Common Errors in Linear Regression

Kwang Woo Ahn, PhD, Assistant Professor

Medical College of Wisconsin, Division of Biostatistics

Friday, November 9, 2012

12:00-1:00 pm

The Medical College of Wisconsin is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education for physicians.

The Medical College of Wisconsin designates this Live activity for a maximum of 1.0 *AMA PRA Category 1 Credit(s)* ™. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Hours of Participation for Allied Health Professionals

The Medical College of Wisconsin designates this activity for up to 1.0 hours of participation for continuing education for allied health professionals.

Financial Disclosure

- In accordance with the ACCME® standard for Commercial Support Number 6, all in control of content disclosed any relevant financial relationships. The following in control of content had no relevant financial relationships to disclose.

Name:	Role in Meeting:
Ruta Brazauskas, PhD	Planning Committee
Haley Montsma, BBA	Planning Committee
Kwang Woo Ahn, PhD	Speaker

Learning Objectives

- Determine the main components of linear regression
- Utilize simple graphs to check linear regression assumptions
- Identify other common errors

Evaluation Forms

Your opinion matters!

Help us plan future meetings, by completing and submitting your evaluation forms.

Thank you.

Linear regression

- Linear regression analysis is a statistical method for investigating linear relationships between response variable and explanatory variables. For example,
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.
  - Y is a response variable, X_i is an explanatory variable, ϵ is an error term, and β_i is parameters we need to estimate.
- Explanatory variables are also called independent variables, covariates, regressors, and predictors.
- A response variable is also called a dependent variable.
- In linear regression, the main focus is estimating β_i's and seeing whether they are significant.

Interpretation

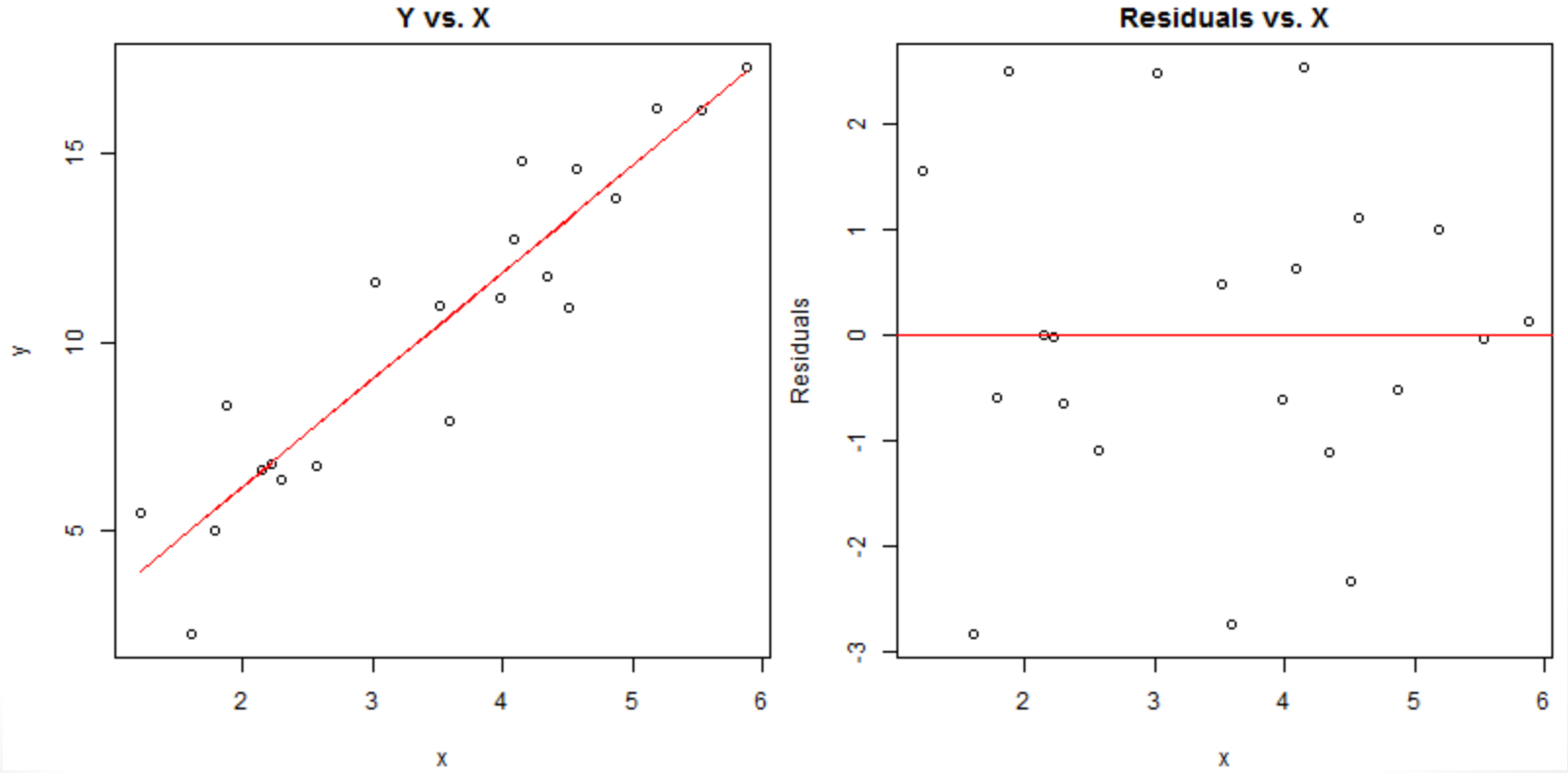
- Assume that we investigate the relationship between rat's length (response variable) and weight/temperature (predictors).
- Consider the following model:
  - Length (mm) = 5 + 2*weight (g) + 0.3*temperature (F).
  - 1 gram increase of weight adds 2mm of length when temperature is held fixed.
  - 1 degree increase of temperature adds 0.3 mm of length when weight is held fixed.

Assumption #1 - Linearity

- Linear regression assumes that the relationship between Y and X_i's is linear.
- Residuals = Observed – Fitted.
- To investigate the linearity assumption, check
  - Plot of Y vs. X_i
  - Scatter plot of residual vs. X_i for all i.
  - Scatter plot of residual vs. the fitted values

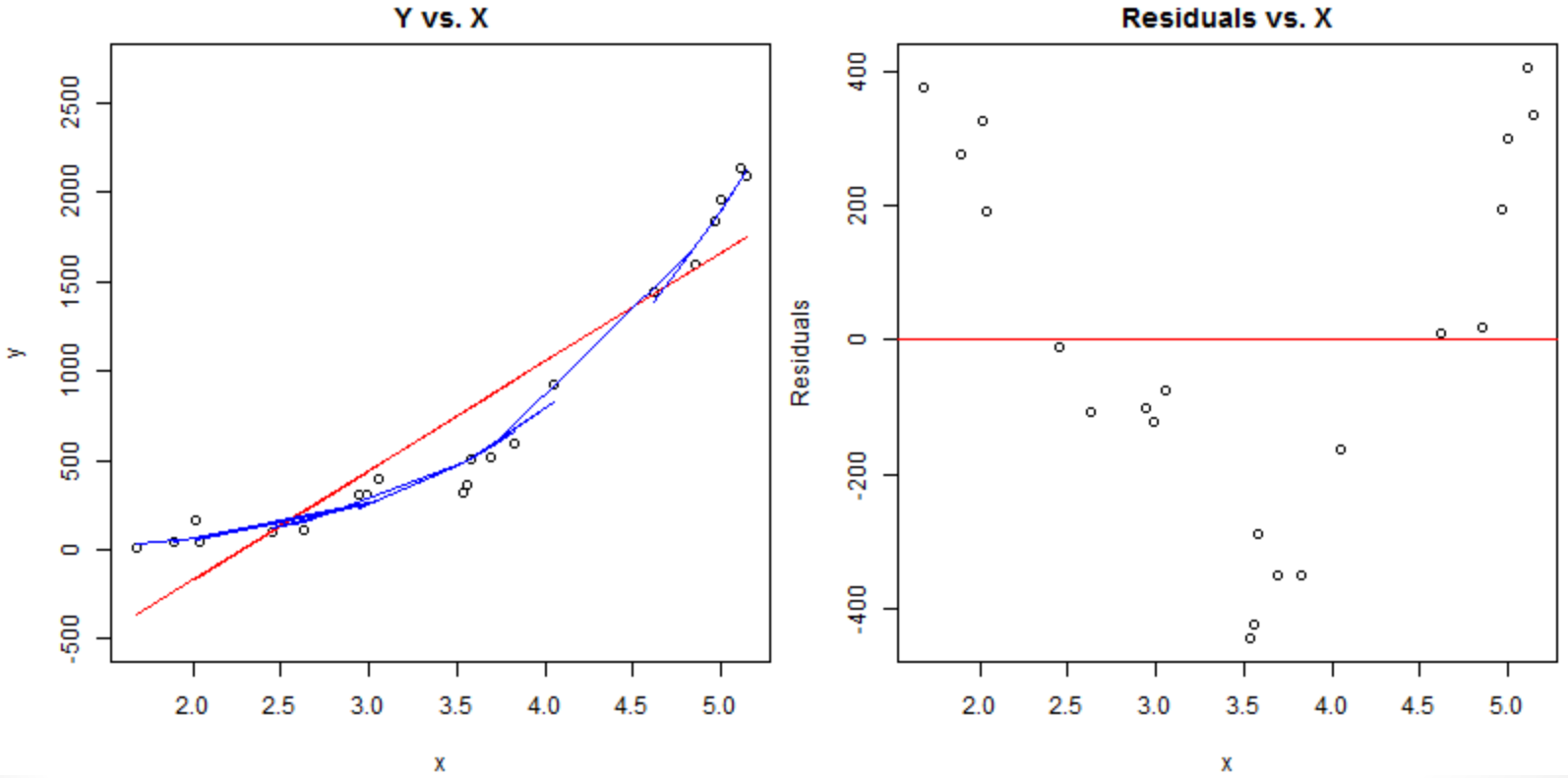
Example – Linearity

- Linearity assumption is satisfied:



Example – Linearity

- Linearity assumption is NOT satisfied:

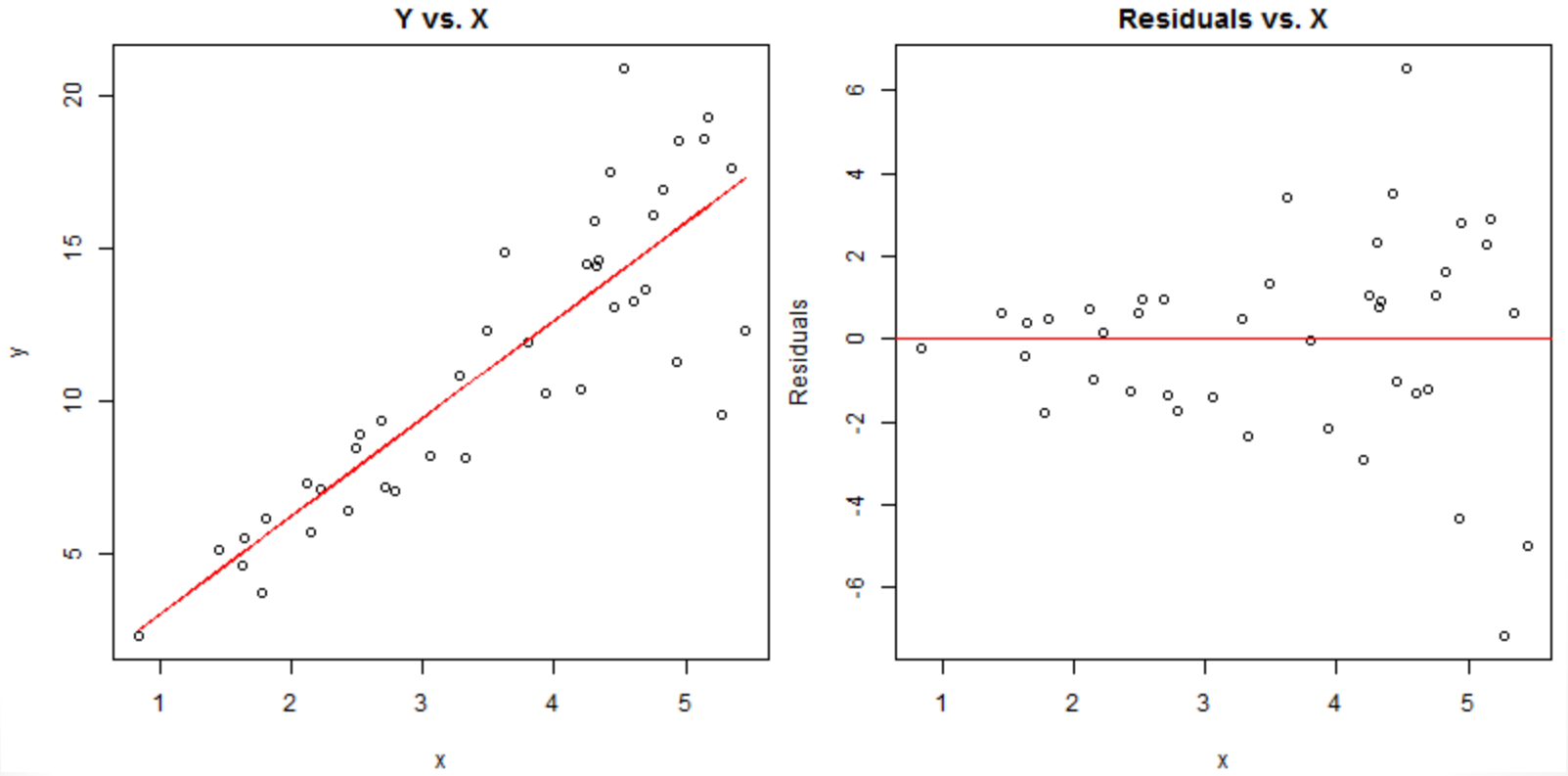


Assumption #2: IID Normal

- The errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are assumed to be independently and identically distributed (iid) normal with mean 0 and variance σ^2.
  - Check whether the errors have a constant variance.
  - Check whether the errors are normally distributed.
  - Check whether the errors are independent.

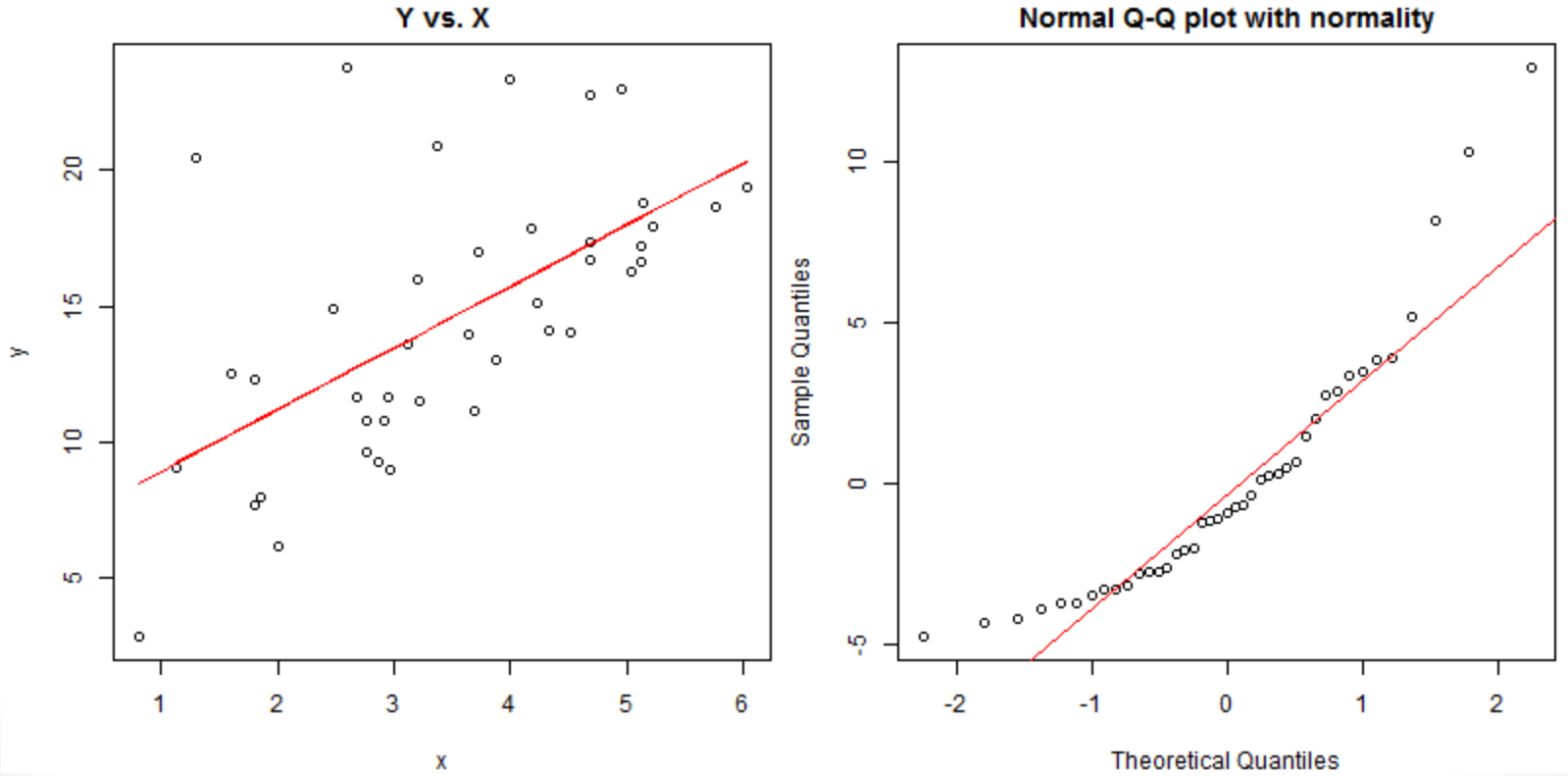
Example – Constant Variance

- The constant variance assumption is violated:



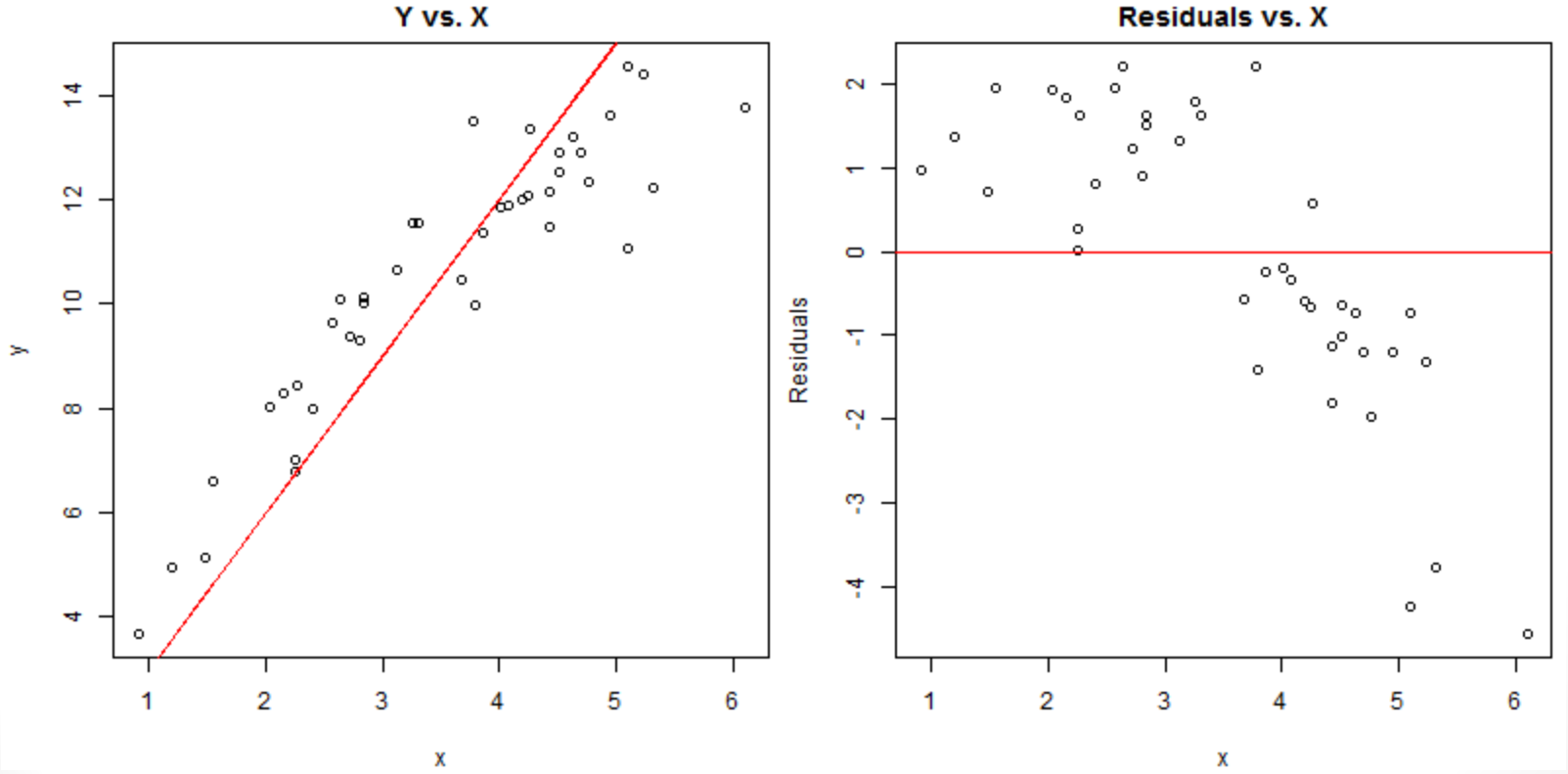
Example - Normality

- The normality assumption is violated:



Example – Independence

- The errors are not independent:

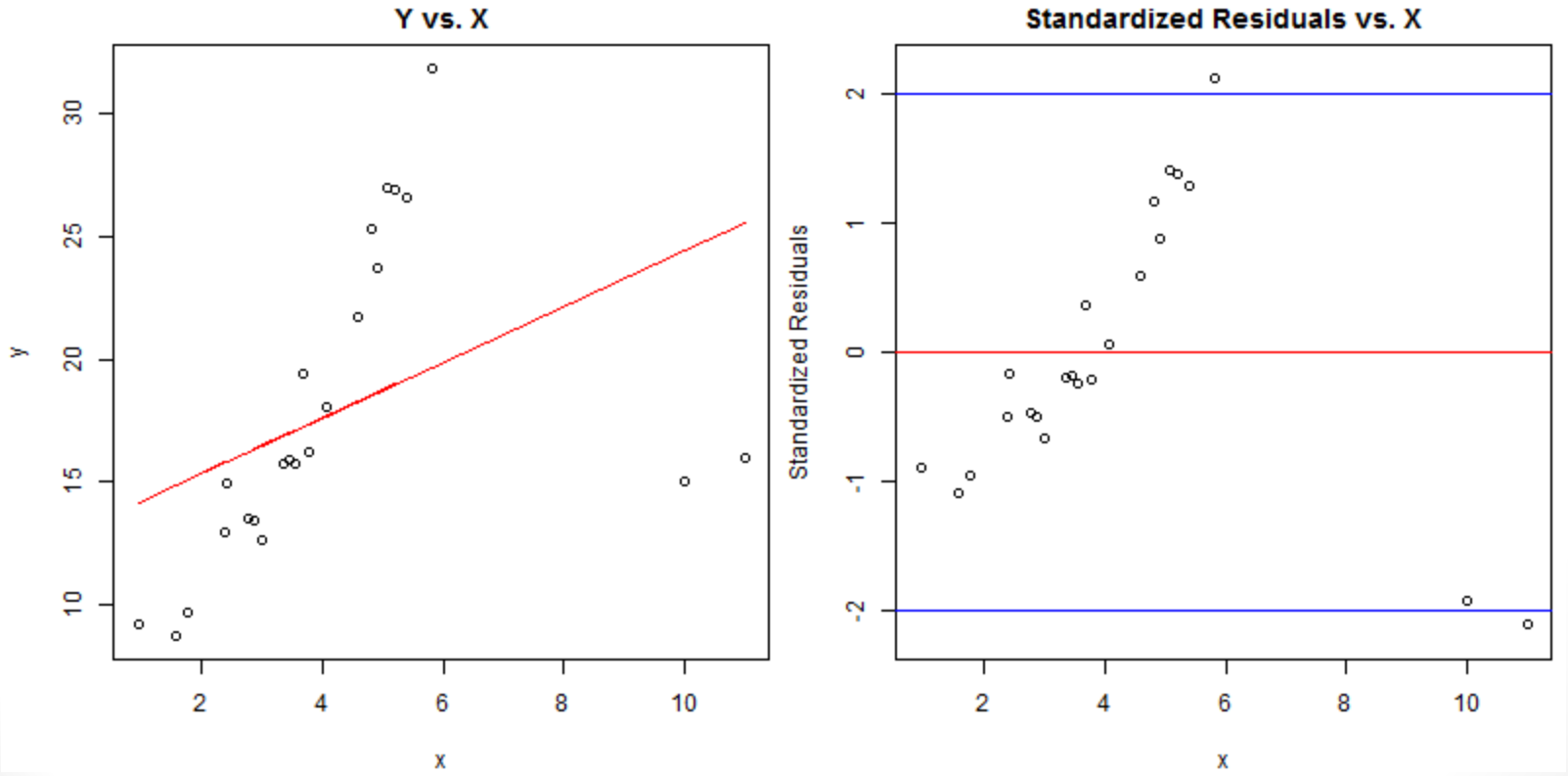


Outliers

- Outliers may significantly affect the regression results.
- To detect outliers, one may use standardized residuals, which are residuals divided by estimates of the standard error of the residuals.
- The observation with the standardized residual greater than 2 or less than -2 is a potential outlier.

Example – Outliers

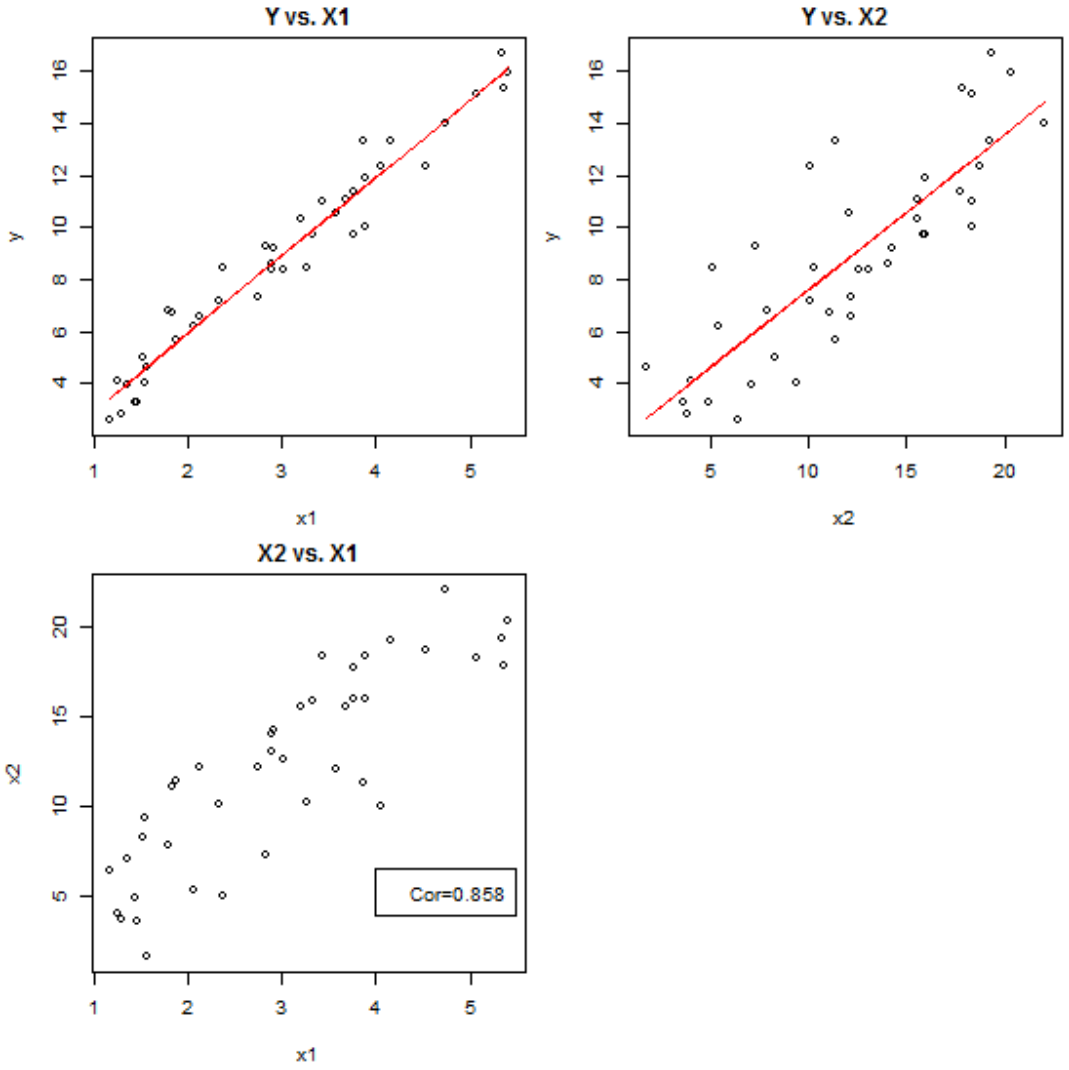
- X=10 and x=11 are outliers.



Multicollinearity

- When the predictor variables are strongly correlated, the regression results may be misleading.
- The estimated coefficients are very sensitive to the addition or deletion of correlated predictors in general.
- The regression coefficients may show large sampling errors, which results in dropping them in the model.
- In practice, multicollinearity may be present if
  - The effect of the predictors is not consistent with what you expected;
  - Predictors that were expected to be significant do now show any significance.

Example - Multicollinearity



Example – Continued

Coefficient of X1 from Y vs. X1	P-value
2.99	<0.001

Coefficient of X2 from Y vs. X2	P-value
0.594	<0.001

Coefficients from Y vs. X1 and X2	P-value
3.036 (X1)	<0.001
-0.012 (X2)	0.804

Remedies

- Data transformation is widely used to cure the violations of linearity, heterogeneity of variance, and normality assumptions. Data transformation needs to be carefully examined.
- For outliers, deletion of them or data transformation is often helpful.

Remedies- Continued

- For correlated errors, time series models are often used to model error terms.
- For collinear data, dropping one of the variables might be helpful. Principal component regression might be useful as well. Increasing sample size is always preferred.

Conclusion

- Graphical methods to check assumptions are reviewed.
- Using linear regression analysis without checking assumptions might draw incorrect results.
- If some of the assumptions are violated, remedies need to be carefully examined.

Free Drop-in Consulting

- **Medical College of Wisconsin**:
  Tuesdays and Thursdays
  Time: 1:00 PM—3:00 PM
  Building: Health Research Center
  Room: H2400 Biostatistics

- **MCW Cancer Center**
  Wednesdays 10:00 AM—12:00 PM
  Fridays 1:00 PM—3:00 PM
  Building: MCW Clinical Cancer Center
  Room: Clinical Trials Support Room
  CLCC: 3236 (Enter through C3233)

- **Froedtert Pavilion**:
  Mondays & Wednesdays
  Time: 1:00 PM—3:00 PM
  Building: Froedtert Pavilion
  Room: L772A- TRU Offices (Lower Level)

- **Clement J. Zablocki VA Medical Center:**
  1st & 3rd Monday of the month
  Time: 9:00 AM—11:00 AM
  Building: 111, 5th Floor B-wing
  Room: 5423

- **Marquette University:**
  Every Tuesday
  Time: 8:30 AM—10:30 AM
  Building: School of Nursing—Clark Hall
  Room: Office of Research and Scholarship: 112D
  Contact: Jessica Pruszynski, PhD to make an appointment
  Please note: Priority given to MU Nursing and Dental School personnel

Questions?