Model Diagnostics

Stat 230: Applied Regression Analysis

PDF version of slides

RMarkdown demo

Conditions required for inference

Our model must be valid for inference to be valid

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

Conditions to check:

- Linear relationship is appropriate
- Errors are independent and identically distributed (iid)
- Errors are normally distributed
- Variance of the errors doesn't depend on x

Residuals

Definition: $e_i = \widehat{\varepsilon}_i = y_i - \widehat{y}_i$

Properties:

- sum to zero $\implies$ mean is 0
- **uncorrelated** with x and $\widehat{y}$
- normally distributed
- $SD(e_i) = \widehat{\sigma}\sqrt{1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$

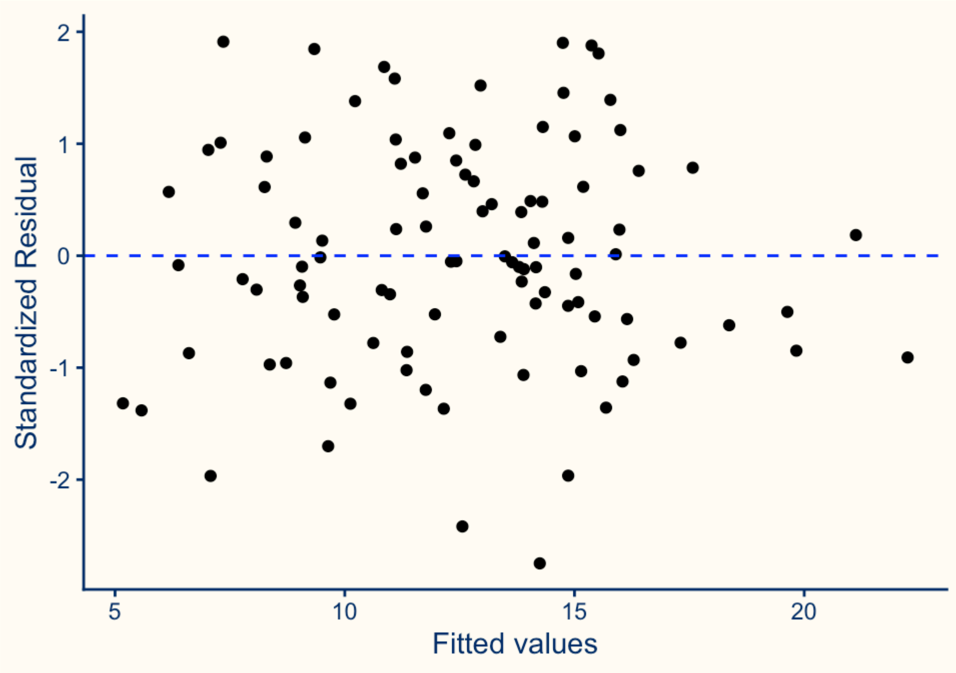
Standardized residuals

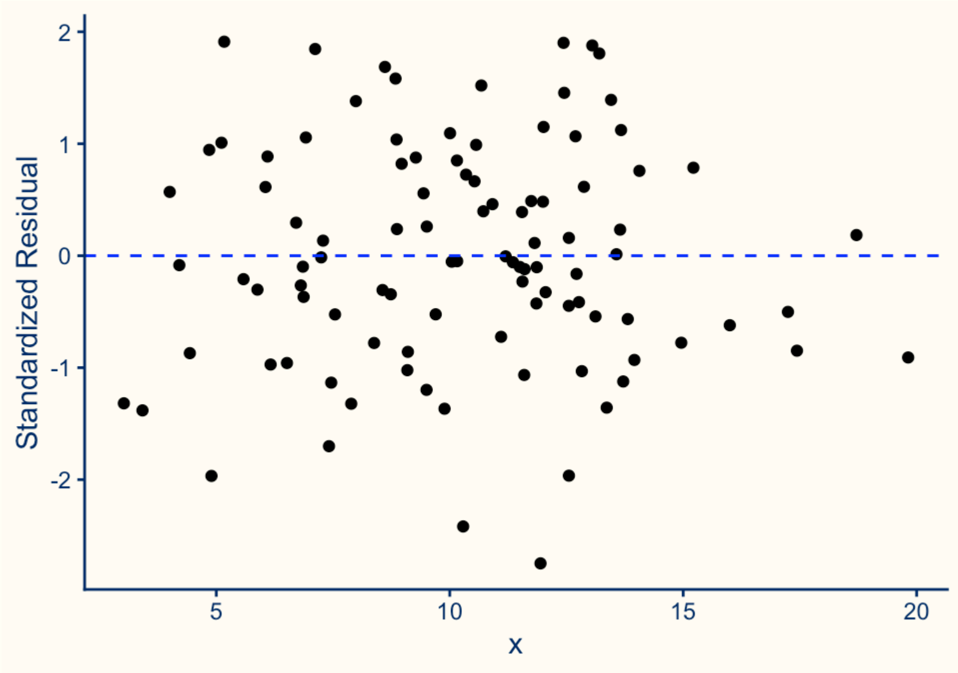
$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{\sum(x_i - \overline{x})^2}}}$$

Properties:

- sum to zero $\Longrightarrow$ mean is 0
- **uncorrelated** with x and $\hat{y}$
- normally distributed
- $SD(r_i) = 1$

A “good” residual plot





Your turn

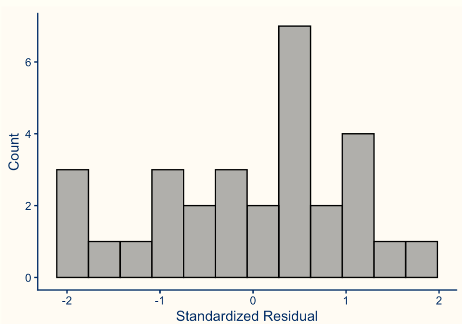
- Work in groups
- On the whiteboards, sketch a plot of y vs. x and a corresponding residual plot that would indicate a violation of the

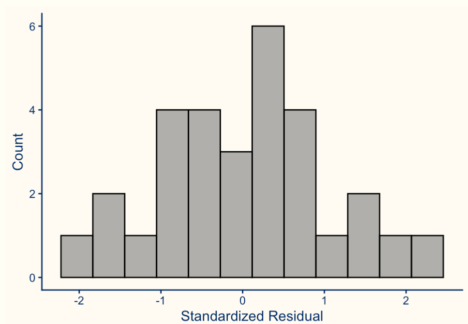
  1. linearity condition
  2. constant variance condition

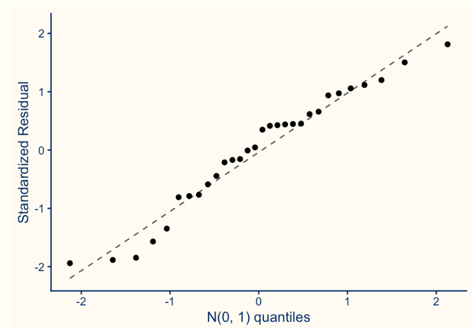
Assessing normality

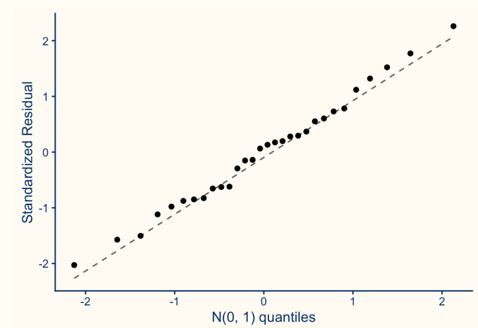
- histogram of residuals
- normal Q-Q plot of residuals

Examples of “good” plots:





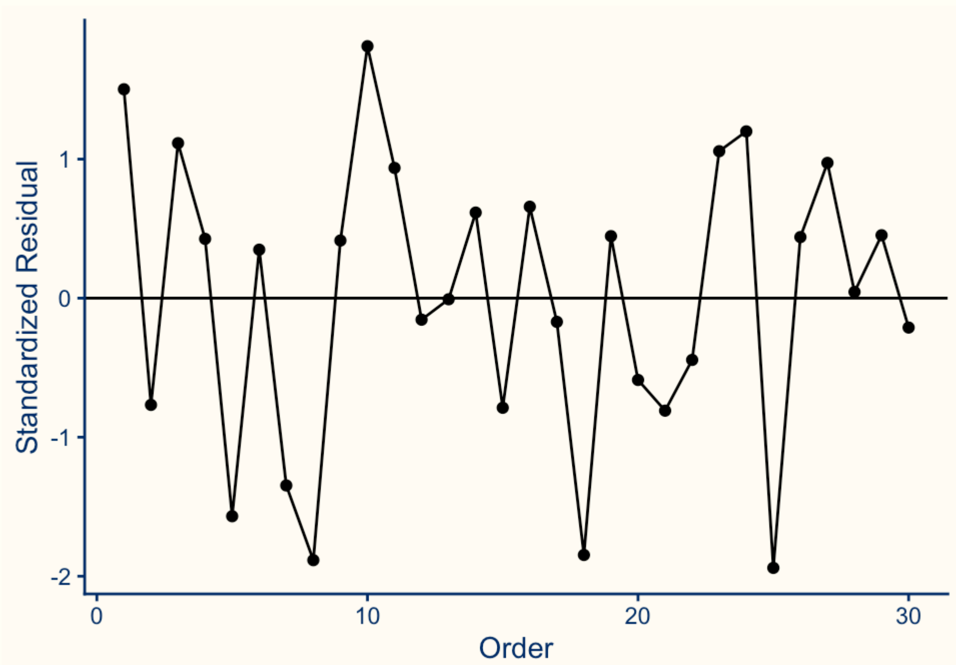


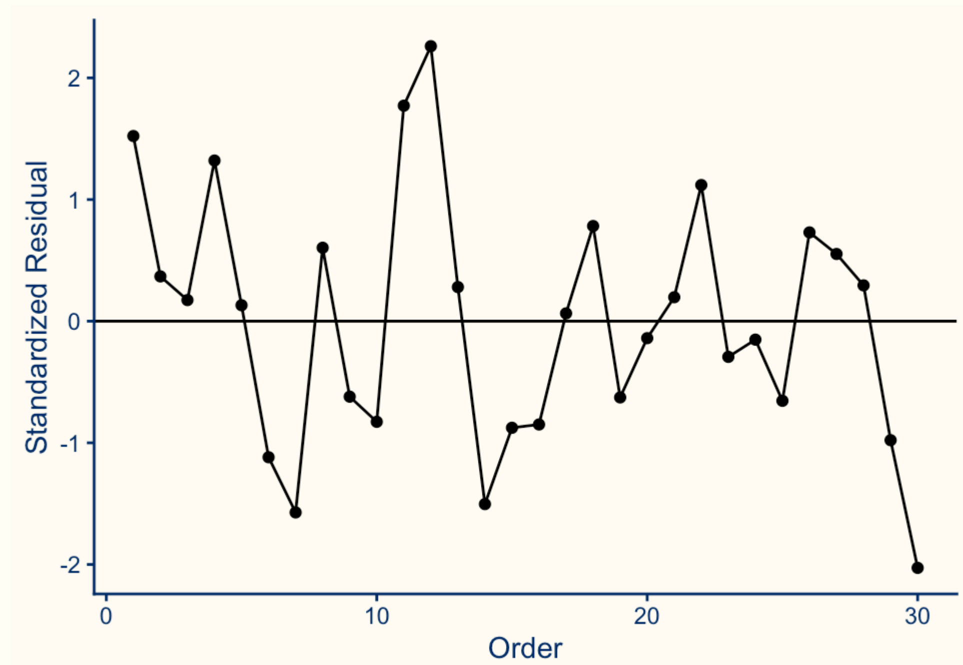


Assessing independence

- plot residuals vs. variable inducing dependence (e.g. time, location, subject ID)

Examples of "good" plots:





Your turn

Work through Example 1 on the worksheet

What happens the conditions aren't valid?

- **Linearity:** if nonlinear, everything breaks!
- **Independence:** estimates are still unbiased (i.e. we fit the right line) but measures of the accuracy of those estimates (the SEs) are typically too small
- **Normality:** estimates are still unbiased (i.e. we fit the right line), SEs are correct BUT confidence/prediction intervals are wrong (we can't use t-distribution)
- **Constant error variance:** estimates are still unbiased but standard errors are wrong (and we don't know how wrong)

What do we do if our assumptions are violated?

1. Change our assumptions (hard, need more stats)
2. Transform y, x, or both
3. Change the type of inference (remember the bootstrap?)

Transforming variables can

- Address non-linear patterns (i.e., linear on transformed scale)
- Stabilize variance
- Correct skew
- Minimize the effects of outliers

Applying transformations

To apply a transformation, we calculate a new variable and use it in place of the original variable in our model

Examples

$$
\begin{aligned}
\log(y) &= \beta_0 + \beta_1 x + \varepsilon \\
y &= \beta_0 + \beta_1 \sqrt{x} + \varepsilon \\
\log(y) &= \beta_0 + \beta_1 \sqrt{x} + \varepsilon
\end{aligned}
$$

Your turn

Work through Example 2 on the worksheet

Review of logarithms

The logarithm $\log_b(x)$ is a function that is the exponent (power) that the base, b, must be raised to produce the value x:

- $\log_{10}(100) = 2$ since $10^2 = 100$
- $\log_{10}(10) = 1$ since $10^1 = 10$
- $\log_2(1) = 0$ since $2^0 = 1$
- $\log_2(0.5) = -1$ since $2^{-1} = \frac{1}{2}$

Review of logarithms

- Takes in only positive numbers, i.e. $x > 0$
- The log of products is the sum of the logs

$$\log_b(mx) = \log_b(m) + \log_b(x)$$

- The log of quotients is the difference of the logs

$$\log_b\left(\frac{m}{x}\right) = \log_b(m) - \log_b(x)$$

- The log of powers is the exponent times the log

$$\log_b(x^p) = p\log_b(x)$$