, including its assumptions, estimation, and interpretation. In this essay, we will explore the key concepts of linear regression.

The first concept discussed in the chapter is the simple linear regression model. This model assumes a linear relationship between a dependent variable Y and a single independent variable X. The model can be written as:

$Y = \beta 0 + \beta 1X + \varepsilon$

where β0 and β1 are the intercept and slope coefficients, respectively, and ε is the error term representing the unexplained variation in Y. The goal of linear regression is to estimate the coefficients β0 and β1 that minimize the sum of squared errors between the observed values of Y and the predicted values based on X.

The linear regression has many advantages:

- It is simple.
- Provides a clear explanation. Weights tell us the importance of each variable. Can get an idea of the relative importance of features.
- Are the building blocks of neural networks.
- Can be easily parallelized.
- Are used as benchmarks because they work really well.

To estimate the coefficients, we use the method of least squares. This performance criterion is based on minimizing vertical errors. This involves finding the values of β0 and β1 that minimize the residual sum of squares (RSS) between the observed values of Y and the predicted values based on X. The RSS is defined as:

$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

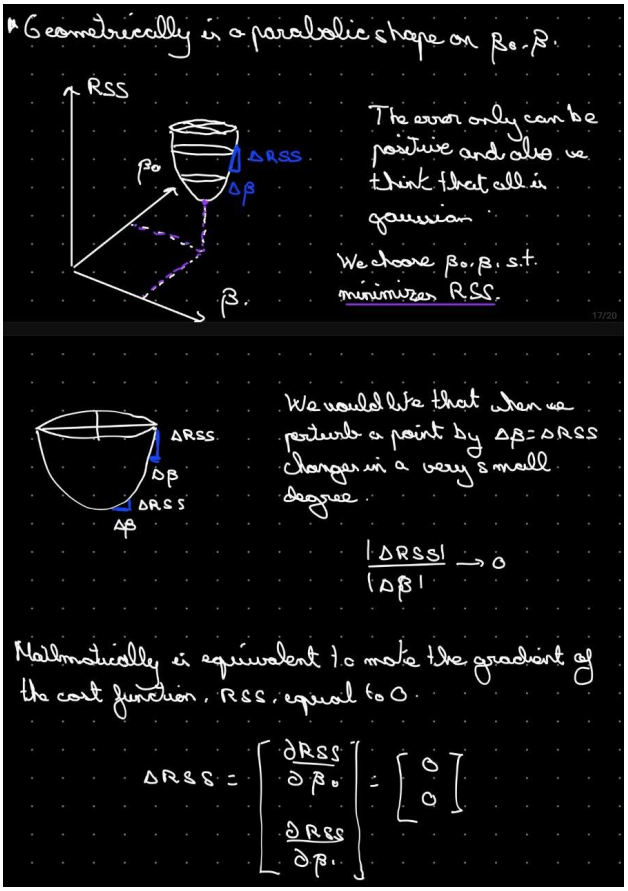
where yi is the observed value of Y for the ith observation, ŷi is the predicted value of Y based on the ith observation of X, and n is the sample size. This criterion is equivalent to moving a ruler in a 2D plane in order to find the best straight line that minimizes the distance to the data points.

The conditions of minimizing RSS can be written in a pair of equations (in 1-D linear regression) called the normal equations. Let $(y_i - \hat{y}_i)^2 = e_i$, then the normal equations state:

$$\sum_{i=1}^{n} e_i = 0$$

This concept can be interpreted geometrically in the case of single dimension linear regression with the following drawings.

$$\sum_{i=1}^{n} e_i \cdot X_i = 0$$



Once we have estimated the coefficients, we can use the model to make predictions for new values of X. The predicted value of Y based on a new value of X is given by:

$\hat{y} = \beta 0 + \beta 1X$

Multiple linear regression extends this concept to more dimensions. This model extends the simple linear regression model to include more than one independent variable. The model can be written as:

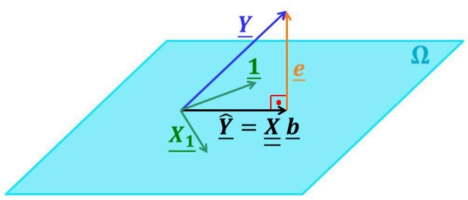
$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta pXp + \varepsilon$

where β0 is the intercept, β1 to βp are the coefficients for the independent variables X1 to Xp, and ε is the error term. The goal of multiple linear regression is to estimate the coefficients β0 to βp that minimize the RSS.

To estimate the coefficients, we use the same method of least squares as in simple linear regression. The only difference is that we now have to estimate p + 1 coefficients instead of two. We can also use the model to make predictions for new values of X. The predicted value of Y based on new values of X1 to Xp is given by:

$\hat{y} = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta pXp$

The normal equations can be extended to multiple linear regression:



Normal equations

$$\underline{e} \perp \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \Rightarrow \left\{ \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right\} = 0 \Rightarrow \sum e_i = 0$$

$$\underline{e} \perp \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \Rightarrow \left\{ \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \right\} = 0 = \sum e_i X_i \Rightarrow \underline{X_1}' \underline{e} = 0$$

We have commented on the assumptions of linear regression in class. The first assumption is linearity, which assumes that the relationship between the dependent variable and the independent variable(s) is linear. The second assumption is independence, which assumes that the observations are independent of each other. The third assumption is homoscedasticity, which assumes that the variance of the error term ε_i is constant for all values of X. The fourth assumption is normality, which assumes that the error term ε_i follows a normal distribution. As professor Monte has said before: 'Assumptions are the mother of all fuck ups.' It makes no sense to apply linear regression to data that doesn't satisfy these conditions.

ε_i	e_i
Unknown	Known
Linearity	
$E(\varepsilon_i) = 0$	$E(e_i) = 0$
Constant variance	
$V(\varepsilon_i) = \sigma^2$	$V(e_i) = \sigma^2 \cdot (1 - h_{ii})$
Normality	
$N(0, \sigma)$	$N(0, \sigma_{e_i})$
Independence	
Yes	No

To test whether this assumption is satisfied one can make an analysis of the residuals e_i, which behave in a very similar manner as the errors ε_i.

One question was left to be answered, and it is how to assess the performance of a linear regression model. One way to do this is by calculating the coefficient of determination (R^2), which measures the proportion of the variation in the dependent variable that is explained by the independent variable(s). R^2 ranges from 0 to 1, with higher values indicating a better fit. Another way to assess performance is by using hypothesis testing to determine whether the coefficients are significantly different from zero via hypothesis testing or ANOVA.