# Statistics mini-course

1 декабря 2024 г.   10:44

1. Hypothesis testing is based on our samples of populations. Null hypothesis is an assumption of two samples of one population are same. Alternate hypothesis - there is a difference between samples. It can also refer to models, samples or populations.
2. T-value indicates how many standard errors your mean is away from hypothesized mean.
3.
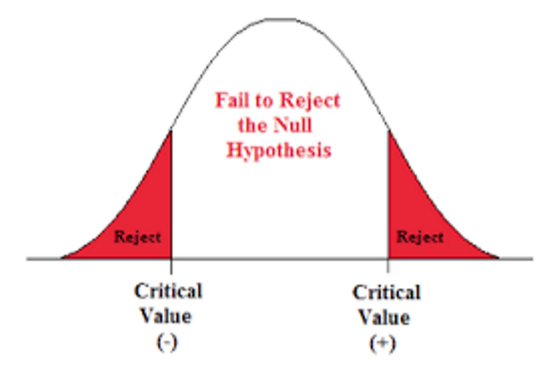
4. In terms of distribution. Area taken by integral from t-value to +inf is a p-value which shows us a probability of t-value being that or more amount of standard errors. Area from -t to +t is the area of null hypothesis and shows us the probability of samples being same
5. In terms of linear regression - T-test is made for every coefficient, where null hypothesis is coefficient being 0 aka does not effect dependent variable at all.
6. P-value shows us how likely our data to appear by chance
7. T-value is an intermidiate step before p-value. It shows us how far our coefficient from zero in terms of standard error of coefficient.
8. Standard error of the coefficient can be calculated using RSS(Residual sum of squares) the sum of differences from estimated and real values squared
9.

Residual Sum of Squares

Residual Sum of Squares measures the extent of variability of observed data not predicted by the regression model.

$RSS = e1^2 + e2^2 + e3^2 + e4^2 + e5^2$

e = Error or residue
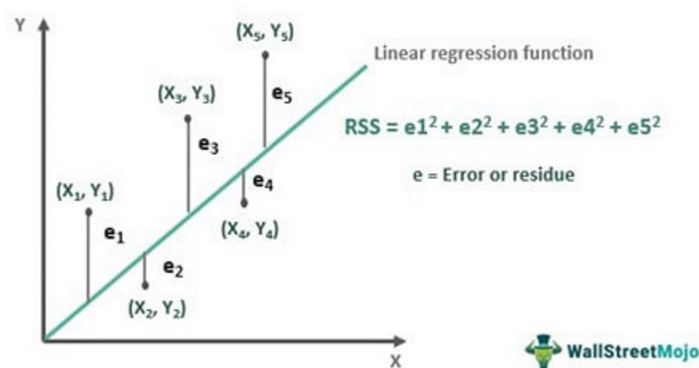
10. T-value of coefficient is in certain way presents our data.

1. F-statistics
2. We have 2 hypothesis.
3. H0: all coefficients from our model are 0, so y=a, where a=const
4. H1: at least one of coefficients is not 0, so at least y=a+bi*xi
5. RSS0 - RSS of H0
6. RSS1 - RSS of H1
7. RSS0 is always greater or equal than RSS1. The question - is it significantly larger?
8. How to measure this significance? F-value!
9. F-value = (RSS0-RSS1)/RSS1 * (n-p-1)/p, where p is number of coefficients, n is number of observations in our sample.

10. n - p -1 is number of degrees of freedom for RSS. Since we have p independent predictors and one for intercept we partially determine p+1 of the residuals. That's why we only have n - p -1 freedom degrees for residuals
11. Or F-statistics = MSR/MSE, where MSR - is mean RSS or average explained variation, MSE - mean Squared error or average unexplained variation
12. F-value basically describes how well the explanatory function of your model do