Linear Regression Models

Segment 2 – Multiple Linear Regression Model

Topic 2 – Accuracy of Ordinary Least Squares Model and Estimators

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

Topics

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

1. Interpretation of OLS Estimators
2. Accuracy of the Coefficient Estimates
3. Accuracy of the Model: R^2 and Adjusted R^2 Statistic

Interpretation of OLS Estimators

Interpretation of OLS Estimators

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

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age.$

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age.$
- $\hat{\beta}_0$ is the predicted $price$ when both predictors $livingArea$ and age are equal to 0.

Interpretation of OLS Estimators

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.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age.$
- $\hat{\beta}_0$ is the predicted $price$ when both predictors $livingArea$ and age are equal to $0.$
- What about $\hat{\beta}_1$?

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age.$
- $\hat{\beta}_0$ is the predicted $price$ when both predictors $livingArea$ and age are equal to 0.
- What about $\hat{\beta}_1$? It is the change in the predicted $price$ for a 1 unit increase in $livingArea$ while keeping the remaining predictor age fixed:

Interpretation of OLS Estimators

- Suppose that we consider the *SaratogaHouses* dataset with $price$ as the response and $livingArea$ and age as the predictors.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age.$
- $\hat{\beta}_0$ is the predicted $price$ when both predictors $livingArea$ and age are equal to 0.
- What about $\hat{\beta}_1$? It is the change in the predicted $price$ for a 1 unit increase in $livingArea$ while keeping the remaining predictor age fixed:

$$\begin{cases} \widehat{price}_{\text{old}} & = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 age \\ \widehat{price}_{\text{new}} & = \hat{\beta}_0 + \hat{\beta}_1 \left(livingArea + 1\right) + \hat{\beta}_2 age \end{cases} \Rightarrow \widehat{price}_{\text{new}} - \widehat{price}_{\text{old}} = \hat{\beta}_1.$$

Interpretation of OLS Estimators

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels:

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$

Interpretation of OLS Estimators

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

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have the same interpretation as before.

Interpretation of OLS Estimators

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

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have the same interpretation as before.
- What about $\hat{\beta}_2$?

Interpretation of OLS Estimators

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts
  $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have the same interpretation as before.
- What about $\hat{\beta}_2$? it is the difference between the predicted price of a hot air-heated house and the predicted price of an electric-heated house (reference level) with the same living area:

Interpretation of OLS Estimators

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

- Suppose now that we consider $price$ as the response and $livingArea$ and $heating$ as the predictors.
- Note that the predictor $heating$ is categorical with three levels: (1) `electric` (2) `hot air` (3) `hot water/steam`.
- Multiple linear regression model (MLRM) predicts $\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2 heatinghotair + \hat{\beta}_3 heatinghotwater/steam.$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have the same interpretation as before.
- What about $\hat{\beta}_2$? it is the difference between the predicted price of a hot air-heated house and the predicted price of an electric-heated house (reference level) with the same living area:

$$\underbrace{[\hat{\beta}_0 + \hat{\beta}_1 livingArea + \hat{\beta}_2]}_{\widehat{price}_{\text{hot air}}} - \underbrace{[\hat{\beta}_0 + \hat{\beta}_1 livingArea]}_{\widehat{price}_{\text{electric}}} = \hat{\beta}_2.$$

Accuracy of the Coefficient Estimates

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased:

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\beta} - \beta] = \mathbf{0}$.

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\beta} - \beta] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\beta}$:

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1},$$

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\beta} - \beta] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\beta}$:

cov$(\hat{\beta}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$, where $\sigma^2 \approx \frac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}$.

Accuracy of the Coefficient Estimates

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

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\beta} - \beta] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\beta}$:

$$\text{cov}(\hat{\beta}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \frac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates,

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\beta} - \beta] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\beta}$:

$$\text{cov}(\hat{\beta}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \tfrac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates, also called their standard errors (SE),

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \tfrac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates, also called their standard errors (SE), can be used to calculate confidence intervals (CI) for population coefficient parameters β_j:

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \tfrac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates, also called their standard errors (SE), can be used to calculate confidence intervals (CI) for population coefficient parameters β_j: a 95% CI for β_j is

Accuracy of the Coefficient Estimates

- It can be shown that the OLS coefficient estimates for an MLRM are unbiased: $E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$.
- The variances of the OLS coefficient estimates are the diagonal elements of the covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, \text{ where } \sigma^2 \approx \frac{1}{n-(p+1)} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{RSS}.$$

- The standard deviations of the OLS coefficient estimates, also called their standard errors (SE), can be used to calculate confidence intervals (CI) for population coefficient parameters β_j: a 95% CI for β_j is $\left[\hat{\beta}_j - 1.96 \times SE(\hat{\beta}_j),\ \hat{\beta}_j + 1.96 \times SE(\hat{\beta}_j)\right]$.

Accuracy of the Coefficient Estimates

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.
- If the p-value is smaller than a threshold, typically 0.05, then

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.
- If the p-value is smaller than a threshold, typically 0.05, then we reject the null hypothesis, and conclude that the jth predictor contributes to the linear model.

Accuracy of the Coefficient Estimates

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.
- If the p-value is smaller than a threshold, typically 0.05, then we reject the null hypothesis, and conclude that the jth predictor contributes to the linear model.
- A better hypothesis test, when the number of predictors is large, is the F test,

Accuracy of the Coefficient Estimates

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

- In order to check if there is a relationship between the response and a particular predictor, standard errors can be used to perform hypothesis tests on the population coefficient parameters.
- Recall that the p-value associated with the coefficient estimate $\hat{\beta}_j$ is a measure of how likely it is to observe that particular value of the estimate assuming that the null hypothesis about the corresponding population coefficient parameter ($\beta_j = 0$) is true.
- If the p-value is smaller than a threshold, typically 0.05, then we reject the null hypothesis, and conclude that the jth predictor contributes to the linear model.
- A better hypothesis test, when the number of predictors is large, is the F test, in which the null hypothesis is that all population coefficients are zeros except the intercept.

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.
- However, the R^2 statistic of a model can always be increased by adding more yet insignificant predictors.

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.
- However, the R^2 statistic of a model can always be increased by adding more yet insignificant predictors.
- Models with too many predictors over fit the data and typically do not perform well on unseen data.

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.
- However, the R^2 statistic of a model can always be increased by adding more yet insignificant predictors.
- Models with too many predictors over fit the data and typically do not perform well on unseen data.
- Adjusted R^2 statistic is a measure which penalizes the addition of predictors. It is the proportion of variance in the response explained by the linear model built using predictors that *actually* affect the response:

Accuracy of the Model: R^2 and Adjusted R^2 Statistic

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

- The R^2 statistic varies between 0 and 1, and is a measure of the variability in the response Y that the MLRM (built using the predictors $X_1, X_2, \ldots, X_p$) is able to explain.
- However, the R^2 statistic of a model can always be increased by adding more yet insignificant predictors.
- Models with too many predictors over fit the data and typically do not perform well on unseen data.
- Adjusted R^2 statistic is a measure which penalizes the addition of predictors. It is the proportion of variance in the response explained by the linear model built using predictors that *actually* affect the response: $R^2_{\text{adj}} = 1 - \left[\frac{(1-R^2)(n-1)}{n-(p+1)}\right].$