Simple Linear Regression

Model Assessment(Continued), Prediction Intervals and Sums of Squares

Ravleen Bajaj

SLR Model Assumptions to Check

Key assumptions about the errors:

Last Time:

- Independence.
  Violations occur when:
  - Repeated measurements on the same individual.
  - Data are ordered in time (time series) or space (spatial data).
  - Clustered or family data.
- Mean 0.
- Common SD (homoscedasticity).

SLR Model Assumptions to Check

Key assumptions about the errors:

Last Time:

- Independence.
  Violations occur when:
  - Repeated measurements on the same individual.
  - Data are ordered in time (time series) or space (spatial data).
  - Clustered or family data.
- Mean 0.
- Common SD (homoscedasticity).

Today:

- Normality.

IMPORTANT: You should not assess normality if the first three assumptions are not reasonable!

Regression Error Assumptions

- For the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

and

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \tilde{\varepsilon}_i$$

For a correct model, if ε_i have a MVN then even $\tilde{\varepsilon}_i$ will have a MVN and the $\tilde{\varepsilon}_i$'s will be:

  - normally distributed
  - approx. uncorrelated
  - mean zero
  - approx. same variance

- We'll use **Q–Q plots** to visualize whether the residuals(and hence errors) follow a normal distribution.

Q–Q Plot Concept

- A **Quantile–Quantile plot** compares ordered quantiles of residuals to theoretical quantiles of a Normal(0,1) distribution.
- If residuals are normal:

  Points cluster around a straight line

- Deviations indicate:
  - Curvature $\Rightarrow$ Skewness
  - Tail deviations $\Rightarrow$ Heavy or light tails $\Rightarrow$ not normal dist.

Example

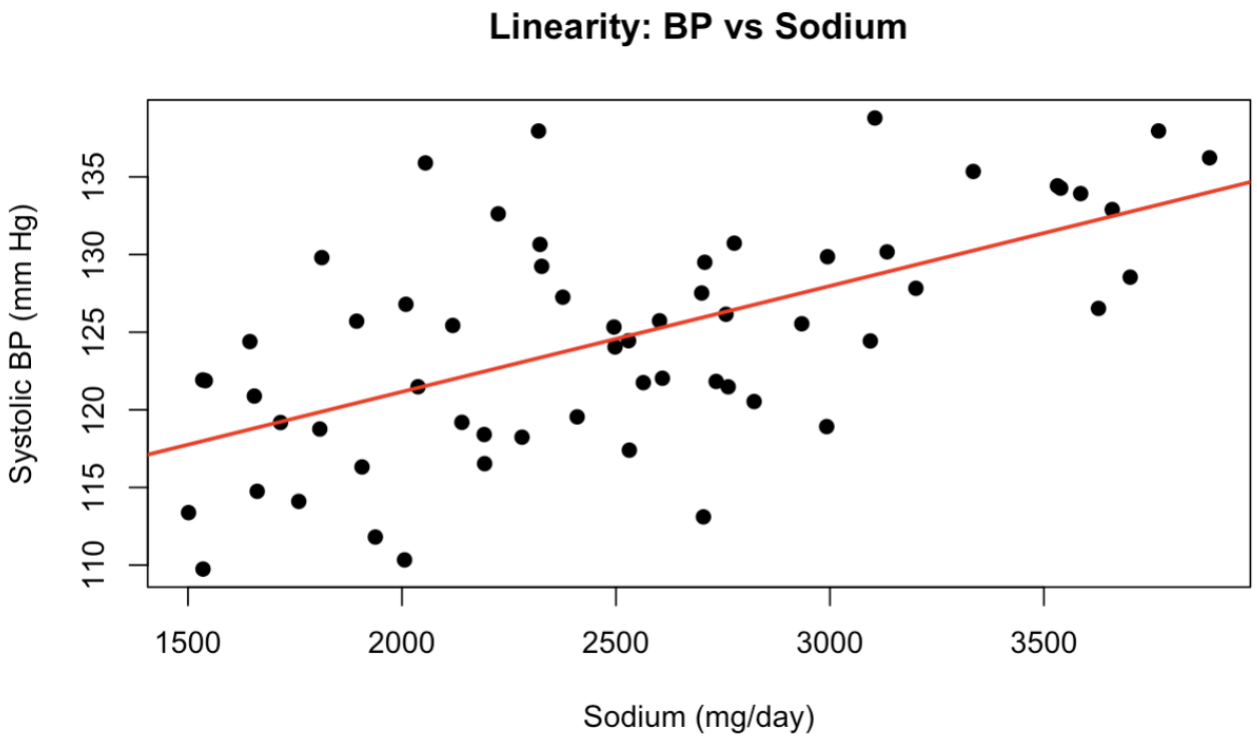
- We are interested to see the dependence of systolic blood pressure (mm Hg) among adults on their daily sodium intake (mg)
- Number of Participants: 60
- For each individual i:

$$Y_i = \text{systolic BP (mm Hg)}, \qquad x_i = \text{daily sodium intake (mg)}$$

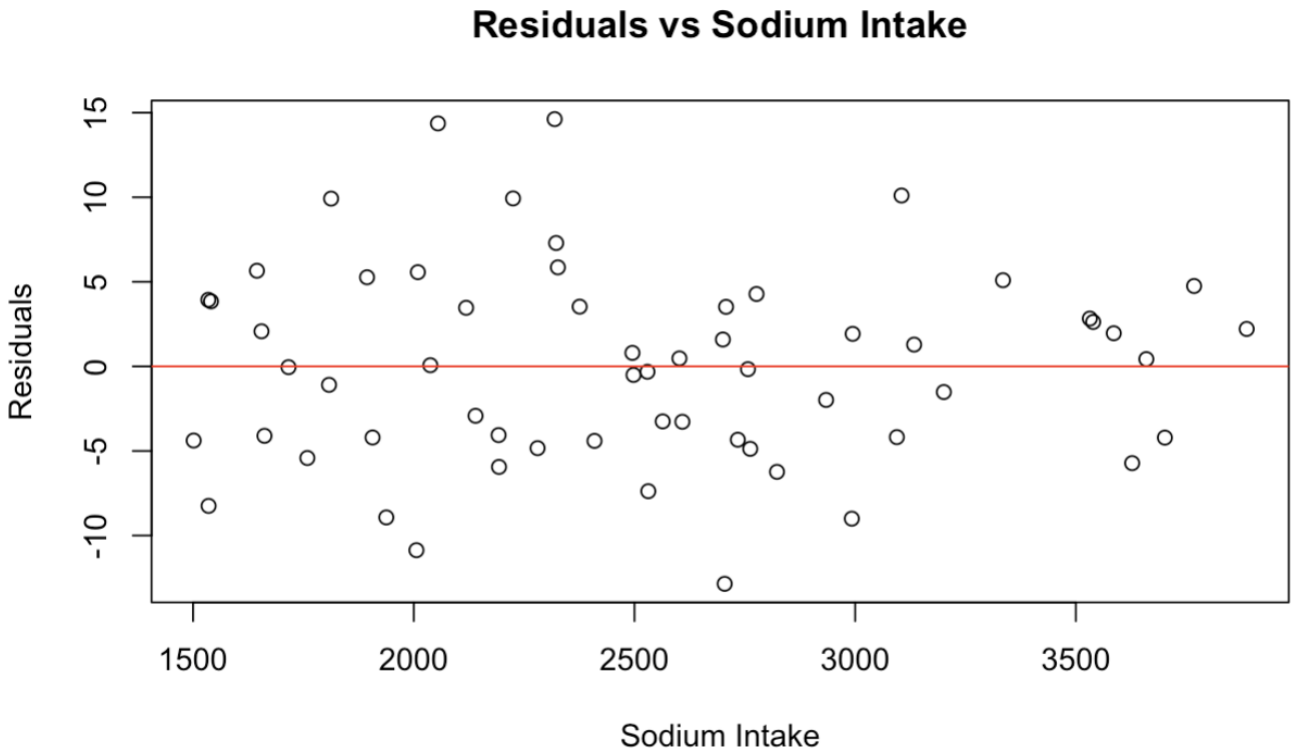
- Model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and } \varepsilon_i's \text{ are independent.}$$

Systolic BP vs Sodium Intake



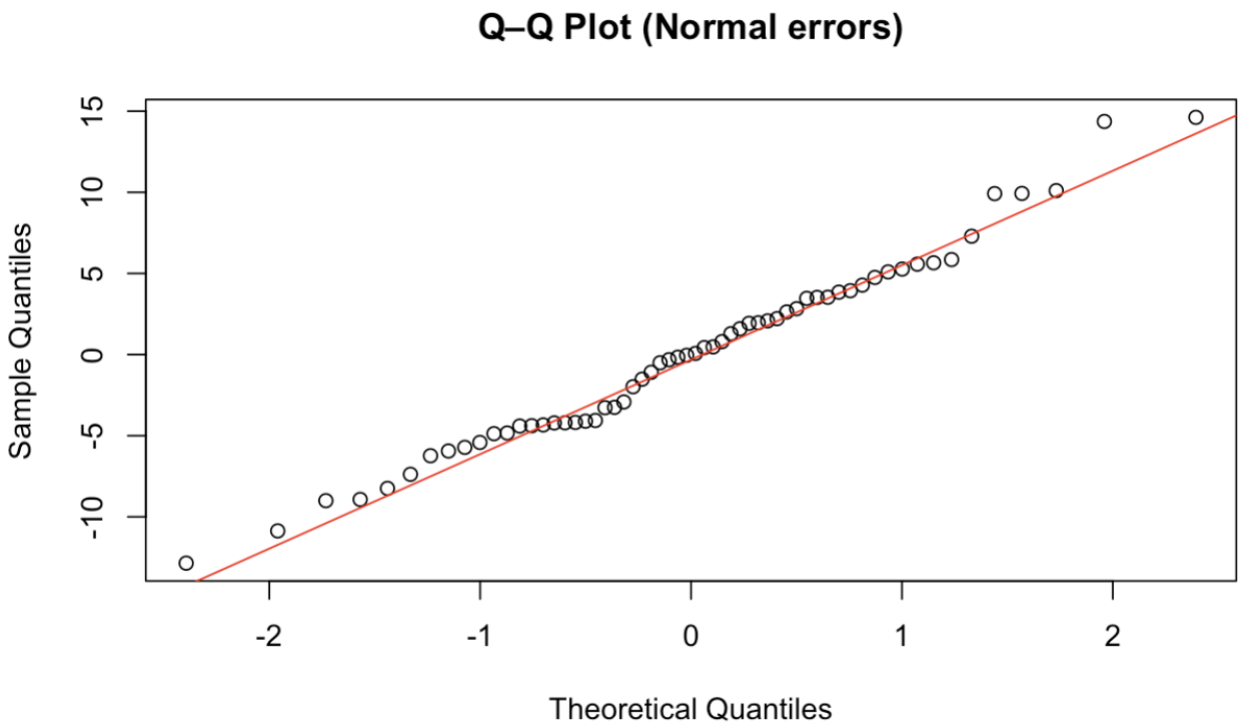
Residuals vs Sodium Intake



- No trend $\Rightarrow$ Linearity reasonable.
- Vertical spread of the residuals is approximately constant $\Rightarrow$ Common SD reasonable.

Case 1: i.i.d. Normal Errors

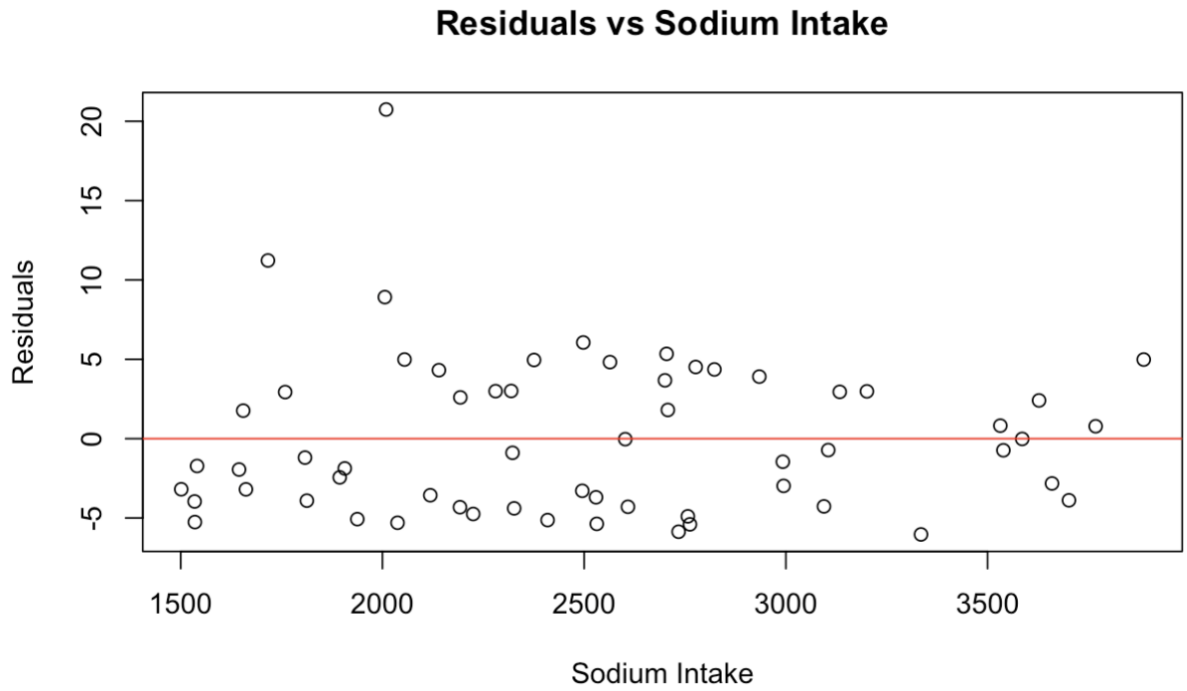
$$\varepsilon_i \sim N(0, 1)$$



- Points roughly cluster around a straight line $\Rightarrow$ residuals are approximately normal.

Case 2: Exponential Errors (Right-Skewed)

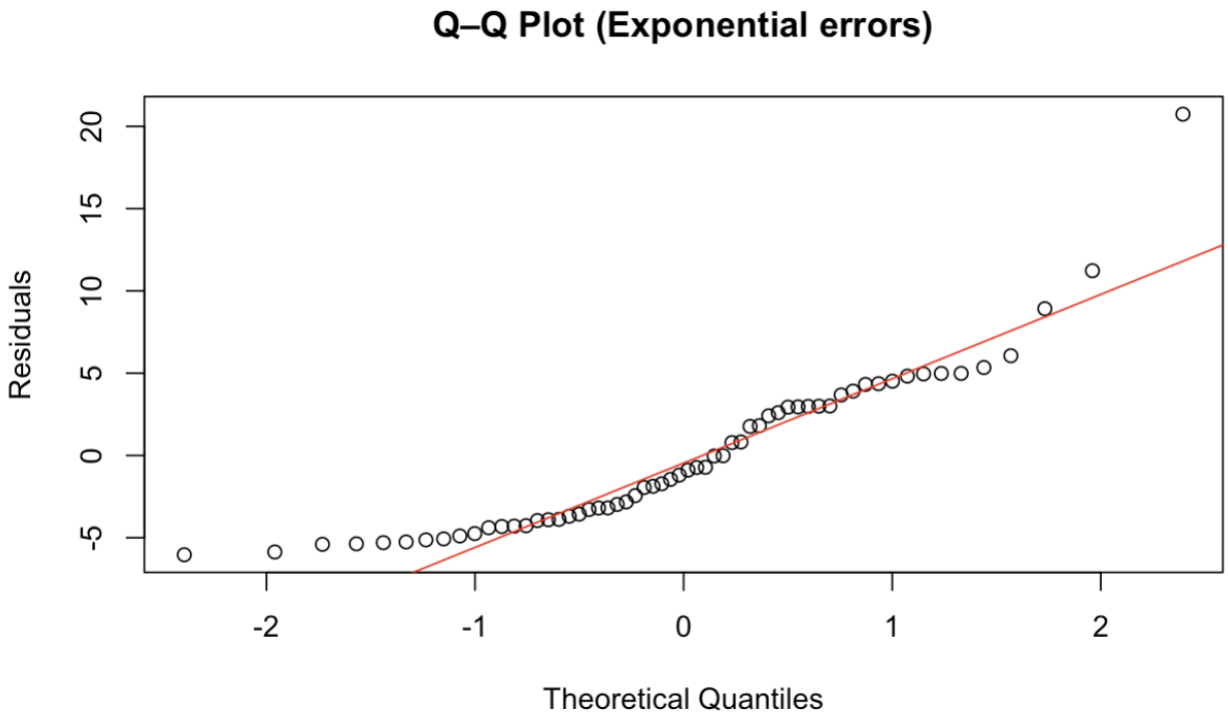
$$\varepsilon_i \sim \text{Exp}(1/6)$$



- Spread of the points above the line is greater than the spread of the points below the line $\Rightarrow$ Right skewed.

Case 2: Exponential Errors (Right-Skewed)

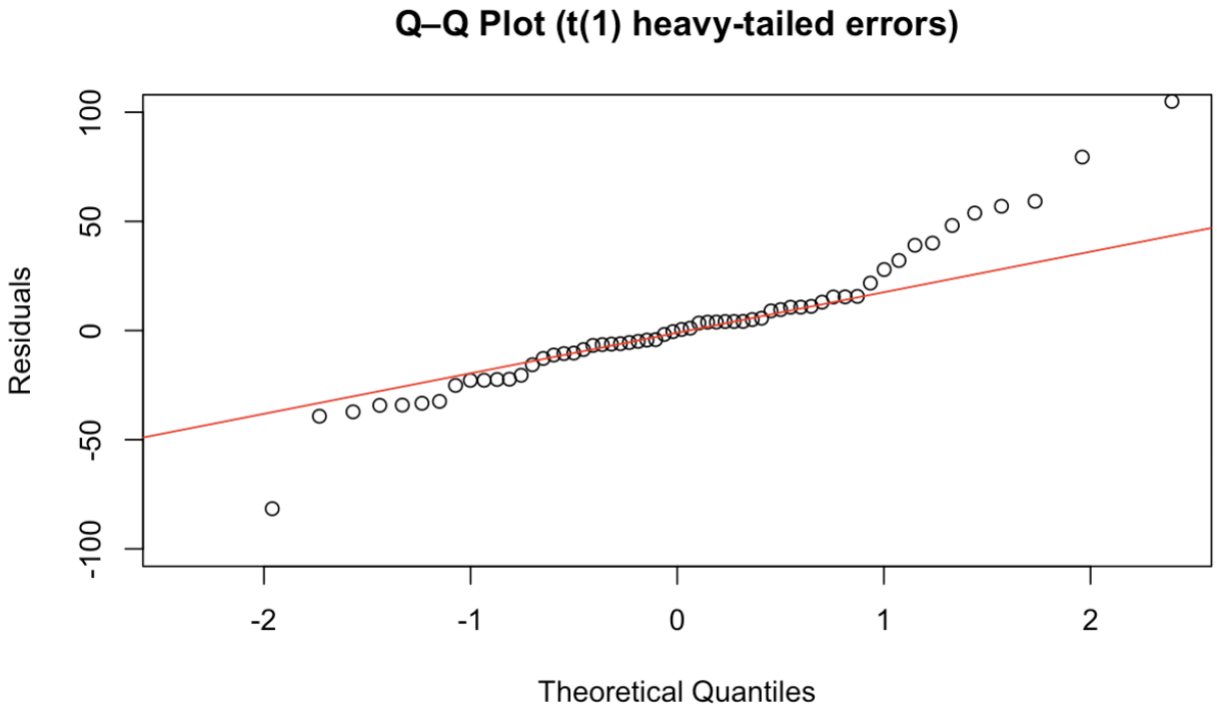
$$\varepsilon_i \sim \text{Exp}(1/6)$$



- Upward curvature at left and right tails $\Rightarrow$ **right-skewed** distribution.
- The "bowl" shape suggests a violation of the normality assumption.

Case 3: i.i.d. t(1) Errors (Heavy-Tailed)

$$\varepsilon_i \sim 10t(1)$$



- Central points near the line, but tails deviate sharply.
- Indicates heavier tails than those of the normal distribution.

Comparison Summary

Distribution	**Shape**	**Q–Q Pattern**	**Interpretation**
Normal(0,1)	Symmetric	$\approx$ straight line	Normality is reasonable
Exp(1)	Right-skewed	Bowl-like curvature	Skewed errors
$t(1)$	Symmetric heavy tails	S-shape	Heavy-tailed errors

Takeaways

- Q–Q plots are a powerful **diagnostic for normality**.
- "Bowl" shaped $\Rightarrow$ Skewness.
- S-shaped $\Rightarrow$ Heavy or light tails.

Motivation: Prediction vs Confidence Intervals

- In SLR, we often want to predict Y for a new value of x.
- Two intervals:
  - **Confidence interval (CI):** uncertainty about the mean response $\hat{\mu}_{Y|x}$.
  - **Prediction interval (PI):** uncertainty about a *new individual's response*.
- Prediction intervals are always **wider** than confidence intervals for a given value of x because they include both uncertainty about the mean ($\mu_{Y|x}$) and uncertainty about the extent of variation of the response around the mean(sigma).

Model Setup

- Recall our simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad \varepsilon_i' s \text{ are independent.}$$

- Suppose we want to predict the blood pressure for a new person with sodium intake x_0.
- Predicted response:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Formula: Prediction Interval

A $100(1-\alpha)\%$ confidence interval for the mean response at x_0:

$$\hat{\mu}_{Y|x_0} \pm t_{n-2,1-\alpha/2}\, \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Formula: Prediction Interval

A $100(1-\alpha)\%$ confidence interval for the mean response at x_0:

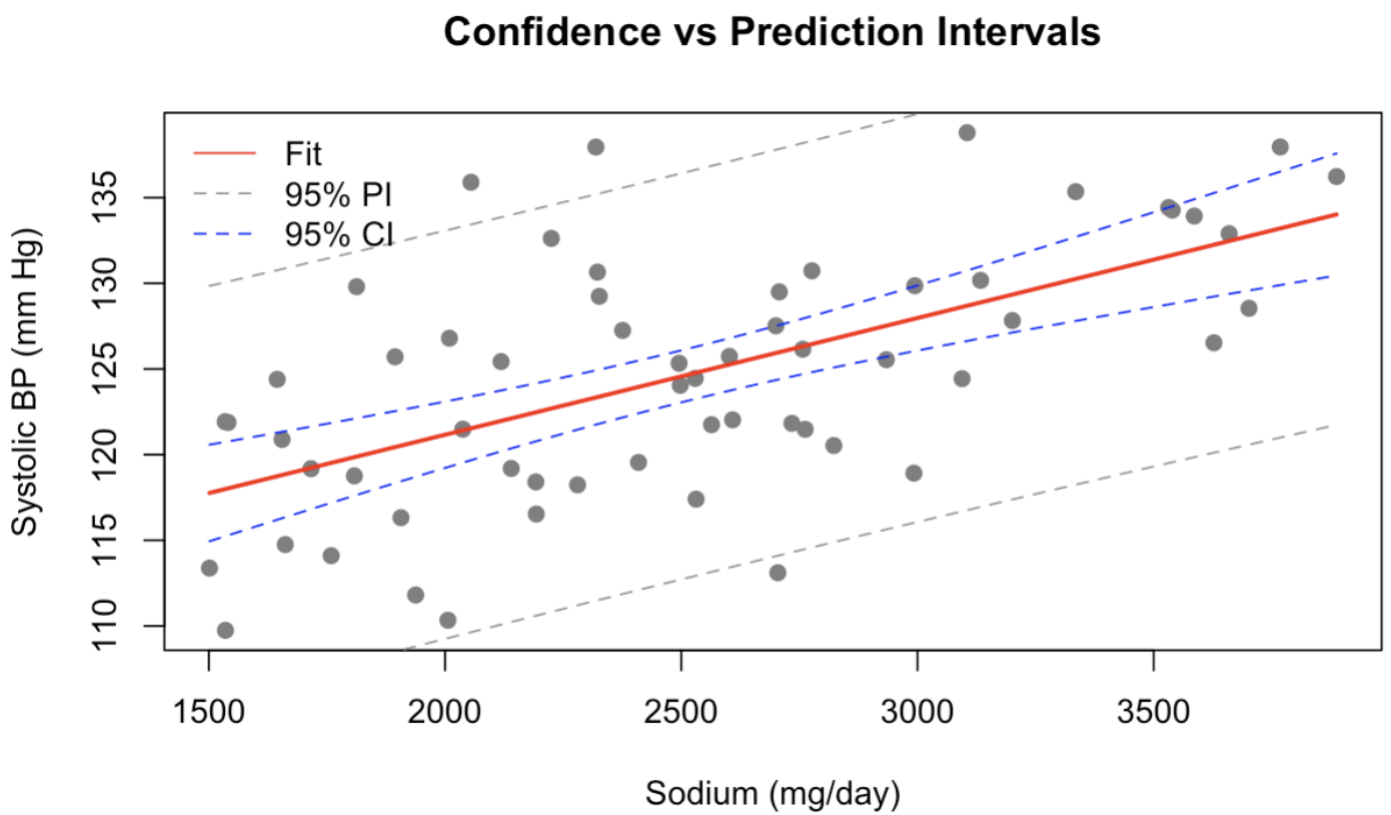
$$\hat{\mu}_{Y|x_0} \pm t_{n-2,1-\alpha/2}\,\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

A $100(1-\alpha)\%$ prediction interval for a new observation Y at x_0

$$\hat{y} \pm t_{n-2,\,1-\alpha/2}\,\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

- Mathematically, the extra "+1" term under the square root makes the PI wider than the CI.

Visualization: Pointwise Confidence vs Prediction Intervals



- The red line is the fitted regression line.
- The blue band shows the 95% CI for the mean response for different values of sodium.
- The gray band shows the 95% PI for BP for different values of sodium.

Key Takeaways

- Prediction intervals give the range where a **new individual response** is likely to fall given the individual's predictor variable value.
- Always wider than confidence intervals for the mean response for that predictor variable value.

Simple Linear Regression Recap

- Model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad \varepsilon_i' s \text{ are independent}$$

- Fitted values:

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Residuals:

$$\hat{\varepsilon}_i = y_i - \hat{\mu}_i$$

Sums of Squares Decomposition

Total Variation in Y

$$\text{Total SS: SST} = \sum (y_i - \bar{y})^2$$

$$\text{Model SS: SSM} = \sum (\hat{\mu}_i - \bar{y})^2$$

$$\text{Residual SS: SSR} = \sum (y_i - \hat{\mu}_i)^2$$

$$\text{SST} = \text{SSM} + \text{SSR}$$

R^2: Coefficient of Determination

Definition

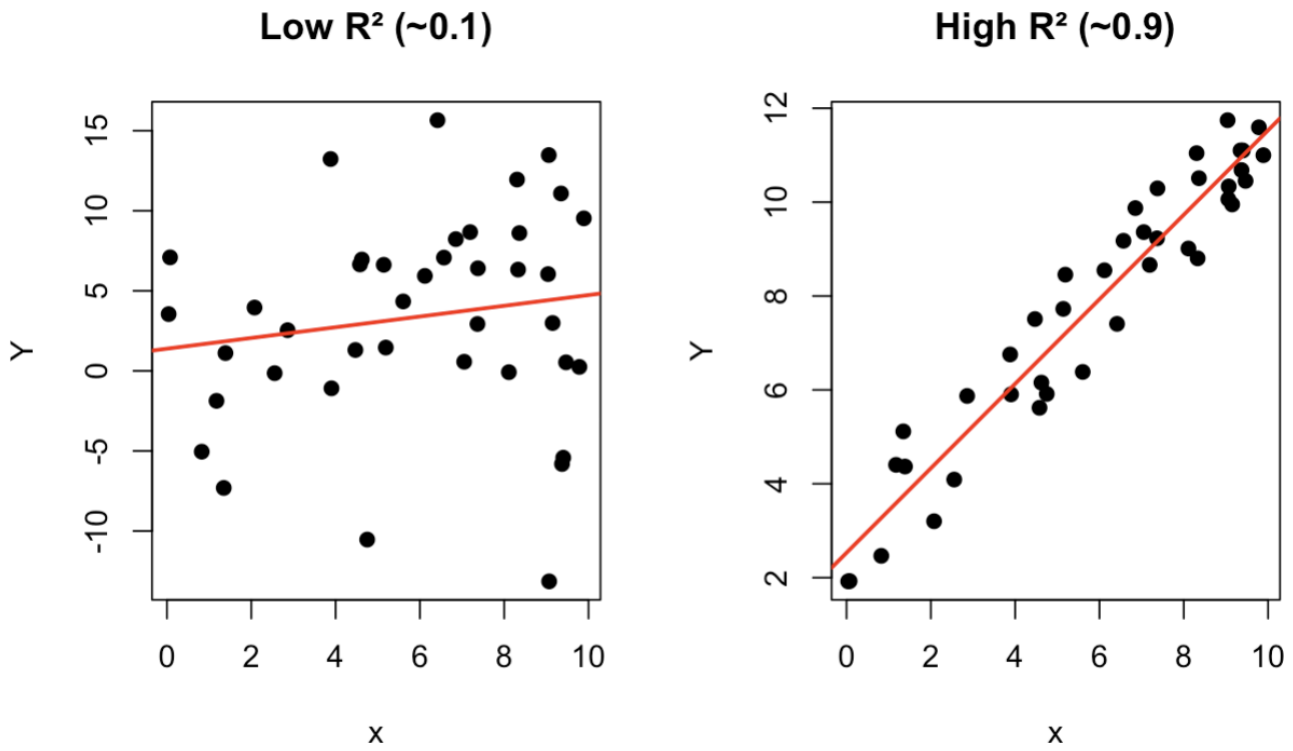
$$R^2 = \frac{\text{SST - SSR}}{\text{SST}} = \frac{\text{SSM}}{\text{SST}}$$

- Proportion of observed variability in the response that is explained by the deterministic part of the model.
- Ranges from 0 to 1.
- In simple linear regression:

$$R^2 = r^2$$

where r is the sample correlation between x and Y.

Visualizing R^2



- Weak linear relationship, large residuals.
- Strong linear pattern, small residuals.

Interpreting Key Quantities

Quantity	**Symbol**	**Interpretation**
Estimated Slope	$\hat{\beta}_1$	Change in estimated mean Y per unit x
Sample Correlation	r	Strength/direction of linear relationship
Fit quality	$R^2 = r^2$	Proportion of observed variation explained

Interpreting Key Quantities

Quantity	**Symbol**	**Interpretation**
Estimated Slope	$\hat{\beta}_1$	Change in estimated mean Y per unit x
Sample Correlation	r	Strength/direction of linear relationship
Fit quality	$R^2 = r^2$	Proportion of observed variation explained

- High $|\widehat{\beta}_1|$ and high $|r| \Rightarrow$ greater magnitude of slope and higher R^2.
- Larger $\hat{\sigma} \Rightarrow$ more scatter, lower R^2.

Sketching Practice for You

- Sketch plots of Y vs. x showing:
  1. Low R^2, high $\hat{\sigma}$
  2. High R^2, low $\hat{\sigma}$
  3. Positive sample correlation ($r > 0$)
  4. Negative sample correlation ($r < 0$)
- Discuss how slope $\widehat{\beta}_1$, r, R^2, and residual spread are visually related.

Thank You!

Questions?