The Analytics Edge FALL 2020

Regression concepts

In class exercise: Week 2

1. Suppose we use linear regression on a training dataset and get a sum of squared error equal to e_1. Using this model, we obtain a sum of squared error in the test set equal to e_2. We add some more new features (predictor variables) in the dataset and refit the model. Select the best option.

   (a) The training error e_1 always decreases or remains the same.

   (b) The training error e_1 always increases or remains the same.

   (c) The test error e_2 always decreases or remains the same.

   (d) The test error e_2 always increases or remains the same.

   Explanation: If we fit a linear model with more explanatory variables on the same training set, then the training error will decrease (or t worst remain the same). It cannot increase. In case of the test set, since the model is not optimizing to fit on that set, error may increase or decrease.

2. The statement, "the p-value is 0.001" is equivalent to the statement that "there is a 0.1% probability that the null hypothesis is true".
   True or False?
   Explanation: The null hypothesis is either true or false; since it is a statement about one of the parameters which is fixed. The randomness is in the sample you observe from the data generated using that fixed model.

3. If you get a p-value of 0.1, it means that when the null hypothesis is true, a value of the test statistic as or more extreme than what was observed occurs in about 10% of all samples.
   True or False?
   Explanation: A p-value of 0.1 for a test statistic value t_0 implies that if you sample from the population under the null hypothesis and create a test statistic, and you do this a 1000 times, say, then you will observe a test statistic value as extreme as t_0 about 100 or fewer times.

4. Suppose we solve a linear regression problem and obtain the optimal estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$. The average value of the residuals with these optimal estimates will be always 0.

   True or False?

   Answer. The estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by minimizing $Q(\boldsymbol{\beta})$ where

   $$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2.$$

   When we minimize $Q(\boldsymbol{\beta})$, taking derivative with respect to β_0 and equating to zero we have

   $$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}) = 0.$$

   Since the optimal solution $\hat{\boldsymbol{\beta}}$ satisfies the above, we have

   $$0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip}) = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i.$$

5. Suppose a 95% confidence interval for the slope of a linear regression of y on x is given by $-3.5 < \beta < -0.5$. Then a two- sided test of the hypothesis $H_0 : \beta = -1$ would result in rejection of H_0 at the 1% level of significance.

   True or False?

   Explanation: Since $H_0 : \beta = -1$ would not be rejected at $\alpha = 0.05$, it would not be rejected at $\alpha = 0.01$.