Simple Linear Regression

Gioele La Manno

École Polytechnique Fédérale de Lausanne (EPFL)

School of Life Science (SV)

March 2025

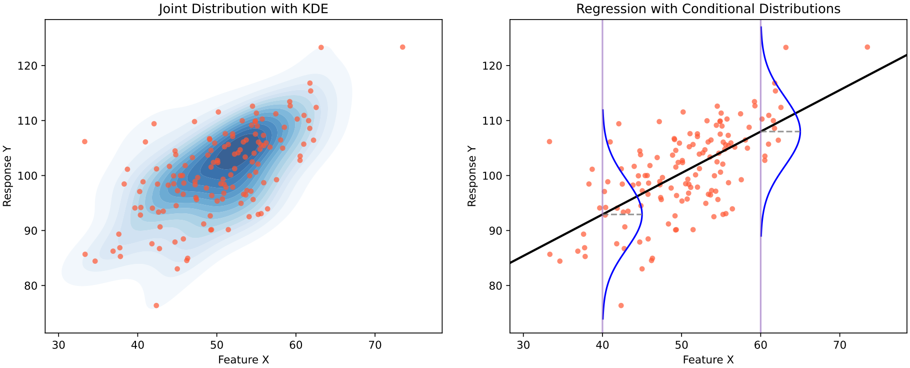
EPFL - BMI - UPLAMANNO

Contents

Simple 1D Linear Regression Analysis

From Joint Distributions to Regression Models

Linear regression can be viewed from a probabilistic perspective, where we model the conditional distribution of one variable given another, typically assuming a Gaussian error distribution.



The Linear Model: A Probabilistic Perspective

The core insight of linear regression is modeling the conditional distribution of Y given X as a normal distribution with linearly changing mean:

$$Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$

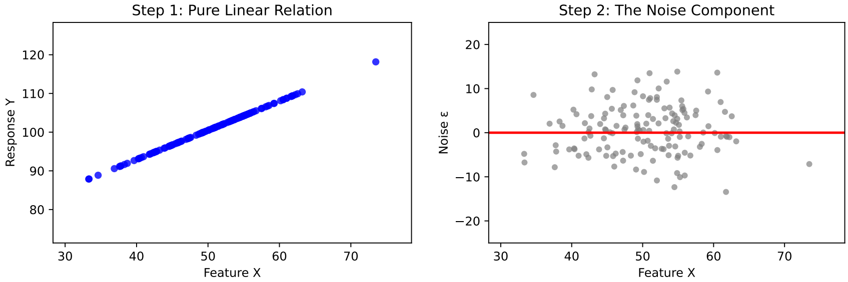
This tells us:

- For any fixed value of X=x, Y follows a normal distribution
- The mean of this distribution is a linear function: $\beta_0 + \beta_1 x$
- The variance remains constant: σ^2

The Data Generation Process in Linear Regression

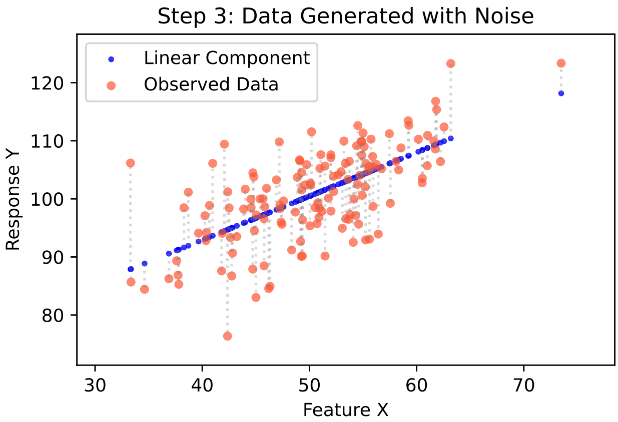
Linear regression assumes data is generated from a deterministic component plus random noise:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$



Real Data: Deterministic Trend + Random Noise

When we observe real data, we see the combination of the deterministic trend and random noise:



Interpretation Through Conditional Expectations

The regression model can be understood through conditional expectations:

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

The conditional expectation of a random variable Y given another random variable X is defined as:

$$E[Y|X] = \int y f_{Y|X}(y|x) dy$$

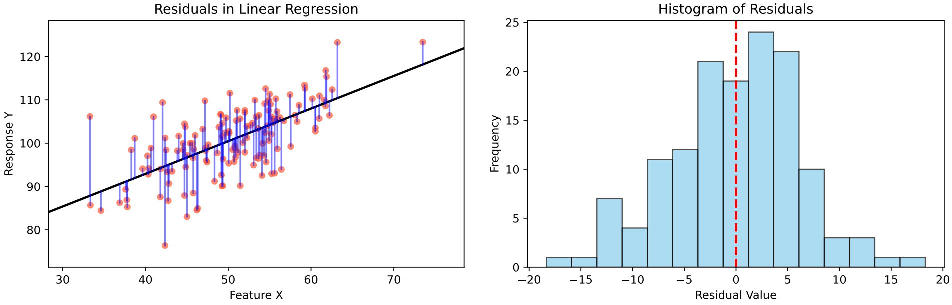
The parameters have clear interpretations:

- β_0 (intercept): The expected value of Y when X=0
- β_1 (slope): The change in the expected value of Y for a one-unit increase in X

Residuals in Linear Regression

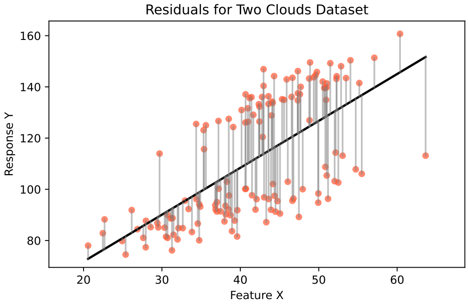
Residuals are the differences between observed and predicted values:

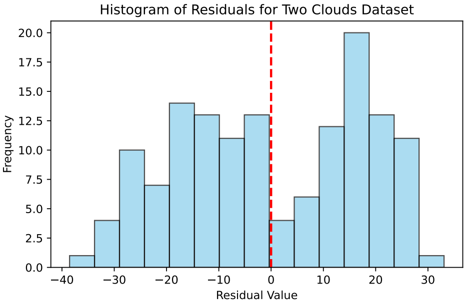
$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$



When Model Assumptions Are Violated

Non-random patterns in residuals can indicate model inadequacy or violated assumptions.





Parameter Estimation: Maximum Likelihood

The likelihood function for the regression model is:

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right)$$

Taking the logarithm and finding the values of β_0 and β_1 that maximize this expression leads to the same results as ordinary least squares.

Derivation of Least Squares Estimators

Starting with the log-likelihood function:

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Setting partial derivatives to zero:

$$\frac{\partial \ell}{\partial \beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \ell}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Least Squares Estimation

The maximum likelihood estimates are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

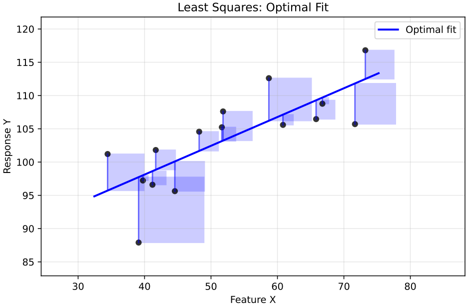
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

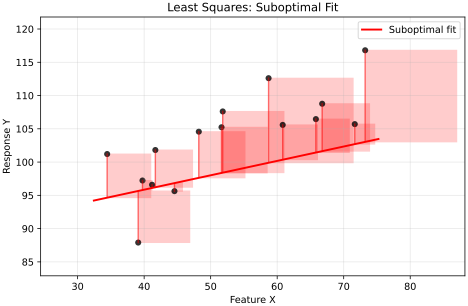
These formulas show that:

- The slope estimate is the ratio of the covariance between X and Y to the variance of X
- The intercept estimate ensures the regression line passes through the point $(\bar{x}, \bar{y})$

The Least Squares Principle

Ordinary least squares finds the line that minimizes the sum of squared vertical distances between observed points and the line.





Properties of the Regression Coefficient Estimates

The regression coefficient estimates have important statistical properties:

- They are unbiased: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$
- They follow a normal distribution in repeated sampling
- Their precision depends on sample size, predictor variability, and error variance

The sampling distribution of the slope estimator is:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Deriving the Sampling Distribution of $\hat{\beta}_1$ - Part 1

Starting with the formula for $\hat{\beta}_1$ and substituting the model equation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Substituting $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$:

$$\begin{aligned} y_i - \bar{y} &= (\beta_0 + \beta_1 x_i + \varepsilon_i) - (\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}) \\ &= \beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \end{aligned}$$

Therefore:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})[\beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

Deriving the Sampling Distribution of $\hat{\beta}_1$ - Part 2

Since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, we can simplify:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The term $\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i$ follows a normal distribution with:

$$E\left[\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right] = 0$$

$$Var\left[\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right] = \sigma^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Therefore:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Hypothesis Testing for Regression Coefficients

To test whether there's a significant linear relationship between X and Y, we test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$.
The test statistic is:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

The standard error of the slope coefficient is calculated as:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

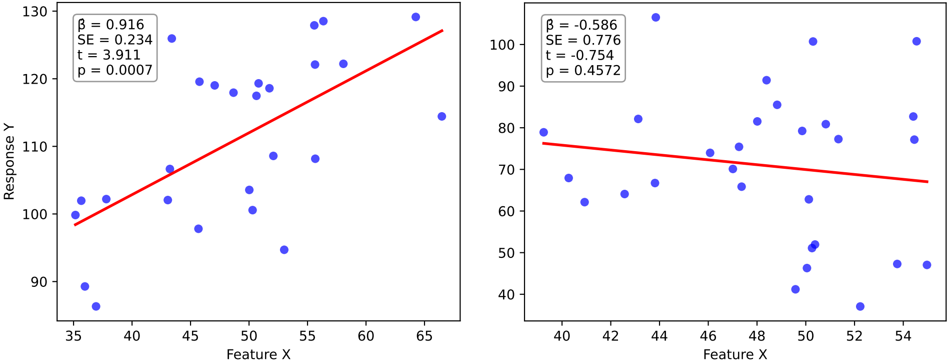
where $\hat{\sigma}^2$ is the estimated error variance:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

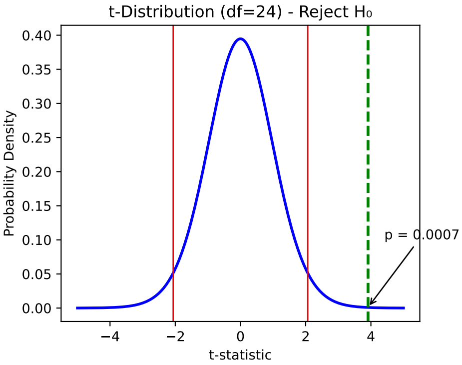
Under the null hypothesis, this follows a t-distribution with n-2 degrees of freedom.

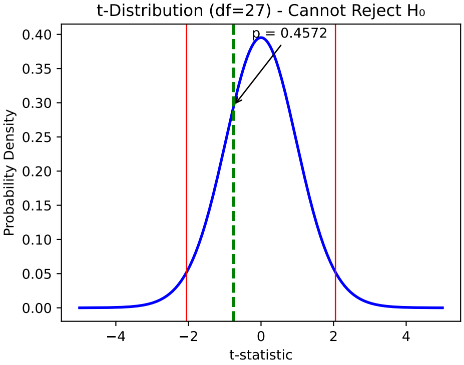
Testing Significance: Two Examples

Let's consider two examples to illustrate the concept of significance testing in linear regression.



Hypothesis Testing Results





Left: t-statistic in rejection region - conclude significant relationship
Right: t-statistic in non-rejection region - insufficient evidence

Confidence Intervals for Regression Parameters

To quantifying the uncertainty in these estimates. Confidence intervals provide a range of plausible values for the true parameters, accounting for sampling variability.

Definition

A confidence interval is a range of values constructed from sample data that is likely to contain the true population parameter with a specified level of confidence.

A $100(1-\alpha)\%$ confidence interval for the slope parameter β_1 is:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \times SE(\hat{\beta}_1)$$

These intervals quantify the precision of our estimates. Where

$SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

Interpreting Confidence Intervals

The correct interpretation of a 95% confidence interval:

> *If we were to repeat our sampling process many times, and calculate a 95% confidence interval from each sample, approximately 95% of these intervals would contain the true parameter value.*

The width of the confidence interval reflects estimation precision and depends on:

- Sample size (n): Larger samples yield narrower intervals
- Error variance (σ^2): Lower variance gives narrower intervals
- Variability in the predictor: Greater variability leads to more precise estimates
- Confidence level (1-α): Higher confidence requires wider intervals

Prediction in Regression Analysis

Beyond parameter estimation, regression models are valuable for prediction:

Definition

The **conditional mean response** is the expected value of Y given $X = x_0$, estimated as $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Also called the "fitted value" or "predicted value".

Definition

A **prediction for an individual response** is an estimate of a single future observation of Y when $X = x_0$.

Key distinction:

- Conditional mean response estimates the average Y for a given X
- Individual prediction accounts for both the regression line and random error
- Individual observations naturally vary around the regression line according to the error distribution—typically $\mathcal{N}(0, \sigma^2)$

Confidence Intervals for Mean Response

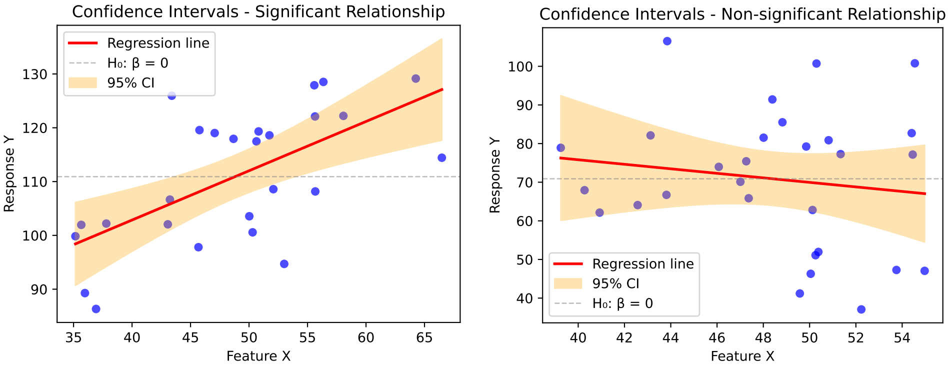
Confidence intervals for the mean response quantify uncertainty about average Y values:

$$\hat{y}_0 \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Properties of these intervals:

- Uncertainty is smallest when $x_0 = \bar{x}$ (center of data)
- Uncertainty increases as x_0 moves away from $\bar{x}$
- Forms a "band" around the regression line
- Width reflects precision of our estimate of the true regression line

Confidence Intervals for Mean Response - Visualization



Derivation of Confidence Intervals for Mean Response

The confidence interval for the mean response at x_0 can be derived from the variance of $\hat{\beta}_1$:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \times \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \hat{\beta}_1 \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x})$$

Since $\hat{y}_0$ is a linear function of $\hat{\beta}_1$, we can derive its variance:

$$\begin{aligned} Var(\hat{y}_0) &= Var(\bar{y} + \hat{\beta}_1(x_0 - \bar{x})) = Var(\bar{y}) + (x_0 - \bar{x})^2 \cdot Var(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \end{aligned}$$

Therefore, the confidence interval for the mean response is:

$$\hat{y}_0 \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Confidence vs. Prediction Intervals

Two types of intervals serve different purposes:

- **Confidence interval for mean response**: Quantifies uncertainty about the average value of Y for a given X value

$$\hat{y}_0 \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- **Prediction interval for individual observation**: Includes both uncertainty in the regression line and random variability of individual observations

$$\hat{y}_0 \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Note the additional "1" under the square root for prediction intervals, representing the inherent variability of individual observations.

Sources of Uncertainty in Prediction

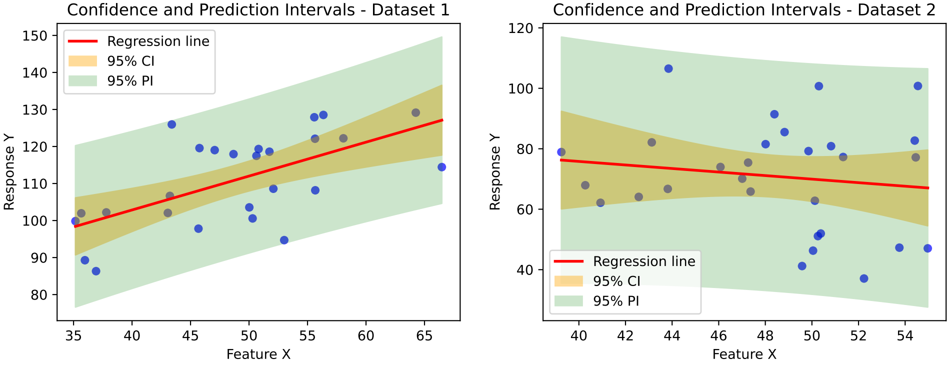
Prediction intervals are wider than confidence intervals because they account for two sources of uncertainty:

1. **Uncertainty in the estimated regression line**: Captured by the terms $\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
2. **Random variability of individual observations**: Captured by the "1" term, which comes from $Var(\varepsilon) = \sigma^2$

Mathematically, if $e_{\text{pred}} = Y_{\text{new}} - \hat{y}_0$, then:

$$Var(e_{\text{pred}}) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

Visualizing Both Types of Intervals



Inner bands (confidence intervals) show uncertainty about the mean response.
Outer bands (prediction intervals) show uncertainty about individual observations.

Choosing the Right Interval for Your Question

- Use **confidence intervals** when interested in the average effect: "What is the average protein content for cells of size 120 μm^3?"
- Use **prediction intervals** when forecasting individual outcomes: "What range of protein content might we observe in the next cell of size 120 μm^3?"

Both intervals have important applications in biological research:

- Confidence intervals help assess general trends and relationships
- Prediction intervals guide experimental design and set expectations for individual outcomes

Common Misconceptions About Confidence Intervals

When interpreting confidence intervals, be aware of these common misunderstandings:

- A 95% confidence interval does *not* mean there is a 95% probability that the true parameter falls within the interval
- Narrower confidence intervals don't always indicate better statistical estimates if model assumptions are violated
- Non-overlapping confidence intervals between groups don't automatically indicate statistically significant differences (this test is too conservative)

Formal hypothesis testing provides the appropriate framework for determining significance.

Limitations in Biological Applications

Several important caveats apply when using regression intervals in biological contexts:

- They assume the model is correct in its functional form (linearity)
- They assume homoscedasticity (constant error variance across all X values)
- They may not account for all sources of biological variability
- Extrapolation beyond the range of observed X values is particularly risky in biological systems, which often exhibit non-linear responses outside observed ranges

Sample Question 1

In simple linear regression, the model $Y = \beta_0 + \beta_1 X + \varepsilon$ assumes that the error term ε follows which distribution?

A) Uniform distribution
B) Student's t-distribution
C) Normal distribution with mean 0 and constant variance
D) Chi-square distribution
E) Exponential distribution

Sample Question 2

Which of the following would increase the precision (reduce the standard error) of the slope estimate?

A) Collecting data points with x-values close to the mean $\bar{x}$

B) Increasing the error variance σ^2

C) Reducing the sample size n

D) Increasing the spread of x-values around their mean

E) Focusing on values of x that produce the largest residuals

Sample Question 3

Consider the regression bands shown in the figure. If we observed a new data point at X = 3 with Y = 15, which of the following statements would be correct?

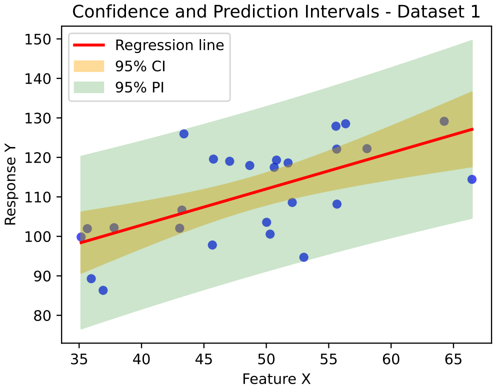
A) This observation provides evidence that the regression model is incorrect

B) This observation falls within the 95% prediction interval but outside the 95% confidence interval

C) This observation is considered an outlier because it falls outside both intervals

D) The probability that the true mean response at X = 3 equals 15 is 95%

E) We's expect 95% of observations at X = 3 to fall within the inner band



Sample Question 4

In constructing a confidence interval for the slope parameter in simple linear regression, which of these factors would make the interval narrower?

A) Decreasing the sample size

B) Increasing the confidence level from 95% to 99%

C) Smaller variability in the response variable (smaller σ^2)

D) Collecting data points with x-values very close to each other

E) Using a one-tailed rather than two-tailed test

Sample Question 5

A researcher measures enzyme activity (Y) as a function of substrate concentration (X) and fits a simple linear regression model. The 95% prediction interval at $X = 5$ is [10, 30], while the 95% confidence interval for the mean response at $X = 5$ is [15, 25]. Which of the following statements is correct?

A) The confidence interval is wider because it accounts for more sources of uncertainty

B) The estimate of the mean response at $X = 5$ is 15

C) If the experiment were repeated many times, about 95% of individual observations at $X = 5$ would fall between 10 and 30

D) The true mean response at $X = 5$ has a 95% probability of falling between 15 and 25

E) The prediction interval and confidence interval would become identical with a large enough sample size