Outline

1. Model Assumptions for SLR
2. Model Diagnostics: Graphical Techniques
3. Remedial Measures: Transformation

Model Assumptions

- A straight line relationship between the response variable Y and the explanatory variable X:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{where} \quad E(\varepsilon_i) = 0$$

- Equal variance:

$$Var(\varepsilon_i) = \sigma^2.$$

- Independence:

$$Cov(\varepsilon_i, \varepsilon_{i'}) = 0 \quad \text{for} \quad i \neq i'.$$

- Normal distribution:

$$\varepsilon_i \sim N(0, \sigma^2).$$

Robustness of Model Assumptions

Departure	$\hat{\beta}/\hat{\mu}_h$	s.e.	$\hat{Y}_{h(new)}$	s.e.
Linearity	S	S	S	S
Equal variance	R	S	R	S
Independence	R	S	R	S
Normality	R	R	R	S
Outliers	S	S	S	S

S = sensitive; R = robust.

Outline

Model Diagnostics

- Correct inference hinges on model assumptions.
- **Model diagnostics** are to evaluate the model assumptions and determine how reasonably they are met.
- A main idea for model diagnostics is to examine the residuals.
- Consider graphical approaches: Subjective but informative.

Graphical Techniques

- Exploratory data analysis (EDA).
  - exploration of X and Y.
  - May not be as effective for model diagnostics.
- Recall for $i = 1, \ldots, n$
  - the ith fitted value: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - the ith residual: $e_i = Y_i - \hat{Y}_i$

What does e_i estimate/predict:

$$\varepsilon_i = Y_i - \mathbb{E}(Y_i) \sim_{\text{i.i.d}} N(0, \sigma^2)$$

Properties of Residuals

- Sample mean: $\bar{e} = 0$.
  Why?
  $$\bar{e} = \frac{\sum_{i=1}^{n} e_i}{n} = 0.$$

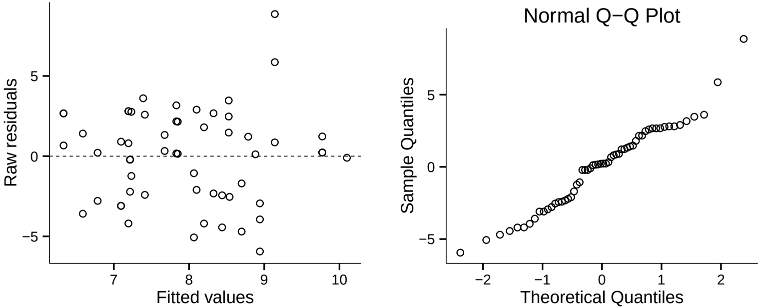
- Sample variance: $\hat{\sigma}^2$.
  Why?
  $$\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \hat{\sigma}^2.$$

- Dependence (Assignment 3(a) and 3(b))
  Why?
  $$\sum_{i=1}^{n} e_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} X_i e_i = 0.$$

Residual Plots

- Departures from model assumptions can be difficult to detect directly from X and Y.
- Thus consider residual plots.
    - Plot e_i against X_i.
    - Plot $|e_i|$ against X_i.
    - Plot e_i^2 against X_i.
    - Plot e_i against $\hat{Y}_i$.
    - Plot e_i against time.
    - Box plot of e_i.
    - Normal QQ plot of e_i.

Example: Wetland Species Richness



Types of Residuals

- **Raw residual** (or, **ordinary least squares residual**):

$$e_i = Y_i - \hat{Y}_i.$$

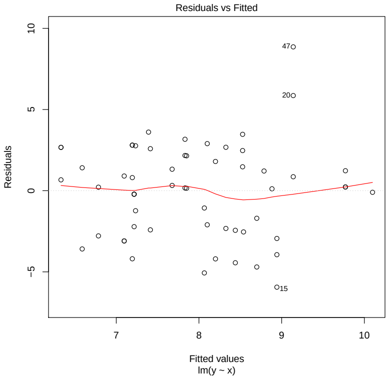
- **standardized residual**:

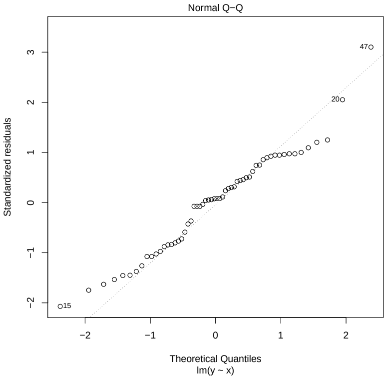
$$r_i = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}, \quad \text{where} \quad p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

where $\hat{\sigma}^2 =$ MSE based on the entire sample. Why?

$$\begin{aligned}
\text{Var}(\boldsymbol{e}) &= \text{Var}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \text{Var}(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) \\
&= \text{Var}(\boldsymbol{Y} - \boldsymbol{X}\underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}}_{\hat{\beta}}) \\
&= \text{Var}\left(\underbrace{(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)}_{\text{non-random}}\boldsymbol{Y}\right) \\
&=
\end{aligned}$$

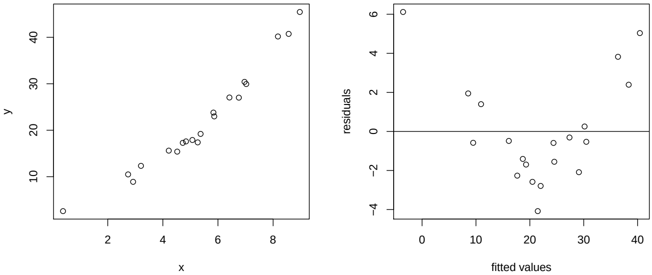
Example: Wetland Species Richness





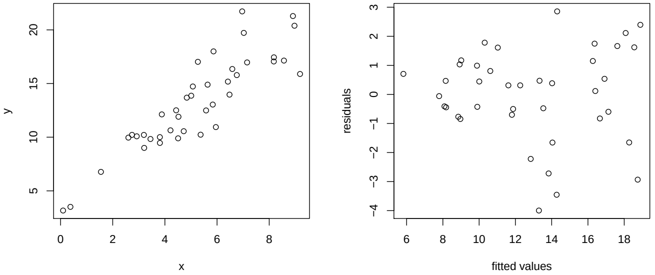
Nonlinearity of Regression Function

- Plot e_i against $\hat{Y}_i$ (or X_i).
- Random scatter indicates no serious departure from linearity.
- Example of departure from linearity: Curved relationship.



Non-equal Error Variance

- Plot e_i against $\hat{Y}_i$ (or X_i).
- Plot $|e_i|$ against $\hat{Y}_i$ (or X_i).
- Plot e_i^2 against $\hat{Y}_i$ (or X_i).
- Random scatter indicates no serious departure from constant variance.
- Example of departure from constant variance: Megaphone/funnel shape.



Nonindependence of Error Terms

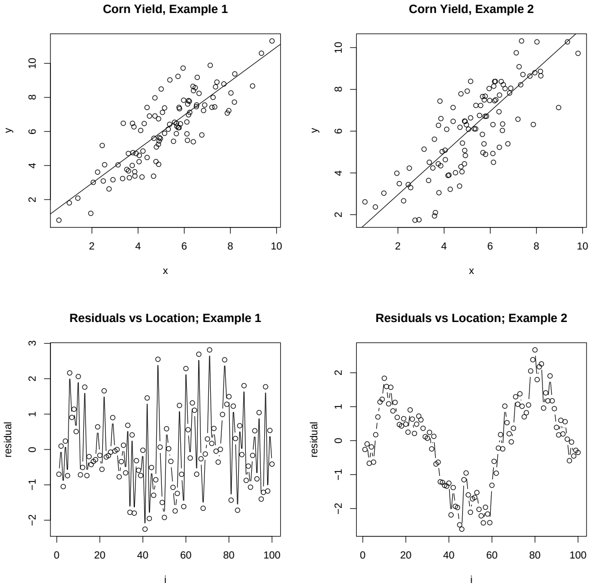
- Possible forms of nonindependence.
  - Observations collected over time and/or across space.
  - Study done on sets of siblings.
- Example of departure from independence:
  - Trend effect
  - Cyclical non-independence

Examples: Corn Yield

For $i = 1, \ldots, n$,

- $i =$ the index of the patch planted to corn.
- The patches are arranged in a long line at the edge of a field.
- $X_i =$ the amount of fertilizer applied to the ith patch.
- $Y_i =$ the corn yield in the ith patch.
- Plot e_i against location i.

Examples: Corn Yield



Nonnormality of Error Terms

Assess whether the residuals $\{e_i\}$ follow from normal.

- Box plot, histogram of e_i.
- Normal QQ plot: compared sorted residuals $e_{(1)}, \ldots, e_{(n)}$ to quantiles from standard normal $N(0, 1)$.

- If the residuals are approximately normal, the normal QQ plot should be approximately linear.
- It is a good idea to examine other departures first.
  other departure affects the distribution
  e.g., distribution of $\{e_i\}$ is subject to independence assumption especially in small sample size

Presence of Outliers

- An outlier refers to an extreme observation.
- Box plot, histogram plot of $\{e_i\}$.
- Plot e_i against $\hat{Y}_i$ (or X_i).
- Random scatter indicates absence of outliers.
- Outliers may convey important information.
  An error. A different mechanism is at work. A significant discovery.

Graphical Techniques: Remarks

- We generally do not plot residuals (e_i) against response (Y_i). Why not?
- Residual plots may provide evidence against model assumptions, but do not generally validate assumptions.
- For data analysis in practice:
  - Fit model and check model assumptions (an iterative process).
  - Generally do not include residual plots in a report, but include a sentence or two to explain model diagnostics employed and findings obtained. such as "Standard model diagnostics did not indicate any violations of the assumptions for this model."
- For this class, include residual plots in homework assignments and reports.
- As much art as science.
  No golden rules. No magic formulas. Decision may be difficult for small sample size.

Outline

Remedial Measures

Basic approaches: replace with a more complex model or transform so SLR model is appropriate.

- Nonlinearity of regression function:
  - Transformation.
  - Polynomial regression.
  - Nonlinear regression.
- Nonequal error variance:
  - Transformation.
  - Weighted least squares.
- Nonindependence of error terms:
  - Models with correlated error terms.
- Nonnormality of error terms.
  - Transformation.
  - Nonparametric methods.
  - Generalized linear models.
- Presence of outliers:
  - Removal of outliers (with caution).
  - Robust estimation.

Example: Surviving Bacteria

Data consist of number of surviving bacteria after exposure to X-rays for different periods of time.

- Let t denote time (in number of 6-minute intervals)
- let n denote number of surviving bacteria (in 100s) after exposure to X-rays for t time.

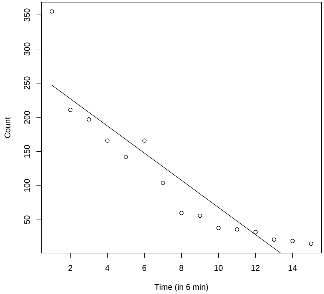
t	1	2	3	4	5	6	7	8
n	355	211	197	166	142	166	104	60

t	9	10	11	12	13	14	15
n	56	38	36	32	21	19	15

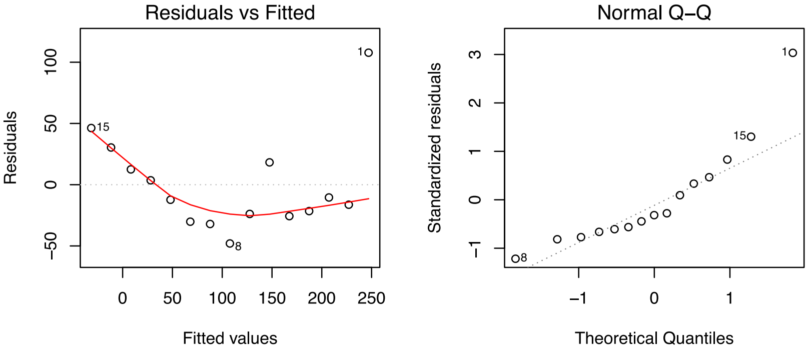
Example: Surviving Bacteria

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	267.010	22.170	12.044	2.0e-08 ***
t	-19.893	2.438	-8.158	1.8e-06 ***

Residual standard error: 40.8 on 13 degrees of freedom
Multiple R-squared: 0.8366, Adjusted R-squared: 0.824
F-statistic: 66.56 on 1 and 13 DF, p-value: 1.804e-06



Example: Surviving Bacteria



Example: Surviving Bacteria

- Here there is a theoretical model:

$$n_t = n_0 e^{\beta t},$$

where

  - t is time,
  - n_t is the number of bacteria at time t,
  - n_0 is the number of bacteria at the start ($t = 0$), and
  - β is a decay rate with $\beta < 0$.

- Consider a log transformation:

$$\ln(n_t) = \ln(n_0) + \beta t = \alpha + \beta t,$$

by setting $\alpha = \ln(n_0)$.
That is, we log-transformed n_t and the result is a linear model.

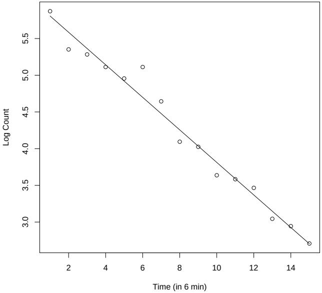
Example: Surviving Bacteria

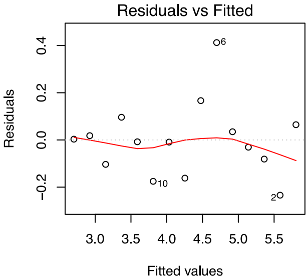
The transformed data are as follows.

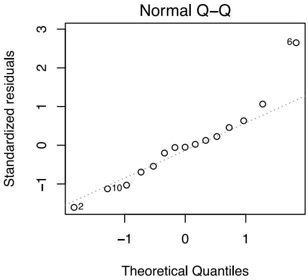
t	1	2	3	4	5	6	7	8
$\ln(n)$	5.87	5.35	5.28	5.11	4.96	5.11	4.64	4.09

t	9	10	11	12	13	14	15
$\ln(n)$	4.03	3.64	3.58	3.47	3.04	2.94	2.71

Example: Surviving Bacteria







Example: Surviving Bacteria

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.028695	0.088259	68.31	< 2e-16 ***
t	-0.221629	0.009707	-22.83	7.1e-12 ***

Residual standard error: 0.1624 on 13 degrees of freedom
Multiple R-squared: 0.9757, Adjusted R-squared: 0.9738
F-statistic: 521.3 on 1 and 13 DF, p-value: 7.103e-12

How to interpret β ?
How to interpret α ?
Inference for n_0 is not straightforward.

$$\hat{n}_0 = e^{\hat{\alpha}} = 415.30$$

but $E(\hat{n}_0) \neq n_0$.