Lecture 11

§4.3 Modeling Linear Trends

1. §4.3 Modeling Linear Trends

Definition.

When a scatterplot displays a linear relationship, we summarize the pattern using a straight line. The **regression line** (also known as the *line of best fit* or the *least-squares regression line*) is a mathematical model used for describing the linear trend and making predictions about observed or future values.

The equation of the regression line is written as:

$$\boxed{\hat{y} = a + bx}$$

where:

- $\hat{y}$ (pronounced "y-hat") is the **predicted value** of the response variable.
- a is the **y-intercept**, the predicted value of y when $x = 0$.
- b is the **slope**, which measures the predicted change in y for each one-unit increase in x.
- x is the value of the explanatory variable.

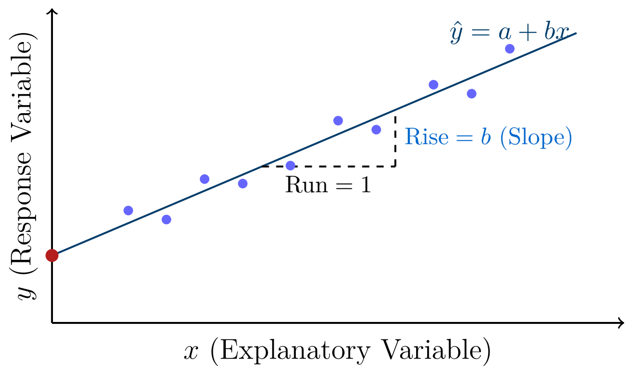
1.1 Terminology for the Variables

The input and output variables have specific names depending on the statistical context:

Variable	Synonyms and Roles
Input Variable (x)	• **Independent variable** • **Predictor variable** • **Explanatory variable**
Output Variable (y or $\hat{y}$)	• **Dependent variable** • **Predicted variable** • **Response variable**

1.2 Visualizing the Slope and Intercept

The geometry of the regression line illustrates how the y-intercept a establishes the starting value and the slope b dictates the rate of change:



Interpretation Templates for Slope and Intercept:

- **Slope (b):** "For each additional 1 unit increase in x, the model predicts that y will change (increase or decrease) by an average of b **units**."
- **Intercept (a):** "When the explanatory variable x is 0, the predicted value of the response variable y is a."
  Practical note: The y-intercept only makes practical sense if $x = 0$ is a realistic value and within the range of observed data.

1.3 Formulas for Finding the Line of Best Fit

The least-squares regression line minimizes the sum of squared vertical errors. Rather than guessing, the slope b and intercept a are calculated directly from summary statistics:

$$\boxed{b = r\left(\frac{s_y}{s_x}\right)} \quad \text{and} \quad \boxed{a = \bar{y} - b\bar{x}}$$

- r is the sample correlation coefficient.
- s_x and s_y are the sample standard deviations of x and y.
- $\bar{x}$ and $\bar{y}$ are the sample means of x and y.
- Notice that the line **always** passes through the center point of the data: $(\bar{x}, \bar{y})$.

Example 1: Predicting Weight from Height

Recall our data from Lecture 10 concerning the heights (x, in inches) and weights (y, in pounds) of six adult women:

Height (x)	61	62	63	64	66	68
Weight (y)	104	110	141	125	170	160

From our previous calculations, we know:

$$\bar{x} = 64.0 \text{ in}, \quad s_x \approx 2.61 \text{ in}, \quad \bar{y} = 135.0 \text{ lbs}, \quad s_y \approx 25.97 \text{ lbs}, \quad r \approx 0.907$$

Step 1: Calculate the Slope b

$$b = r\left(\frac{s_y}{s_x}\right) = 0.907 \times \left(\frac{25.97}{2.61}\right) \approx 0.907 \times 9.950 \approx \boxed{9.02}$$

Interpretation: For every additional 1 inch in height, a woman's weight is predicted to increase by approximately 9.02 **pounds**.

Step 2: Calculate the y-intercept a

$$a = \bar{y} - b\bar{x} = 135.0 - (9.02)(64.0) = 135.0 - 577.28 = \boxed{-442.28}$$

Thus, the linear regression model is:

$$\boxed{\hat{y} = -442.28 + 9.02x} \quad \text{or} \quad \widehat{\text{Weight}} = -442.28 + 9.02(\text{Height})$$

Step 3: Making a Prediction

Predict the weight of a woman who is 65 inches tall:

$$\hat{y} = -442.28 + 9.02(65) = -442.28 + 586.30 = \boxed{144.02 \text{ lbs}}$$

1.4 Residuals: Measuring Prediction Errors

An individual observation rarely falls precisely on the regression line. The difference between what actually happened and what the model predicted is called the **residual**.

Definition of Residual:

$$\boxed{e = y - \hat{y} = \text{Observed } y - \text{Predicted } y}$$

- **Positive residual ($y > \hat{y}$):** The observation lies *above* the line (the model underestimated the actual value).
- **Negative residual ($y < \hat{y}$):** The observation lies *below* the line (the model overestimated the actual value).

Example 2: Computing a Residual

From our data table, a woman who is 63 inches tall actually weighs 141 pounds ($y = 141$).

1. Find her predicted weight:
$$\hat{y} = -442.28 + 9.02(63) = 126.0 \text{ lbs}$$
2. Compute the residual:
$$e = y - \hat{y} = 141 - 126.0 = \boxed{+15.0 \text{ lbs}}$$

   The woman weighs 15.0 pounds more than predicted by the regression line.

1.5 The Danger of Extrapolation

Warning: Beware of Extrapolation!
Extrapolation is using a regression line to make predictions for x-values that lie far outside the range of values used to build the model.

- We have no evidence that the relationship remains linear outside our observed range.
- *Example:* Using our height-weight model to predict the weight of a newborn baby who is 20 inches tall gives:

$$\hat{y} = -442.28 + 9.02(20) = -261.88 \text{ lbs (impossible!)}$$

2. Simple Linear Regression in StatCrunch

Example.

Open the **`popdensity and crime`** dataset in StatCrunch. Use the **"Simple Linear"** tool under the

Stat → **Regression** menu to find the regression line for the following pairs of columns. Interpret the slope and intercept where appropriate.

Steps to Follow in StatCrunch

1. Open the **`popdensity and crime`** dataset in StatCrunch.
2. Go to the top menu and select:

$$\textbf{Stat} \longrightarrow \textbf{Regression} \longrightarrow \textbf{Simple Linear}$$

3. In the dialog box:
   - Select the specified column for the **X-Variable** (Predictor/Explanatory).
   - Select the specified column for the **Y-Variable** (Response/Dependent).
4. Leave all other settings at their defaults and click **Compute!**.
5. Locate the fitted equation, the intercept value (a), and the slope coefficient (b) in the output table.

Part 1: The pop1990 and pop2000 Columns

- **X-Variable:** `pop1990` **Y-Variable:** `pop2000`

Regression Equation:

$$\widehat{\text{pop2000}} = \underline{\hspace{4cm}} + \underline{\hspace{4cm}}\,(\text{pop1990})$$

Interpretations:

1. **Slope (b):** ____

   ____
2. **Intercept (a):** Is it appropriate to interpret the intercept here? Why or why not?

   ____

   ____

Part 2: The pop2000 and totcrimerate Columns

- **X-Variable:** `pop2000` **Y-Variable:** `totcrimerate`

Regression Equation:

$$\widehat{\text{totcrimerate}} = \underline{\hspace{4cm}} + \underline{\hspace{4cm}}\,(\text{pop2000})$$

Interpretations:

1. **Slope (b):** ____

   ____
2. **Intercept (a):** Is it appropriate to interpret the intercept here? Why or why not?

   ____

   ____

Part 3: The pop2000 and Rank Pop Columns

- **X-Variable:** `pop2000` **Y-Variable:** `Rank Pop`

Regression Equation:

$$\widehat{\text{Rank Pop}} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ + \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (\text{pop2000})$$

Interpretations:

1. **Slope (b):** ____

   ____

2. **Intercept (a):** Is it appropriate to interpret the intercept here? Why or why not?

   ____

   ____