MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

Linear Regression Models

Segment 1 – Simple Linear Regression Model

Topic 5 – Accuracy of the Model

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Topics

1. Residual Standard Error
2. R2 Statistic
3. Summary

Residual Standard Error

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.
- In an SLRM $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, the random error term prevents a perfect prediction even if the population coefficients are exactly known.

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.
- In an SLRM $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, the random error term prevents a perfect prediction even if the population coefficients are exactly known.
- The RSE is an in-sample estimate of the standard deviation σ of the random error term ε:

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.
- In an SLRM $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, the random error term prevents a perfect prediction even if the population coefficients are exactly known.
- The RSE is an in-sample estimate of the standard deviation σ of the random error term ε: RSE $= \sqrt{\frac{1}{n-2} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}}$.

Residual Standard Error

- The residual standard error (RSE) is a measure of lack of fit of the model to the data.
- In an SLRM $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, the random error term prevents a perfect prediction even if the population coefficients are exactly known.
- The RSE is an in-sample estimate of the standard deviation σ of the random error term ε: RSE $= \sqrt{\frac{1}{n-2} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}}$.
- The RSE is the amount by which the response will deviate from the true regression line on an average.

R^2 Statistic

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1:

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model?

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM;

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is residual sum of squares (RSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}\right)\right)^2$.

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n}\left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is residual sum of squares (RSS) = $\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n}\left(y^{(i)} - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}\right)\right)^2$.
- The R^2 statistic is defined as

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is residual sum of squares (RSS) = $\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}\right)\right)^2$.
- The R^2 statistic is defined as $\frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$.

R^2 Statistic

- Suppose we have a dataset with n samples and want a simple model to predict the response Y from a single predictor X_1: the average model $\hat{y}^{(i)} = \bar{y}_n$.
- What is the error associated with this model? Total sum of squares (TSS) = $\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n}\left(y^{(i)} - \bar{y}_n\right)^2$.
- Suppose now we build an SLRM; the associated error is residual sum of squares (RSS) = $\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n}\left(y^{(i)} - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)}\right)\right)^2$.
- The R^2 statistic is defined as $\frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$.
- The R^2 statistic varies between 0 & 1 is and is a measure of the variability in the response Y that the SLRM (built using the predictor X_1) is able to explain.

Summary

Summary

Different ways to assess the accuracy of an SLRM using residual standard error (RSE) and R^2 statistic.