Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Assumptions and Diagnostics

1

About This Lesson

2

Multiple Linear Regression: Model

Data: $\{(x_{1,1}, \dots, x_{1,p}), y_1\}, \dots, \{(x_{n,1}, \dots, x_{n,p}), y_n\}$

Model: $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i,\ i = 1, \dots, n$

Assumptions:

- *Linearity/Mean Zero Assumption*: The relationship between the response variable and each predicting variable is linear. (For each $j, j = 1, \dots, p$, y_i and x_{ij} are linearly related, $i = 1, \dots, n$.) $\mathrm{E}(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1, \dots, \varepsilon_n\}$ are independent random variables
- Assumption that $\varepsilon_i \sim$ Normal *for confidence/prediction intervals, hypothesis testing*

Georgia Tech

3

Properties of the Errors & Residuals

Properties of (true) errors:

- $\mathrm{E}(\varepsilon_i) = 0$
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$
- $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathrm{I}$

Properties of the (estimated) residuals:

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$
- $\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$ (or $\mathrm{E}(\hat{\varepsilon}_i) = 0$)
- $\mathbf{V}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$ (or $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{i,i})$
  - Where $\mathbf{H}$ is the hat matrix, and $h_{i,i}$ is the *i*-th element on its diagonal

Georgia Tech

4

Properties of the Errors & Residuals

Properties of (true) errors:

- $\mathrm{E}(\varepsilon_i) = 0$
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$
- $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathrm{I}$

Properties of the (estimated) residuals:

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$
- $\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$ (or $\mathrm{E}(\hat{\varepsilon}_i) = 0$)
- $\mathbf{V}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$ (or $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{i,i})$)
  - Where $\mathbf{H}$ is the hat matrix, and $h_{i,i}$ is the *i*-th element on its diagonal

- **While the true errors have constant variance, the estimated residuals do not.**
- **To use the estimated residuals for assessing the model assumptions, we need to standardize:**

$$r_i = \hat{\varepsilon}_i / \left(\hat{\sigma}\sqrt{1 - h_{i,i}}\right)$$

Georgia Tech

5

Residuals Analysis

Standardized Residual Values: $r_i = \hat{\varepsilon}_i / \left(\hat{\sigma}\sqrt{1 - h_{i,i}}\right)$

Graphical assessment of MLR assumptions:

- Plot standardized residuals r_i against each predictor
  - *Linearity*
- Plot standardized residuals r_i against fitted values
  - *Constant Variance & Independence*
- QQ normal plot & histogram
  - *Normality*

Georgia Tech

6

Residuals Analysis

Standardized Residual Values: $r_i = \hat{\varepsilon}_i / \left(\hat{\sigma}\sqrt{1 - h_{i,i}}\right)$

Graphical assessment of MLR assumptions:

- Plot standardized residuals r_i against each predictor
  - *Linearity*
- Plot standardized residuals r_i against fitted values
  - *Constant Variance & Independence*
- QQ normal plot & histogram
  - *Normality*

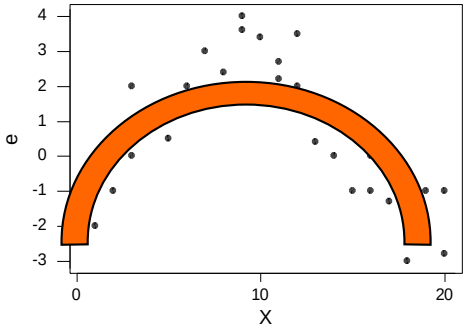
- **We evaluate the normality assumption using the residuals, not the response variable.**
- **We do not check the predicting variables for normality.**
- **However, if the distribution of a predicting variable is strongly skewed, it is possible that the linearity assumption with respect to that variable will not hold.**

Georgia Tech

7

Residual Analysis: Linearity Assumption

Linearity: Plot the residuals against each predicting variable.



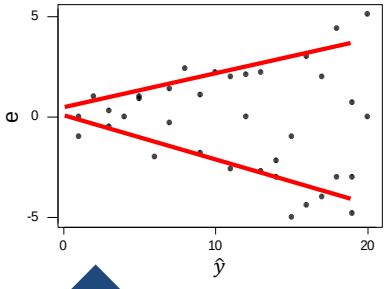
This shows that there may be a non-linear relationship between *X* and *Y*.

Georgia Tech

8

Residual Analysis: Constant Variance Assumption

Constant Variance: Plot the residuals against fitted values.

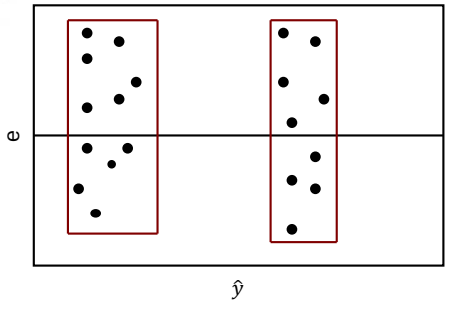


The residuals show larger variance as the predicting variable increases.

Here, it is an example for which σ^2 is not constant.

Residual Analysis: Independence Assumption

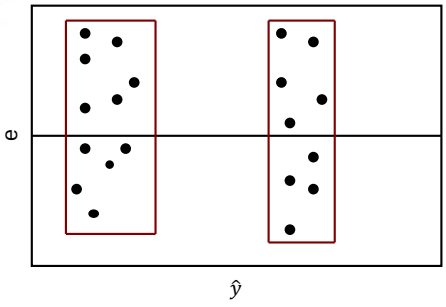
Independence *(uncorrelated errors)***:** Plot the residuals against fitted values.



- There are clusters of residuals.
- The independence assumption does not hold.

Residual Analysis: Independence Assumption

Independence *(uncorrelated errors)*: Plot the residuals against fitted values.



- There are clusters of residuals.
- The independence assumption does not hold.

- **Using residual analysis, we are actually checking for uncorrelated errors, not independence.**
- **Independence is a more complicated matter. If the data are from a randomized experiment, then independence holds, but most data are from observational studies.**
- **We commonly correct for selection bias in observational studies using controlling variables.**

Georgia Tech

11

Checking the Assumption of Normality

One way to check this assumption in a regression is using a **Normal Probability (Q-Q) Plot**

y-axis:	e_i
x-axis:	$\Phi^{-1}\left(\frac{r_i - 3/8}{n+1/4}\right)$

r_i = rank of e_i (between 1, n)
Φ = CDF of Normal Distribution

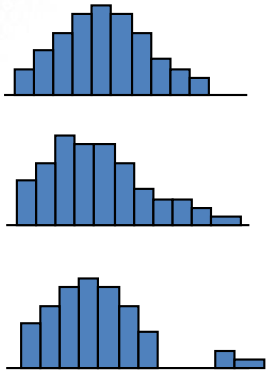
- Let the R statistical software do this for you!
- A straight line in a normal probability plot implies that the assumption is valid
- **Curviture (especially at the ends)** shows non-normality

Georgia Tech

12

Residual Analysis: Normality Assumption

A complementary approach for checking for the normality assumption is by plotting the histogram of the residuals.



Normality Assumption:
The residuals should have an approximately symmetric, unimodal distribution, with no gaps in the data.

Georgia Tech

13

Predicting Variable Transformation

- If the model fit is inadequate, it does not mean that a regression is not useful.
- One problem might be that one or more predicting variable *X* might not have a linear relationship with the response variable *Y*.
- To model the nonlinear relationship, we transform X by some nonlinear function such as

or

$$f(x) = x^a$$

$$f(x) = \log(x)$$

Georgia Tech

14

Normality Transformation

Problem: Constant variance or/and normality assumption

Solution: Transform the response variable from $\boldsymbol{y}$ to $\boldsymbol{y}^*$ via

$$\boldsymbol{y}^* = \boldsymbol{y}^{\lambda}$$

where the value of λ depends on how $\mathrm{Var}(\boldsymbol{y})$ changes as x changes.

$\sigma_y(x) \propto const$	$\lambda = 1$	(don't transform)
$\sigma_y(x) \propto \sqrt{\mu_x}$	$\lambda = 1/2$	$\boldsymbol{y}^* = \sqrt{\boldsymbol{y}}$
$\sigma_y(x) \propto \mu_x$	$\lambda = 0$	$\boldsymbol{y}^* = \ln(\boldsymbol{y})$
$\sigma_y(x) \propto \mu_x^2$	$\lambda = -1$	$\boldsymbol{y}^* = \frac{1}{\boldsymbol{y}}$

Georgia Tech

15

Outliers in Regression

A data point far from the majority of the data (in *y* and/or any *x*) may be called an *outlier*, especially if it does not follow the general trend of the rest of the data.

- Data points that are far from the means of the *X*s or near the edge of the observation space are called *leverage points*.
- A data point that is far from the means of *y* and/or an *x* is called an *influential point* if it influences the fit of the regression.
- Excluding a leverage point may or may not the regression fit significantly, thus a leverage point may or may not be an influential point.

The upshot: Sometimes there are good reasons to exclude subsets of data (e.g., errors in data entry or experimental errors). Sometimes an outlier belongs in the data. Outliers should always be examined.

Georgia Tech

16

Checking for Outliers

Cook's Distance: $\mathrm{D}_i = \frac{(\hat{\boldsymbol{Y}}_{(i)} - \hat{\boldsymbol{Y}})^{\mathrm{T}}(\hat{\boldsymbol{Y}}_{(i)} - \hat{\boldsymbol{Y}})}{(k+1)\hat{\sigma}^2}$

where $\hat{\boldsymbol{Y}}_{(i)}$ are the fitted values from the model fitted without the *i*th observation (i.e., excluding the *i*th observation from the data) and $\hat{\boldsymbol{Y}}$ are the fitted values from the model fitted with the *i*th observation (i.e., including all observations).

Cook's Distance measures how much the estimated parameter values in the regression model change when the *i*th observation is removed.

Rule of Thumb: $\mathrm{D}_i > 4/n$, $\mathrm{D}_i > 1$, OR any "large" D_i should be investigated.

Georgia Tech

17

Checking for Outliers

Cook's Distance: $\mathrm{D}_i = \frac{(\hat{\boldsymbol{Y}}_{(i)} - \hat{\boldsymbol{Y}})^{\mathrm{T}}(\hat{\boldsymbol{Y}}_{(i)} - \hat{\boldsymbol{Y}})}{(k+1)\hat{\sigma}^2}$

where $\hat{\boldsymbol{Y}}_{(i)}$ are the fitted values from the model fitted without the *i*th observation (i.e., excluding the *i*th observation from the data) and $\hat{\boldsymbol{Y}}$ are the fitted values from the model fitted with the *i*th observation (i.e., including all observations).

Cook's Distance measures how much the estimated parameter values in the regression model change when the *i*th observation is removed.

Rule of Thumb: $\mathrm{D}_i > 4/n$, $\mathrm{D}_i > 1$, OR any "large" D_i should be investigated.

- Outliers: are those few observations with much larger Cook's distance than the rest of observations;
- If a large number of outliers, then they probably point to a heavy tailed distribution rather than truly extreme values.

Georgia Tech

18

Summary

19