Lecture 5 - **Simple Linear Regression**

BIOENG-210 Course Notes
Prof. Gioele La Manno

March 2024

1 Simple 1D Linear Regression Analysis

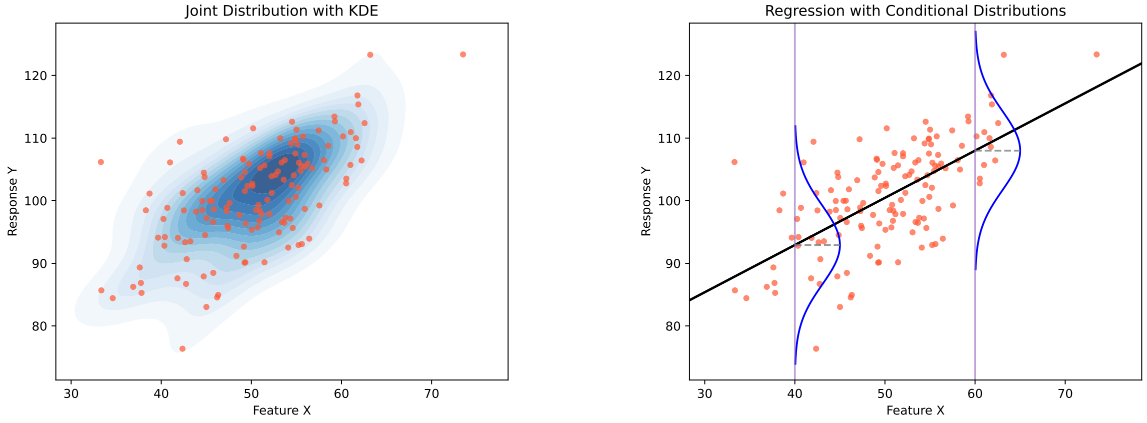
1.1 From Joint Distributions to Regression Models

In a previous lecture, we explored joint distributions and how they capture the relationship between two random variables. Linear regression can be viewed as a natural extension of these concepts, where we focus specifically on modeling the conditional distribution of one variable given another.

Recall that for two random variables X and Y with joint density $f_{X,Y}(x, y)$, the conditional density of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

Linear regression essentially models this conditional distribution, making specific assumptions about its form. Specifically, linear regression assumes that the conditional distribution of Y given X is a normal distribution with a mean that is a linear function of X.



(a) Joint distribution between two variables X and Y

(b) Regression line with and conditional distributions

Figure 1: **Connection between joint distributions and regression**

- **Left panel (Joint distribution):** This figure illustrates the joint probability distribution between two continuous variables X and Y. The scattered points show actual observations, while the underlying heat map represents the estimated joint density. Note the correlation between the variables - as X increases, Y tends to increase as well, though with considerable variation.

- **Right panel (Conditional distributions):** This figure shows how linear regression relates to the joint distribution. The blue line represents the best-fit regression line through the data. The key insight is shown by the vertical Gaussian curves at two different X values - these represent the conditional distributions P(Y—X) at those specific X values. Each curve is centered exactly at the regression line, demonstrating that the regression line represents the mean of Y given X. The constant width of these distributions illustrates the homoscedasticity assumption (constant variance regardless of X value). This visualization captures the essence of linear regression as modeling the conditional expectation E[Y—X].

Linear regression thus provides a way to characterize how the distribution of Y changes systematically with X. Rather than modeling the full joint distribution, it focuses specifically on how the center of the conditional distribution shifts as a function of the predictor variable.

1.2 The Linear Model: A Probabilistic Perspective

The core insight of linear regression is that we can model the conditional distribution of Y given X as a normal distribution whose mean is a linear function of X, while its variance remains constant. Using formal probabilistic notation:

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

This concise statement encapsulates the entire linear regression model. It tells us that:

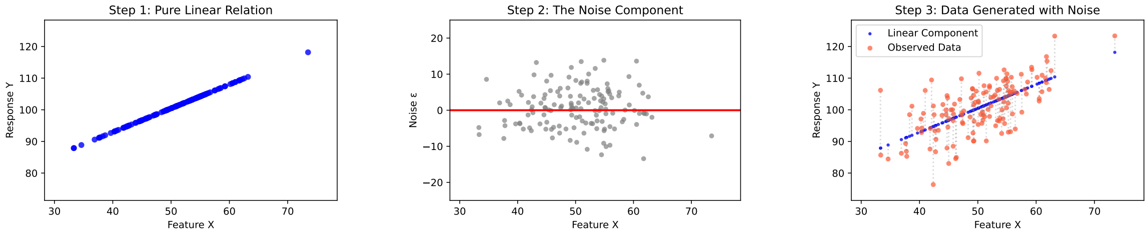
- For any fixed value of $X = x$, the variable Y follows a normal distribution
- The mean of this distribution is $\beta_0 + \beta_1 x$ (a linear function of x)
- The variance of this distribution is σ^2 (constant across all values of x)

Using the model notation common in statistical literature, we write:

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

Where ε_i represents the random noise or error term.

Note that this decomposition separates the deterministic part of the model $(\beta_0 + \beta_1 X_i)$ from the random noise (ε_i) is important but not should be taken for granted. It is only possible for the normal distribution which has the property that the sum of normal random variables is also normal (e.g. for poisson distribution this is not true).



(a) 1 - Pure Linear relation

(b) 2 - The noise

(c) 3 - Data generated with noise

Figure 2: **The conceptual model assumed to generate the data**

1.3 Interpretation Through Conditional Expectations

Another way to understand regression is through conditional expectations. While the concept of conditional expectation builds on our previous discussions , it is worthwhile to establish its definition.

Definition 1.1. The conditional expectation of a random variable Y given another random variable X is defined as:

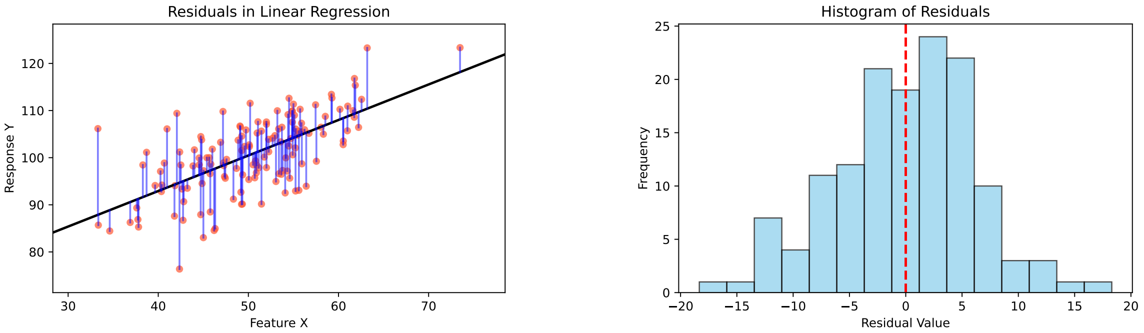
$$E[Y|X] = \int y f_{Y|X}(y|x)dy$$

The linear regression model assumes:

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

This means that the expected value of Y given a specific value of X is a linear function of x. 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



(a) Residuals in linear regression

(b) Histogram of residuals

Figure 3: **Residuals in linear regression**

When fitting the model to data, we observe the difference between the actual and predicted values, which we call residuals:

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

These residuals are the empirical realizations of the error terms in our model. The distinction is that ε_i represents the theoretical error in the population model, while $\hat{\varepsilon}_i$ represents the observed residual in our specific sample. Analyzing these residuals is crucial for assessing model fit and checking regression assumptions.

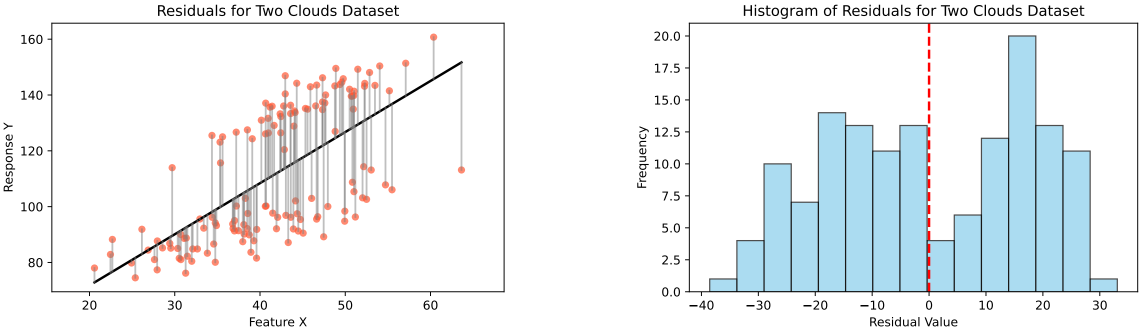
1.4 Parameter Estimation: Maximum Likelihood and Least Squares

Given this probabilistic model, how do we estimate the unknown parameters β_0, β_1, and σ^2 from data? Two equivalent approaches lead to the same result:

1.4.1 Maximum Likelihood Estimation

The likelihood function for n independent observations is:

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



(a) Data showing two clouds of points

(b) Bimodal Histogram of residuals

Figure 4: **Residuals not following normal model**

Taking the logarithm:

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

Maximizing this log-likelihood with respect to the parameters requires setting the partial derivatives to zero:

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

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

From the first equation, we get:

$$\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \beta_1 \bar{x}$$

This equation confirms an important geometric property: the regression line must pass through the point $(\bar{x}, \bar{y})$.

Substituting this into the second equation, we get the so called **normal equations**:

$$\sum_{i=1}^{n} x_i(y_i - \bar{y} + \beta_1\bar{x} - \beta_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i + \beta_1\bar{x}\sum_{i=1}^{n} x_i - \beta_1\sum_{i=1}^{n} x_i^2 = 0$$

Note that $\sum_{i=1}^{n} x_i = n\bar{x}$, so:

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + \beta_1(n\bar{x}^2 - \sum_{i=1}^{n} x_i^2) = 0$$

Solving for β_1:

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

Recognizing the numerator as the covariance between X and Y, and the denominator as the variance of X

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Least Squares Estimators for Simple Linear Regression

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

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

1.4.2 Connection to Least Squares

We can also view the estimation of a more familiar ”least squares” procedure. Looking at our log-likelihood function:

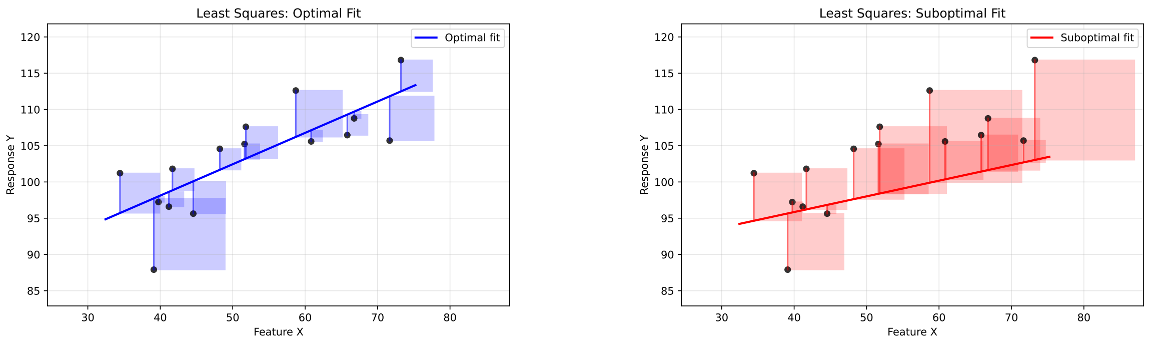
$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

When maximizing this expression, only the last term depends on β_0 and β_1. Therefore, maximizing the log-likelihood is equivalent to maximizing:

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

Since σ^2 is positive, and considering the minus sign, this is equivalent to minimizing the residual sum of squares (RSS) as a function of β_0 and β_1:

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 = \text{RSS}(\beta_0, \beta_1)$$



(a) Optimal least squares fit

(b) Suboptimal line with larger squared residuals

Figure 5: **Least squares: Minimizing the sum of squared vertical distances**

1.5 Properties of the Estimates

The regression coefficient estimates have several important properties:

- They are unbiased: $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$
- They are consistent: as sample size increases, they converge to the true parameter values
- They are efficient: among unbiased estimators, they have minimum variance
- They follow a normal distribution in repeated sampling (when errors are normal)

The sampling distribution of the slope estimator is particularly important for inference:

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

Since σ^2 is typically unknown, we estimate it and use a t-distribution for inference:

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

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

1.5.1 Demonstration of the Sampling Distribution Formula - (Optional content!)

Let's derive the sampling distribution of $\hat{\beta}_1$ to understand why it follows a normal distribution with the variance given above.

Recall that our model is:

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

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently.

We established earlier that:

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

Let's substitute the model equation into this formula:

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

where $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$.

Now we can substitute this into our formula for $\hat{\beta}_1$:

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

This shows that $\hat{\beta}_1$ equals the true parameter β_1 plus a term involving the errors.

We can simplify the numerator further by noting that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, which means:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) &= \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i - \bar{\varepsilon}\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\end{aligned}$$

Therefore:

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

Now, to find the distribution of $\hat{\beta}_1$, we need to understand the distribution of $\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i$.

Since $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently, and the x_i values are fixed (not random), each term $(x_i - \bar{x})\varepsilon_i$ follows a normal distribution with:

$$\begin{aligned}E[(x_i - \bar{x})\varepsilon_i] &= (x_i - \bar{x})E[\varepsilon_i] = 0 \\ Var[(x_i - \bar{x})\varepsilon_i] &= (x_i - \bar{x})^2 Var[\varepsilon_i] = (x_i - \bar{x})^2\sigma^2\end{aligned}$$

The sum $\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i$ is therefore a sum of independent normal random variables, which is also normally distributed with:

$$\begin{aligned}E\left[\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right] &= \sum_{i=1}^{n}E[(x_i - \bar{x})\varepsilon_i] = 0 \\ Var\left[\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i\right] &= \sum_{i=1}^{n}Var[(x_i - \bar{x})\varepsilon_i] = \sigma^2\sum_{i=1}^{n}(x_i - \bar{x})^2\end{aligned}$$

Therefore:

$$\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i \sim \mathcal{N}(0, \sigma^2\sum_{i=1}^{n}(x_i - \bar{x})^2)$$

Consequently, the sampling distribution of $\hat{\beta}_1$ is:

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

1.5.2 Interpretation of the Variance Formula

This confirms our earlier statement about the sampling distribution of $\hat{\beta}_1$. The variance formula shows that:

- As sample size n increases, the variance decreases, making the estimate more precise
- Greater variability in the predictor variable x (larger $\sum_{i=1}^{n}(x_i - \bar{x})^2$) leads to smaller variance and more precise estimates
- Higher error variance σ^2 leads to higher variance in our estimate of β_1

Since we typically do not know σ^2, we estimate it from the data using $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2}$. When we use this estimate, the standardized statistic $\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$ follows a t-distribution with $n-2$ degrees of freedom rather than a standard normal distribution.

1.6 Inference in Linear Regression

With the sampling distribution established, we can perform various inferential tasks:

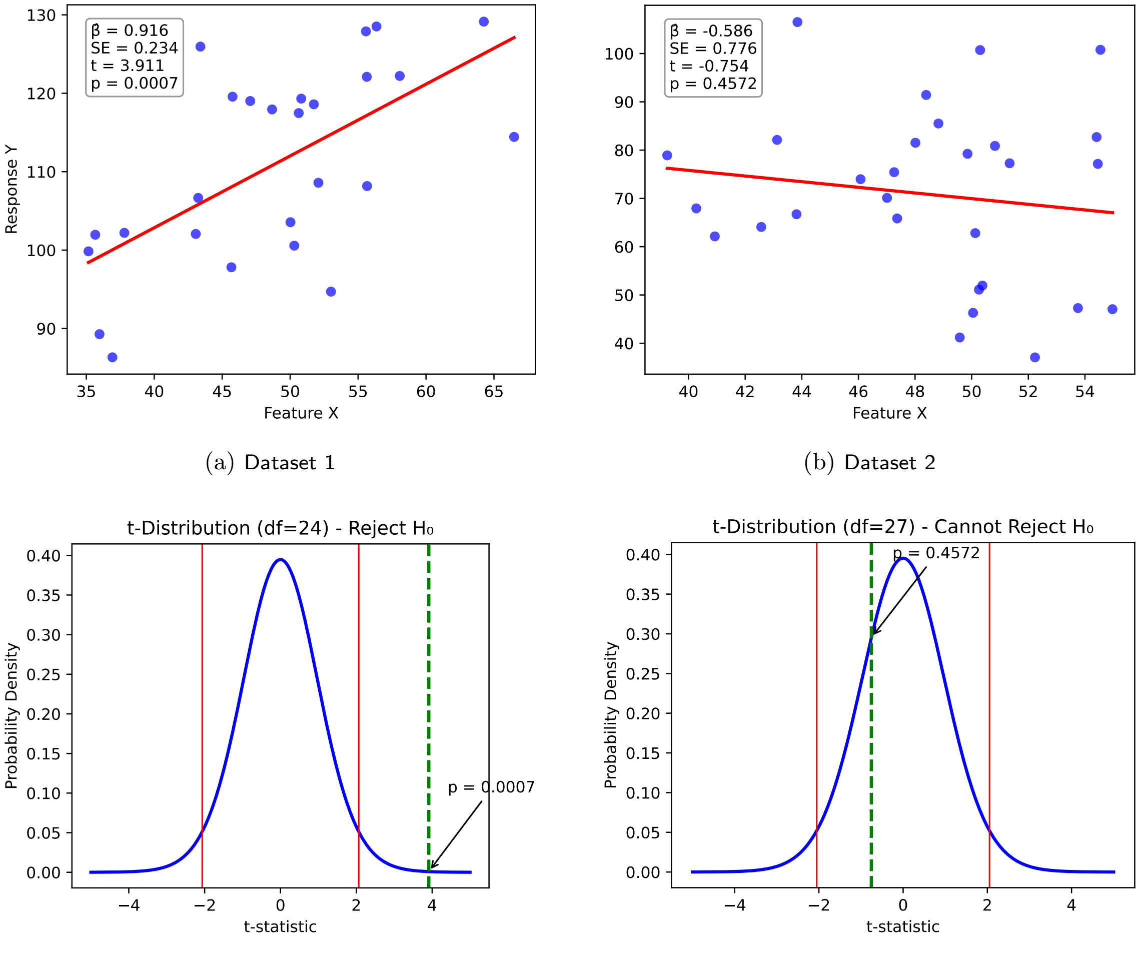
1.6.1 Hypothesis Testing for Regression Coefficients

To test whether there's a significant linear relationship between X and Y, we test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$.

The test statistic is:

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

Under the null hypothesis, this follows a t-distribution with $n - 2$ degrees of freedom. We can calculate a p-value and make decisions just as in our earlier discussion of hypothesis testing.



(a) Dataset 1

(b) Dataset 2

(c) Dataset 1 test: Rejecting the null hypothesis

(d) Dataset 2 test: Failing to reject the null hypothesis

Figure 6: **Testing significance of regression slope**

1.7 Confidence Intervals for Regression Parameters

Having estimated the regression parameters, we now turn to quantifying the uncertainty in these estimates. Confidence intervals provide a range of plausible values for the true parameters, accounting for sampling variability.

Definition 1.2. 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.

For the slope parameter β_1 in simple linear regression, a $100(1-\alpha)\%$ confidence interval is given by:

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

where:

- $\hat{\beta}_1$ is our point estimate of the slope
- $t_{\alpha/2,n-2}$ is the critical value from the t-distribution with $n-2$ degrees of freedom
- $SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$ is the standard error of the slope estimate

Similarly, we can construct a confidence interval for the intercept β_0:

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2} \times SE(\hat{\beta}_0)$$

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

1.7.1 Interpretation of Confidence Intervals

The correct interpretation of a 95% confidence interval is:

> 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.

For example, if our 95% confidence interval for β_1 is [0.3, 0.7], we can state: "We are 95% confident that the true change in the expected value of Y for a one-unit increase in X is between 0.3 and 0.7 units."

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

- Sample size (n): Larger samples yield narrower intervals
- Variance of the error term (σ^2): Lower error variance gives narrower intervals
- Variability in the predictor ($\sum_{i=1}^{n}(x_i-\bar{x})^2$): Greater variability leads to more precise estimates and narrower intervals
- Confidence level (1-α): Higher confidence requires wider intervals

1.8 Prediction in Regression Analysis

Beyond parameter estimation, regression models are often used to make predictions for new observations. There are two types of predictions we might want to make:

Definition 1.3. The **conditional mean response** is the expected value of Y given a specific value of $X = x_0$, estimated as $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Also informally called the "fitted value" or "predicted value".

Definition 1.4. A **prediction for an individual response** is an estimate of a single future observation of Y when $X = x_0$.

While the conditional mean response estimates the average value of Y for a given X, an individual prediction accounts for both the systematic component (the regression line) and the random error component (ε). This distinction is critical because individual observations naturally vary around the regression line according to the error distribution—typically $\mathcal{N}(0, \sigma^2)$ in linear regression.

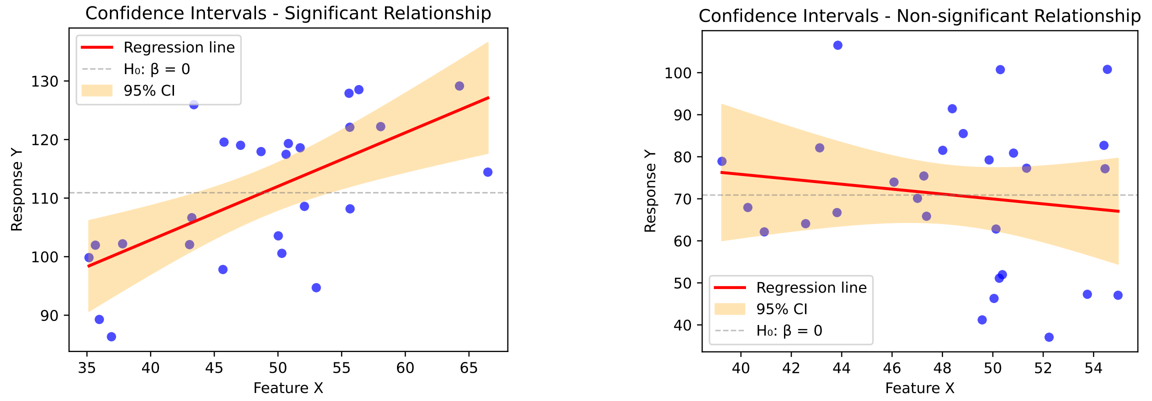
1.8.1 Confidence Intervals for the Mean Response

A confidence interval for the mean response at $X = x_0$ is given by:

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

This interval quantifies our uncertainty about the average value of Y for a given value of $X = x_0$. The uncertainty is smallest when $x_0 = \bar{x}$ and increases as x_0 moves away from $\bar{x}$.

The equation above implies confidence intervals are wider at the edges of the data, where we have less information to estimate the slope and narrower in the center, where we have more data points and more information to estimate the slope One we can visualize CI as a band around the regression line.



(a) Dataset 1: Confidence intervals for the regression line

(b) Dataset 2: Confidence intervals for the regression line

Figure 7: **Visualization of confidence intervals for the regression**

1.9 Demonstration of CI on the mean response - (Optional content!)

The equation above is derived as follows from the general formula for the confidence interval of the slope parameter:

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

The term under the square root can be simplified as:

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

To find the confidence interval for the predicted value at a specific point x_0, we need to consider that:

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

Since $\hat{y}_0$ is a linear function of $\hat{\beta}_1$, the variance of $\hat{y}_0$ can be derived from the variance of $\hat{\beta}_1$:

$$Var(\hat{y}_0) = Var(\bar{y} + \hat{\beta}_1(x_0-\bar{x})) = Var(\bar{y}) + (x_0-\bar{x})^2 \cdot Var(\hat{\beta}_1)$$

Where $Var(\bar{y}) = \frac{\sigma^2}{n}$ and $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
Therefore:

$$Var(\hat{y}_0) = \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

This leads to the confidence interval for the mean response at $X = x_0$:

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

This expression shows that the confidence interval is centered around the predicted value $\hat{y}_0$ and widens as x_0 moves away from $\bar{x}$.

1.9.1 Common Misconceptions About Confidence Intervals

When interpreting confidence intervals in regression analysis, several misconceptions commonly arise that deserve clarification. First, many incorrectly believe that a 95% confidence interval means there is a 95% probability that the true parameter falls within the calculated interval. In reality, the confidence level refers to the reliability of the estimation procedure across repeated sampling, not the probability of the parameter being in any specific interval. Once calculated from a particular dataset, an interval either contains the true value or it doesn't.

Another frequent misunderstanding is that narrower confidence intervals always indicate better statistical estimates. While narrow intervals often suggest precise estimation when properly constructed, artificially narrow intervals can result from violated model assumptions or inappropriate analysis methods. The validity of the confidence interval depends not just on its width but on whether the underlying statistical assumptions are satisfied.

Finally, researchers sometimes conclude that non-overlapping confidence intervals between groups automatically indicate statistically significant differences. This heuristic is actually too conservative in practice; confidence intervals can overlap to some extent while the difference between groups remains statistically significant. Formal hypothesis testing, rather than visual inspection of confidence interval overlap, provides the appropriate framework for determining significance.

1.9.2 Prediction Intervals for Individual Observations

When predicting a single future observation rather than the mean response, we must account for both the uncertainty in the estimated regression line and the random variability around this line. A prediction interval for an individual response at $X = x_0$ is:

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

Notice that this is wider than the confidence interval for the mean response, due to the addition of "1" under the square root. This additional term has a precise statistical interpretation: it represents the inherent variability of individual observations around the regression line.

To understand where this "1" comes from, consider the two distinct sources of uncertainty when predicting a new observation Y_{new} at $X = x_0$:

1. **Uncertainty in the estimated regression line**: This is captured by the terms $\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, which represent our uncertainty about where the true mean response lies.

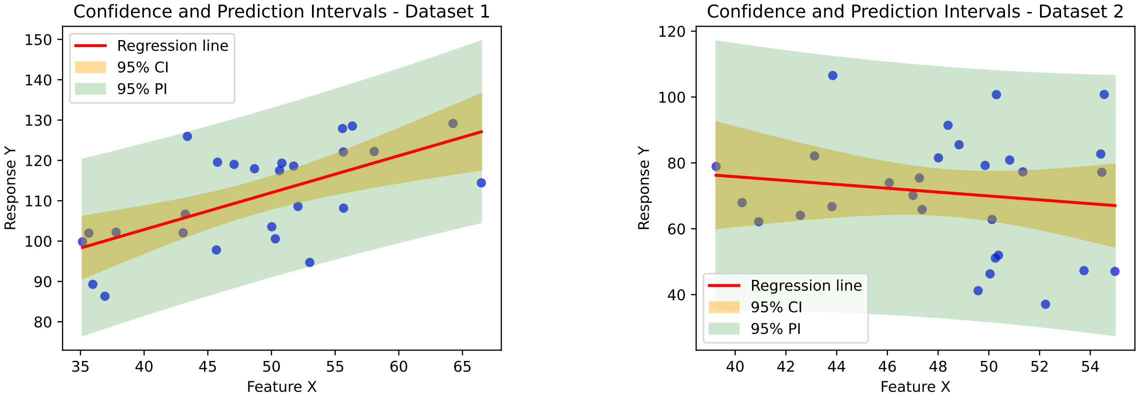
2. **Random variability of individual observations**: This is captured by the "1" term, which comes directly from the variance of the error term $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ in our model. Since $Var(\varepsilon) = \sigma^2$, and we standardize by dividing by σ^2, this gives us the "1" inside the square root.

Mathematically, if we denote the prediction error as $e_{\text{pred}} = Y_{\text{new}} - \hat{y}_0$, its variance is:

$$Var(e_{\text{pred}}) = Var(Y_{\text{new}} - \hat{y}_0) = Var(Y_{\text{new}}) + Var(\hat{y}_0) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

Dividing by σ^2 gives us the expression under the square root in our prediction interval formula.

This is why prediction intervals are always wider than confidence intervals for the mean response—they must account for both the uncertainty in our estimate of the mean response (model uncertainty) and the natural variability of individual observations around that mean (data variability). Even with a perfectly known regression line, individual observations would still vary around it according to the error distribution.



(a) Dataset 1: Confidence and prediction intervals

(b) Dataset 2: Confidence and prediction intervals

Figure 8: **Confidence intervals vs. prediction intervals**

1.9.3 Practical Interpretation and Usage

The choice between confidence and prediction intervals depends on the research question:

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

In biological research, both types of intervals have important applications:

- Confidence intervals help assess general trends and relationships, supporting theory development and hypothesis testing
- Prediction intervals guide experimental design, sample size planning, and set expectations for individual experimental outcomes

The distinction is crucial when communicating results: confidence intervals address population parameters, while prediction intervals address future observations. Failing to distinguish between them can lead to overly optimistic expectations about the precision of individual predictions.

1.9.4 Limitations in Biological Applications

In biological contexts, several caveats apply to these intervals:

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

Despite these limitations, confidence and prediction intervals remain valuable tools for quantifying uncertainty in regression analyses, provided they are used and interpreted appropriately in the biological context.