MAT 616 – Spring 2025
Linear Regression Line Using Calculus

1. Model Setup and Error Function

Assume we have a data set of n points:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n).$$

We model the relationship between x and y with the linear equation:

$$y = \alpha + \beta x.$$

Explanation: Here, α represents the intercept and β represents the slope. This is our assumed linear relationship.

The sum of squared errors (or residuals) is defined as:

$$S(\alpha, \beta) = \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2.$$

Explanation: This function $S(\alpha, \beta)$ measures the total squared difference between the observed values y_i and the predictions from our model $\alpha + \beta x_i$. Minimizing this function yields the best-fit line.

2. Minimizing the Error Function Using Calculus

To find the optimal values of α and β, we take partial derivatives of $S(\alpha, \beta)$ with respect to each parameter and set them equal to zero.

a. Partial Derivative with Respect to α

Differentiate S with respect to α:

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^{n} \Big(y_i - \alpha - \beta x_i \Big) = 0.$$

Explanation: The derivative is computed using the chain rule on the squared term. Setting this derivative to zero finds the condition for the minimum error with respect to α.

Dividing both sides by -2 gives:

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0.$$

Explanation: Removing the constant factor simplifies the equation without affecting the zero condition.

Expanding the sum, we have:

$$\sum_{i=1}^{n} y_i - n\alpha - \beta \sum_{i=1}^{n} x_i = 0.$$

Explanation: Since α is constant with respect to the summation index i, it factors out as $n\alpha$.

Solving for α:

$$n\alpha = \sum_{i=1}^{n} y_i - \beta \sum_{i=1}^{n} x_i \quad \Longrightarrow \quad \alpha = \frac{1}{n} \sum_{i=1}^{n} y_i - \beta \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Define the sample means:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Thus, the expression for the intercept becomes:

$$\alpha = \bar{y} - \beta \bar{x}.$$

This shows that the intercept α is adjusted from the mean of y by the product of the slope β and the mean of x.

b. Partial Derivative with Respect to β

Next, differentiate S with respect to β:

$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} x_i \Big(y_i - \alpha - \beta x_i \Big) = 0.$$

Explanation: Here, the derivative brings down a factor x_i when differentiating βx_i. Setting this derivative to zero gives the condition for the optimal slope.

Substitute the expression for α ($\alpha = \bar{y} - \beta \bar{x}$) into the equation:

$$-2 \sum_{i=1}^{n} x_i \left[y_i - (\bar{y} - \beta \bar{x}) - \beta x_i \right] = 0.$$

Explanation: By substituting α, we eliminate it from the equation so that the expression depends only on β and the data.

Simplify the term inside the brackets:

$$y_i - \bar{y} + \beta \bar{x} - \beta x_i = (y_i - \bar{y}) - \beta (x_i - \bar{x}).$$

Explanation: This groups together the deviations from the means and the terms with β.

Thus, the derivative equation becomes:

$$-2 \sum_{i=1}^{n} x_i \left[(y_i - \bar{y}) - \beta (x_i - \bar{x}) \right] = 0.$$

Explanation: The factor -2 can be cancelled later, as it does not affect the location of the minimum.

Dividing both sides by -2 and expanding:

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

Explanation: This step isolates the terms involving β.

Notice that we can express the denominator term as:

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

Since $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, we have $\bar{x} \sum_{i=1}^{n} x_i = n\bar{x}^2$. Thus,

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

Explanation: This expression represents the total variability of x about its mean.

Similarly, the numerator simplifies as follows. Expand:

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} \left[(x_i - \bar{x}) + \bar{x}\right] (y_i - \bar{y}).$$

Since

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

we obtain:

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

Explanation: The cancellation occurs because the sum of the deviations of y from its mean is zero. This sum is essentially the covariance between x and y.

Thus, our equation reduces to:

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \beta \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = 0.$$

Solving for β yields:

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

Explanation: The numerator represents the covariance between x and y, while the denominator represents the total variability (or variance) of x multiplied by n. Notice that the denominator can be rewritten as:

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

so that the final formula for the slope becomes:

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

This is the standard least-squares estimator for the slope of the regression line.

3. Computing the Intercept α

Now that β is known, substitute it back into the expression for α:

$$\alpha = \bar{y} - \beta\bar{x}.$$

Explanation: This ensures that the regression line passes through the point $(\bar{x}, \bar{y})$, the centroid of the data.

4. Final Regression Line

The derived least-squares estimators are:

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \alpha = \bar{y} - \beta\bar{x}.$$

Thus, the best-fit linear regression model is:

$$y = \alpha + \beta x.$$

Explanation: These equations fully define the regression line using the parameters α and β that minimize the error function.

Summary of the Process

1. **Model Specification:** Begin with the linear model $y = \alpha + \beta x$.
2. **Error Function:** Define $S(\alpha, \beta) = \sum(y_i - (\alpha + \beta x_i))^2$.
3. **First Derivative with respect to α:** Differentiate and set equal to zero to obtain $\alpha = \bar{y} - \beta\bar{x}$.
4. **First Derivative with respect to β:** Differentiate, substitute the value of α, and solve for β.
5. **Final Equations:** Substitute back to get $\alpha = \bar{y} - \beta\bar{x}$ and form the regression line $y = \alpha + \beta x$.