Linear Regression Models

Segment 2 – Multiple Linear Regression Model

Topic 3 – Feature Engineering and Regularization

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

Topics

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

1. Feature Engineering: Centering, Standardization, and Logarithmic Transformation

2. Regularization: Ridge and Lasso

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

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

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- If we center the predictor values (subtract the sample mean), then the coefficient estimates become mote interpretable.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- If we center the predictor values (subtract the sample mean), then the coefficient estimates become mote interpretable.
- With centering of the predictors, the MLRM predicts

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- If we center the predictor values (subtract the sample mean), then the coefficient estimates become mote interpretable.
- With centering of the predictors, the MLRM predicts

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \left(livingArea - \text{mean}(livingArea)\right) + \hat{\beta}_2 \left(age - \text{mean}(age)\right).$$

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

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

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- If we center the predictor values (subtract the sample mean), then the coefficient estimates become mote interpretable.
- With centering of the predictors, the MLRM predicts

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \left(livingArea - \text{mean}(livingArea)\right) + \hat{\beta}_2 \left(age - \text{mean}(age)\right).$$

- The intercept estimate can now be interpreted as the (approximate) average value of $price$ around the average value of $livingArea$ and average value of age.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

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

- In MLRMs, it is helpful to standardize the predictors (subtract the sample mean and divide by sample standard deviation) if there are different scales (units) present in the features.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

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

- In MLRMs, it is helpful to standardize the predictors (subtract the sample mean and divide by sample standard deviation) if there are different scales (units) present in the features.
- For example, in the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $lotSize$ as the predictors, it would be helpful to standardize the predictors.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- In MLRMs, it is helpful to standardize the predictors (subtract the sample mean and divide by sample standard deviation) if there are different scales (units) present in the features.
- For example, in the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $lotSize$ as the predictors, it would be helpful to standardize the predictors.
- The interpretations of the coefficient estimates have to be made with respect to the scaled predictor values.

Feature Engineering: Centering, Standardization, and Logarithmic Transformation

- In MLRMs, it is helpful to standardize the predictors (subtract the sample mean and divide by sample standard deviation) if there are different scales (units) present in the features.
- For example, in the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and $lotSize$ as the predictors, it would be helpful to standardize the predictors.
- The interpretations of the coefficient estimates have to be made with respect to the scaled predictor values.
- The response variable may be logarithmically transformed if it cannot be predicted to be negative (for example, height, weight etc.).

Regularization: Ridge and Lasso

Regularization: Ridge and Lasso

- How can we choose predictors that *explain* the response variable the most?

Regularization: Ridge and Lasso

- How can we choose predictors that *explain* the response variable the most?
- We achieve this by adding a regularization term to the residual sum of squares (RSS) and minimize it:

Regularization: Ridge and Lasso

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

- How can we choose predictors that *explain* the response variable the most?
- We achieve this by adding a regularization term to the residual sum of squares (RSS) and minimize it:

$$\underbrace{\sum_{i=1}^{n}\left(y^{(i)}-\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_p x_p^{(i)}\right)\right)^2}_{\text{RSS}}+\begin{cases}\lambda\sum_{j=1}^{p}|\beta_j|^2 & (\text{ridge})\\ \quad\text{or} & \\ \lambda\sum_{j=1}^{p}|\beta_j| & (\text{lasso}),\end{cases}$$

where $\lambda > 0$ is the strength of regularization.

Regularization: Ridge and Lasso

- How can we choose predictors that *explain* the response variable the most?
- We achieve this by adding a regularization term to the residual sum of squares (RSS) and minimize it:

$$\underbrace{\sum_{i=1}^{n}\left(y^{(i)}-\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_p x_p^{(i)}\right)\right)^2}_{\text{RSS}}+\begin{cases}\lambda\sum_{j=1}^{p}|\beta_j|^2 & (\text{ridge})\\ \quad\text{or} & \\ \lambda\sum_{j=1}^{p}|\beta_j| & (\text{lasso}),\end{cases}$$

  where $\lambda > 0$ is the strength of regularization.
- Both ridge and lasso approaches for regularization shrink the coefficient estimates towards 0 but lasso typically yields a much smaller subset of nonzero coefficient estimates.