Regression Assumptions

Derick Smith

MCSC 6020 G, Fall 2019

Fully Built Linear Regression Model

1. We cleaned our data
2. Performed variable selection
3. Constructed a final model

$$y = \hat{y} + \varepsilon$$
$$\hat{y} = \sum_{j=0}^{N} \left\{ \beta_{j,k} x_k^{j \cdot \omega_{j,k}} \right\}_{\forall k}$$
$$\omega_{j,k} = \{0,1\}$$
$$\varepsilon \overset{iid}{\sim} \mathcal{N}(\mu = 0, \sigma^2)$$

4. Confirmed the validity of our β_i coefficients

$$\begin{aligned} \boldsymbol{H_0} : &\quad \beta_{j,k} \neq 0 \quad \forall j,k \\ \boldsymbol{H_a} : &\quad \beta_{j,k} = 0 \quad \exists j,k \end{aligned}$$

$$\alpha > p \quad \therefore \text{ fail to reject } \boldsymbol{H_0}$$

Fully Built Linear Regression Model

1. We cleaned our data
2. Performed variable selection
3. Constructed a final model

$$
\begin{aligned}
y &= \hat{y} + \varepsilon \\
\hat{y} &= \sum_{j=0}^{N} \left\{ \beta_{j,k} x_k^{j \cdot \omega_{j,k}} \right\}_{\forall k} \\
\omega_{j,k} &= \{0,1\} \\
\varepsilon &\overset{iid}{\sim} \mathcal{N}(\mu = 0, \sigma^2)
\end{aligned}
$$

4. Confirmed the validity of our β_i coefficients

$$
\begin{aligned}
\boldsymbol{H_0}: &\quad \beta_{j,k} \neq 0 \quad \forall j,k \\
\boldsymbol{H_a}: &\quad \beta_{j,k} = 0 \quad \exists j,k
\end{aligned}
$$

$$\alpha > p \quad \therefore \text{ fail to reject } \boldsymbol{H_0}$$

Fully Built Linear Regression Model

1. We cleaned our data
2. Performed variable selection
3. Constructed a final model

$$y = \hat{y} + \varepsilon$$

$$\hat{y} = \sum_{j=0}^{N} \left\{ \beta_{j,k} x_k^{j \cdot \omega_{j,k}} \right\}_{\forall k}$$

$$\omega_{j,k} = \{0, 1\}$$

$$\varepsilon \overset{iid}{\sim} \mathcal{N}(\mu = 0, \sigma^2)$$

4. Confirmed the validity of our β_i coefficients

$$\begin{aligned} \boldsymbol{H_0}: &\quad \beta_{j,k} \neq 0 \quad \forall j,k \\ \boldsymbol{H_a}: &\quad \beta_{j,k} = 0 \quad \exists j,k \end{aligned}$$

$$\alpha > p \quad \therefore \text{ fail to reject } \boldsymbol{H_0}$$

Fully Built Linear Regression Model

1. We cleaned our data
2. Performed variable selection
3. Constructed a final model

$$
\begin{aligned}
y &= \hat{y} + \varepsilon \\
\hat{y} &= \sum_{j=0}^{N} \left\{ \beta_{j,k} x_k^{j \cdot \omega_{j,k}} \right\}_{\forall k} \\
\omega_{j,k} &= \{0,1\} \\
\varepsilon &\overset{iid}{\sim} \mathcal{N}(\mu = 0, \sigma^2)
\end{aligned}
$$

4. Confirmed the validity of our β_i coefficients

$$
\begin{aligned}
\boldsymbol{H_0} : &\quad \beta_{j,k} \neq 0 \quad \forall j,k \\
\boldsymbol{H_a} : &\quad \beta_{j,k} = 0 \quad \exists j,k
\end{aligned}
$$

$$\alpha > p \;\; \therefore \text{ fail to reject } \boldsymbol{H_0}$$

Work's Not Done

- We made assumptions along the way
- In particular for errors:
  - Randomly independent
  - Constant variance
  - Normally distributed with mean zero

Work's Not Done

- We made assumptions along the way
- In particular for errors:
  - Randomly independent
  - Constant variance
  - Normally distributed with mean zero

Work's Not Done

- We made assumptions along the way
- In particular for errors:
  - Randomly independent
  - Constant variance
  - Normally distributed with mean zero

Work's Not Done

- We made assumptions along the way
- In particular for errors:
  - Randomly independent
  - Constant variance
  - Normally distributed with mean zero

Work's Not Done

- We made assumptions along the way
- In particular for errors:
  - Randomly independent
  - Constant variance
  - Normally distributed with mean zero

Test Assumptions

- Randomly independent:
  (x_i, ε_i) plot structureless
- Constant variance:
  (x_i, ε_i) plot points spread rectangularly
- Normally distributed with mean zero:
  Q-Q plot close to one-to-one correlation between experimental and theoretical

Test Assumptions

- Randomly independent:

  $(x_{i,}\varepsilon_i)$ plot structureless

- Constant variance:

  $(x_{i,}\varepsilon_i)$ plot points spread rectangularly

- Normally distributed with mean zero:

  Q-Q plot close to one-to-one correlation between experimental and theoretical

Test Assumptions

- Randomly independent:

  $(x_{i,}\varepsilon_i)$ plot structureless
- Constant variance:

  $(x_{i,}\varepsilon_i)$ plot points spread rectangularly
- Normally distributed with mean zero:

  Q-Q plot close to one-to-one correlation between experimental and theoretical

Transform Model (If Invalid)

- If not structureless:
  - may exist autocorrelation
  - significant modification may be necessary beyond linear regression
- If non-normal:
  - Change flexibility using resampling (e.g. k-fold cross-validation or bootstrapping)
- If non-constant variance (and potentially non-normal):

$$y^* = \sqrt{y}$$
$$y^* = sin^{-1}(y)$$
$$y^* = log(y)$$

Transform Model (If Invalid)

- If not structureless:
  - may exist autocorrelation
  - significant modification may be necessary beyond linear regression
- If non-normal:
  - Change flexibility using resampling (e.g. k-fold cross-validation or bootstrapping)
- If non-constant variance (and potentially non-normal):

$$y^* = \sqrt{y}$$

$$y^* = sin^{-1}(y)$$

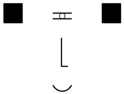
$$y^* = log(y)$$

Transform Model (If Invalid)

- If not structureless:
  - may exist autocorrelation
  - significant modification may be necessary beyond linear regression
- If non-normal:
  - Change flexibility using resampling (e.g. k-fold cross-validation or bootstrapping)
- If non-constant variance (and potentially non-normal):

$$
\begin{aligned}
y^* &= \sqrt{y} \\
y^* &= sin^{-1}(y) \\
y^* &= log(y)
\end{aligned}
$$

Let’s Run Some Code



Summary

- Build your model thoughtfully
- Test your coefficients
- Test your assumptions!
- Transform as needed
- Rinse, repeat

$$\mathbb{Q} \bigcup \varepsilon \; \tau \; \mathbb{I} \; \mathscr{O} \; \eta \; \varsigma ?$$