Linear Regression Models

Segment 3 – Other Considerations in the MLRM

Topic 1 – Confounding and Collinearity: Correlation Matrix & Variance Inflation Factor (VIF)

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

Topics

1. Confounding and Collinearity: Basic Ideas
2. When Does Confounding Arise?
3. Detecting Collinearity: Correlation Matrix
4. Quantifying Collinearity: Variance Inflation Factor (VIF)

Confounding and Collinearity: Basic Ideas

Confounding and Collinearity: Basic Ideas

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

- Confounding:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan’s death rate is higher than that of countries like Vietnam;

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for;

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.
- Structural collinearity:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.
- Structural collinearity: when model is created with correlated predictors.

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.
- Structural collinearity: when model is created with correlated predictors.
- Data collinearity:

Confounding and Collinearity: Basic Ideas

- Confounding: occurs when a third variable that distorts the observed relationship between the predictor and response.
- Example: Japan's death rate is higher than that of countries like Vietnam; before concluding that Japan is a riskier place to live, confounding factors such as *age* need to be accounted for; median age of Japan is much higher than that of, say, Vietnam.
- Collinearity: occurs when predictors are highly correlated such that it is difficult to distinguish their effect on the response.
- Also referred to as multicollinearity or ill-conditioning.
- Structural collinearity: when model is created with correlated predictors.
- Data collinearity: when data comprises correlated predictors.

When Does Confounding Arise?

When Does Confounding Arise?

- Indication bias:

When Does Confounding Arise?

- Indication bias: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.

When Does Confounding Arise?

- Indication bias: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.
- Selection bias:

When Does Confounding Arise?

- Indication bias: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.
- Selection bias: the effect of a trial drug for treating a particular medical condition may be affected by the imbalance between the groups.

When Does Confounding Arise?

- Indication bias: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.
- Selection bias: the effect of a trial drug for treating a particular medical condition may be affected by the imbalance between the groups.
- Recall bias:

When Does Confounding Arise?

- <u>Indication bias</u>: the effect of a trial drug for treating a particular medical condition may differ substantially between those who have the condition and those who do not.
- <u>Selection bias</u>: the effect of a trial drug for treating a particular medical condition may be affected by the imbalance between the groups.
- <u>Recall bias</u>: study participants who have cancer may be more likely to recall being a smoker.

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%
Diabetes (%)	12%	32%

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%
Diabetes (%)	12%	32%
BMI (mean)	24	26

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%
Diabetes (%)	12%	32%
BMI (mean)	24	26

Predictors that are *imbalanced* among the two groups:

Confounding Example

Which of the following are likely to be confounding factors for the hypothesis that high cholesterol food is associated with heart disease?

Factor	Low cholesterol food	High cholesterol food
Smoker (%)	10%	30%
Age (mean years)	42	41
Daily exercise (%)	25%	28%
Diabetes (%)	12%	32%
BMI (mean)	24	26

Predictors that are *imbalanced* among the two groups: Smoker, Diabetes are potential confounders.

Collinearity

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example:

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction:

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent:

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent:
  $\widehat{height} = 30 + 3 \times age1 + 0 \times age2 =$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent: $\widehat{height} = 30 + 3 \times age1 + 0 \times age2 = 30 + 2 \times age1 + 12 \times age2 =$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent: $\widehat{height} = 30 + 3 \times age1 + 0 \times age2 = 30 + 2 \times age1 + 12 \times age2 = 30 + 1 \times age1 + 24 \times age2.$

Collinearity

- Collinearity, an extreme case of confounding, implies an exact linear relationship between predictors.
- Example: predictor $age1$ in years is collinear with the predictor $age2$ in months because $age1 = 12 \times age2 \Rightarrow$ columns of design matrix $\mathbf{X} = \begin{bmatrix} age1 & age2 \end{bmatrix}$ are linearly dependent $\Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist.
- Collinearity poses no problem for prediction: the model $\widehat{height} = \hat{\beta}_0 + \hat{\beta}_1 \times age1 + \hat{\beta}_2 \times age2$ has theoretically infinitely many solutions but all result in the same predicted height.
- For example, the following solutions are all equivalent: $\widehat{height} = 30 + 3 \times age1 + 0 \times age2 = 30 + 2 \times age1 + 12 \times age2 = 30 + 1 \times age1 + 24 \times age2.$
- Quantifying individual effects of collinear predictors is a problem.

Detecting Collinearity: Correlation Matrix

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.
- Given a dataset, sample correlation measure between two predictors $\mathbf{x}_1$ and $\mathbf{x_2} \Rightarrow$

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.
- Given a dataset, sample correlation measure between two predictors $\mathbf{x}_1$ and $\mathbf{x}_2 \Rightarrow$ mean-centering them $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2 \Rightarrow$

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.
- Given a dataset, sample correlation measure between two predictors $\mathbf{x}_1$ and $\mathbf{x_2} \Rightarrow$ mean-centering them $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2 \Rightarrow \rho = \frac{\tilde{\mathbf{x}}_1^{\mathrm{T}}\tilde{\mathbf{x}}_2}{\|\tilde{\mathbf{x}}_1\|\|\tilde{\mathbf{x}}_2\|}$ is in between -1 and 1.

Detecting Collinearity: Correlation Matrix

- Collinearity can be detected by studying the sample correlation matrix of continuous predictors.
- Given a dataset, sample correlation measure between two predictors $\mathbf{x}_1$ and $\mathbf{x_2}$ $\Rightarrow$ mean-centering them $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ $\Rightarrow \rho = \dfrac{\tilde{\mathbf{x}}_1^{\mathrm{T}}\tilde{\mathbf{x}}_2}{\|\tilde{\mathbf{x}}_1\|\|\tilde{\mathbf{x}}_2\|}$ is in between -1 and 1.
- Correlation matrix for some continuous predictors from the `saratogaHouses` dataset:

	livingArea	lotSize	age	landValue	bedrooms	rooms
livingArea	1.00	0.16	-0.17	0.42	0.66	0.73
lotSize	0.16	1.00	-0.02	0.06	0.11	0.14
age	-0.17	-0.02	1.00	-0.02	0.03	-0.08
landValue	0.42	0.06	-0.02	1.00	0.20	0.30
bedrooms	0.66	0.11	0.03	0.20	1.00	0.67
rooms	0.73	0.14	-0.08	0.30	0.67	1.00

Consequences of Correlated Predictors

In model built with correlated predictors:

Consequences of Correlated Predictors

In model built with correlated predictors:

- regression coefficients estimates will change dramatically depending on which correlated predictors are included or not;

Consequences of Correlated Predictors

In model built with correlated predictors:

- regression coefficients estimates will change dramatically depending on which correlated predictors are included or not;
- coefficient estimates for predictors with known strong relationships with the response will not be accurate;

Consequences of Correlated Predictors

In model built with correlated predictors:

- regression coefficients estimates will change dramatically depending on which correlated predictors are included or not;
- coefficient estimates for predictors with known strong relationships with the response will not be accurate;
- standard errors of the coefficients estimates will be (relatively) large;

Consequences of Correlated Predictors

In model built with correlated predictors:

- regression coefficients estimates will change dramatically depending on which correlated predictors are included or not;
- coefficient estimates for predictors with known strong relationships with the response will not be accurate;
- standard errors of the coefficients estimates will be (relatively) large;
- wider confidence intervals for coefficients.

Quantifying Collinearity: Variance Inflation Factor (VIF)

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

Quantifying Collinearity: Variance Inflation Factor (VIF)

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

- If all the predictors are perfectly collinear,

Quantifying Collinearity: Variance Inflation Factor (VIF)

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

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.

Quantifying Collinearity: Variance Inflation Factor (VIF)

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

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.
- The *tolerance* of the predictor which is regressed upon the others is $1 - R^2$.

Quantifying Collinearity: Variance Inflation Factor (VIF)

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

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.
- The *tolerance* of the predictor which is regressed upon the others is $1 - R^2$.
- A small value of the *tolerance* (< 0.1, for example) indicates that the predictor under consideration is highly correlated with the other predictors.

Quantifying Collinearity: Variance Inflation Factor (VIF)

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.
- The *tolerance* of the predictor which is regressed upon the others is $1 - R^2$.
- A small value of the *tolerance* (< 0.1, for example) indicates that the predictor under consideration is highly correlated with the other predictors.
- The variance inflation factor (VIF) of the predictor under consideration is $1/tolerance = 1/(1 - R^2)$.

Quantifying Collinearity: Variance Inflation Factor (VIF)

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

- If all the predictors are perfectly collinear, then the R^2 metric when one predictor is regressed upon the others will be exactly 1.
- The *tolerance* of the predictor which is regressed upon the others is $1 - R^2$.
- A small value of the *tolerance* (< 0.1, for example) indicates that the predictor under consideration is highly correlated with the other predictors.
- The variance inflation factor (VIF) of the predictor under consideration is $1/tolerance = 1/(1 - R^2)$.
- A large vale of VIF (> 10, for example) indicates additional study about the correlation between predictors.

Summary

Summary

- Describe and differentiate between confounding and collinearity.

Summary

- Describe and differentiate between confounding and collinearity.
- Describe how correlation matrix can be used to detect collinearity.

Summary

- Describe and differentiate between confounding and collinearity.
- Describe how correlation matrix can be used to detect collinearity.
- Interpret variance inflation factor (VIF) for quantifying collinearity.