STAC67: Regression Analysis

Lecture 3

Thibault Randrianarisoa

September 16, 2026

University of Toronto Scarborough

Simple linear model with normal errors

- The random errors are sometimes assumed to be normally distributed.

- Simple linear model with normal errors:

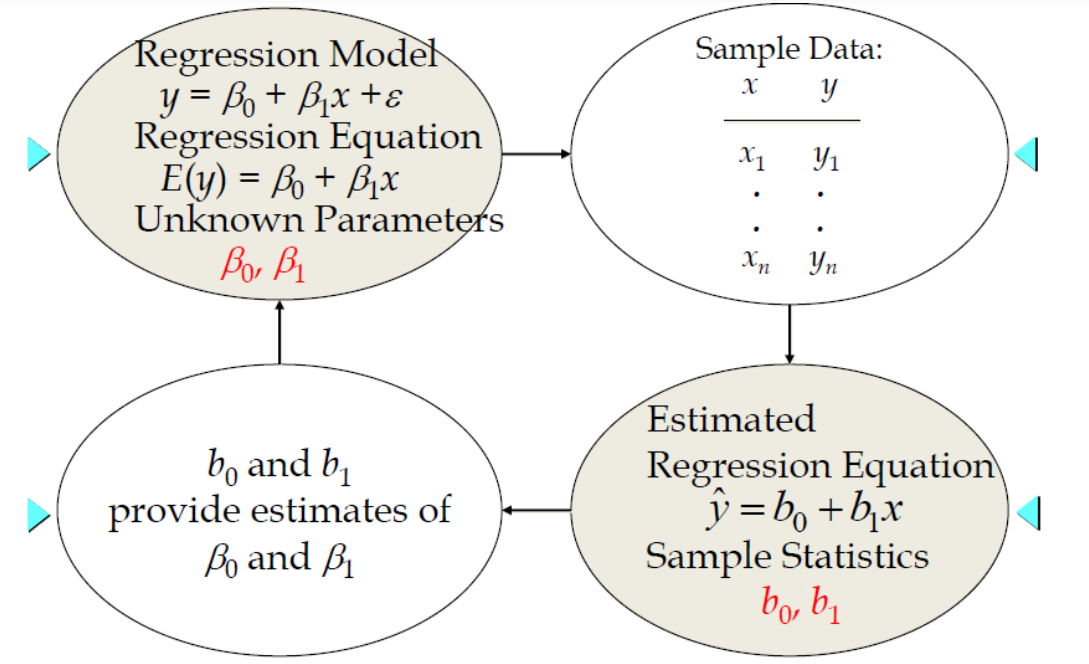
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

  where ϵ_i are independent and identically distributed (i.i.d) with normal distribution with mean 0 and variance σ^2.

- In terms of Y, this means that we have the conditional distribution

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

Parameter Estimation



- Maximum likelihood estimation

$$f(y; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

Likelihood Function:

$$L(\beta_0, \beta_1, \sigma^2) = \Pi_{i=1}^{n} f(y_i; \beta_0, \beta_1, \sigma^2)$$

Parameter Estimation

To find m.l.e for β_0, β_1, and σ^2, we can maximize the following:

Log-likelihood Function:

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

Maximizing Likelihood function w.r.t $\beta_0,\ \beta_1$ is equivalent to minimizing,

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

Parameter Estimation

Method of least squares

Simple linear model:

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

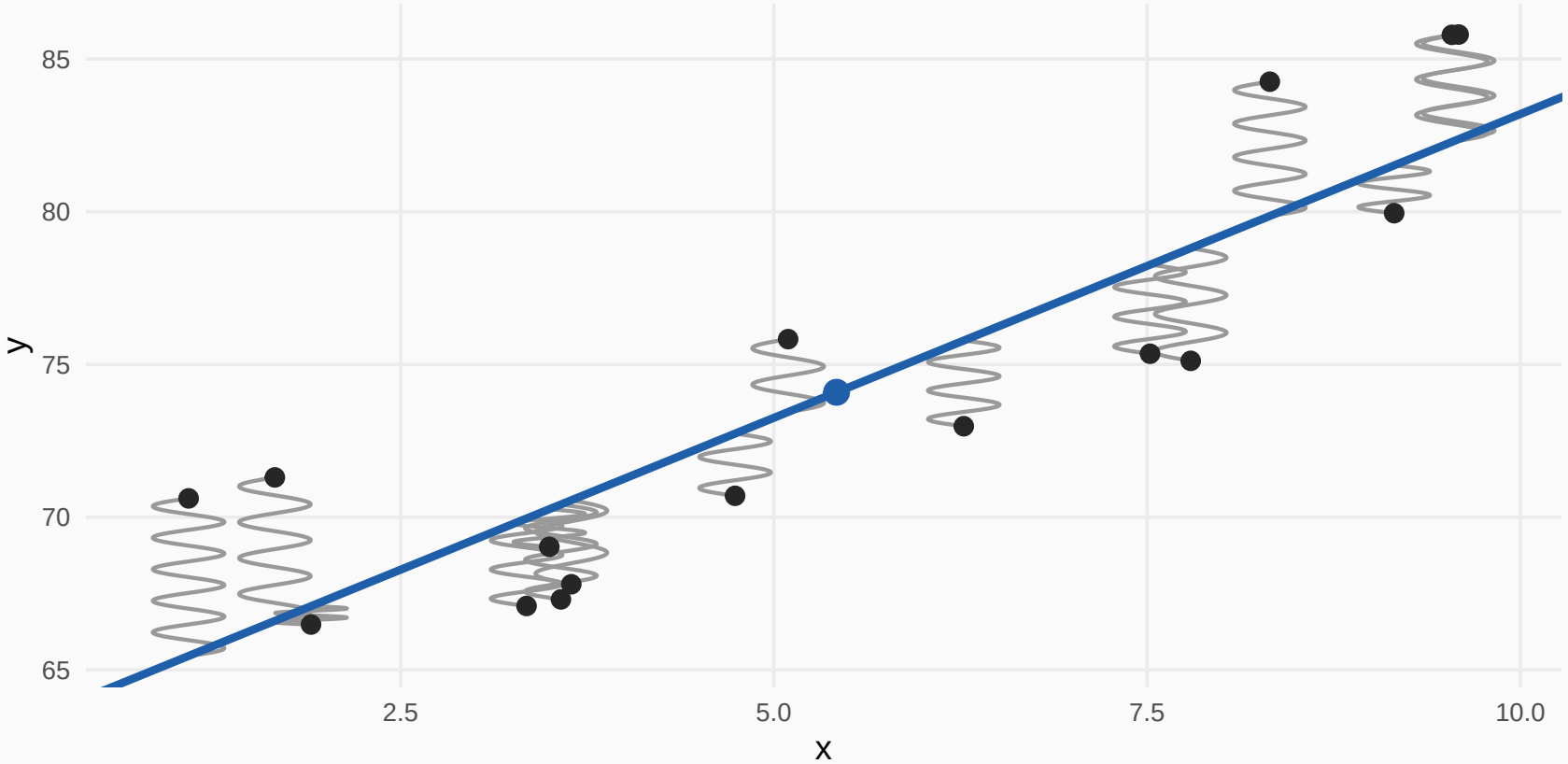
Goal: Find the best estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$ given the data.

- What does it mean, "best"?
- Least squares: best by criterion

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

- Find the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the criterion Q.
  1. Write the normal equations (derivatives of Q set to 0)
  2. Find the solution of the normal equations

Least squares, mechanically



Attach each observation to the line by a spring. A spring stretched by e_i stores energy proportional to e_i^2, so the total energy is $\sum_i e_i^2 = Q(\beta_0, \beta_1)$.

The least squares line is where the springs come to rest.

Least Square Estimators

Minimise $Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ over β_0 and β_1.

Normal equations. Set both partial derivatives to zero:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0, \qquad \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0.$$

Divide the first by -2 and expand: $\sum_i y_i = n\beta_0 + \beta_1 \sum_i x_i$, so

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

The fitted line always passes through the point $(\bar{x}, \bar{Y})$.

Least Square Estimators

Substitute $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ into the second normal equation:

$$\sum_{i=1}^{n} x_i \big(y_i - \bar{Y} - \hat{\beta}_1 (x_i - \bar{x})\big) = 0.$$

Solving for $\hat{\beta}_1$,

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xY}}{S_{xx}}}$$

where $S_{xY} = \sum_i (x_i - \bar{x})(Y_i - \bar{Y})$ and $S_{xx} = \sum_i (x_i - \bar{x})^2$.

Q is a positive quadratic in (β_0, β_1), so this stationary point is the minimum.

Sums of squares, and where r comes in

Notation

For a sample $(x_1, Y_1), \ldots, (x_n, Y_n)$,

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad S_{xY} = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}), \qquad S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

These are sums, not averages. The empirical variances, covariance and correlation are

$$s_x^2 = \frac{S_{xx}}{n-1}, \quad s_{xY} = \frac{S_{xY}}{n-1}, \quad s_Y^2 = \frac{S_{YY}}{n-1}, \quad r = \frac{s_{xY}}{s_x s_Y} = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}}.$$

The factor $n - 1$ cancels in the slope, which is why both forms are seen:

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \frac{s_{xY}}{s_x^2} = r \, \frac{s_Y}{s_x} = r \sqrt{\frac{S_{YY}}{S_{xx}}}.$$

$\hat{\beta}_1$ is linear in the Y_i

The x_i are fixed constants, and $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. So the $\bar{Y}$ in S_{xY} contributes nothing:

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(x_i - \bar{x})Y_i - \bar{Y}\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{=0} = \sum_{i=1}^{n}(x_i - \bar{x})Y_i.$$

Dividing by S_{xx},

$$\boxed{\hat{\beta}_1 = \sum_{i=1}^{n} k_i Y_i, \qquad k_i = \frac{x_i - \bar{x}}{S_{xx}}}$$

A weighted sum of the Y_i, with weights fixed by the design.

Least Square Estimators

Other criteria. Why square the errors?

We could use least absolute deviations estimates, minimizing

$$Q_1(\beta_0, \beta_1) = \sum_{i=1}^{n} |(y_i - \beta_0 - \beta_1 x_i)|$$

- **Convenience.** Q is differentiable everywhere, so the normal equations have a closed-form solution. Q_1 is not differentiable at zero and needs an algorithm (linear programming).
- **Optimality.** Under the assumptions of the next slide, least squares is the best estimator of a whole class.

Gauss-Markov Theorem

Theorem 1 (Gauss–Markov)

Consider the simple linear model

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

Suppose that the following assumptions (called Gauss-Markov assumptions) concerning the random errors are satisfied:

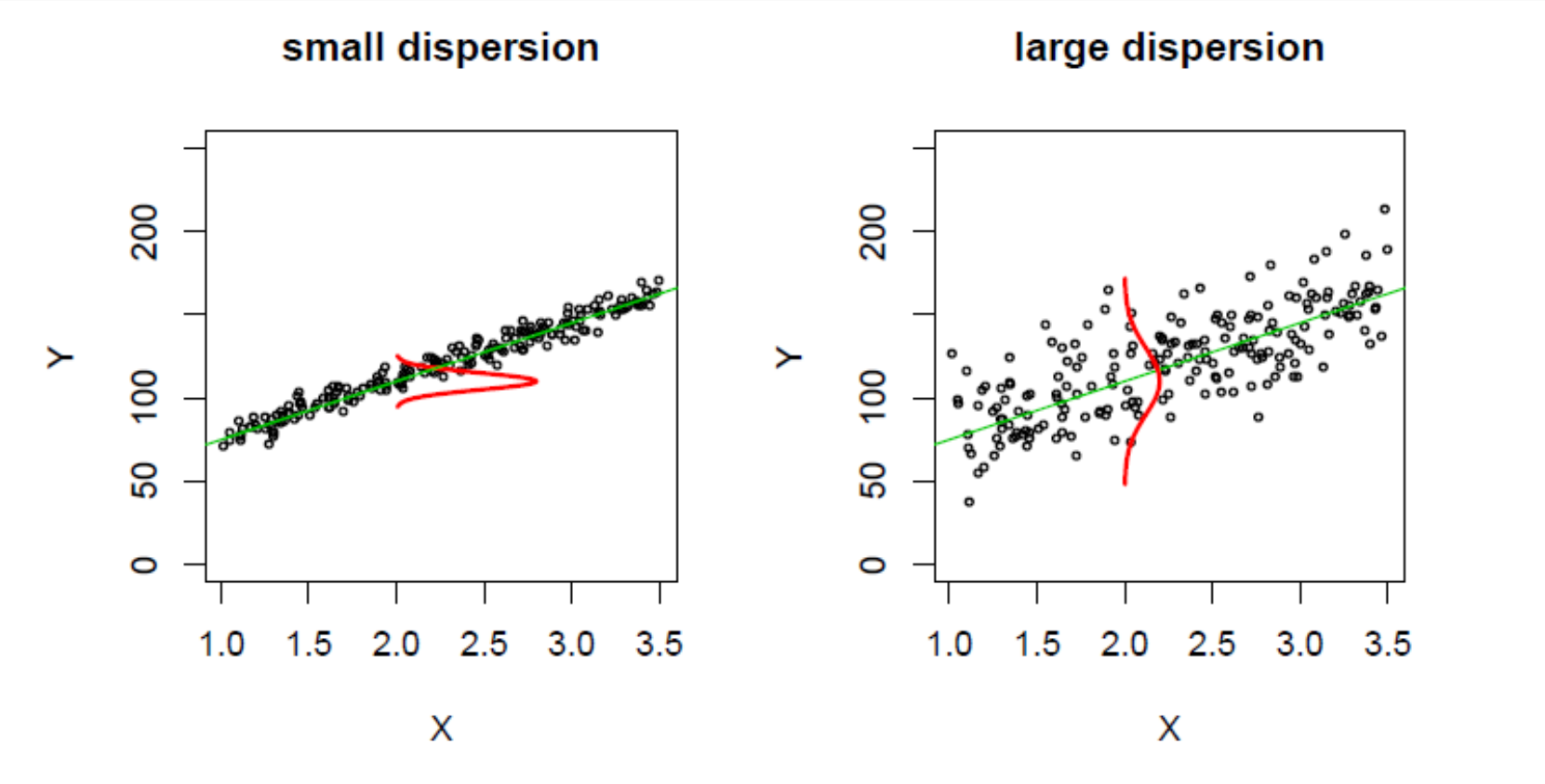
- Mean zero: $E(\epsilon_i) = 0$
- Constant variance: $Var(\epsilon_i) = \sigma^2$
- Uncorrelated: $Cov(\epsilon_i, \epsilon_j) = 0, \;\; i \neq j$

Then the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all unbiased linear estimators (BLUE).

The conclusion holds for **both** estimators separately, and in fact for any linear combination $c_0\beta_0 + c_1\beta_1$: for instance the fitted mean $\hat{\beta}_0 + \hat{\beta}_1 x_0$ is BLUE for $E(Y \mid X = x_0)$. Note that normality is **not** assumed.

The interpretation of σ^2

The variance, σ^2 controls the dispersion of Y around $\beta_0 + \beta_1 X$



Fitted values and Residuals

Definition

Regression equation or fitted regression line

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X,$$

where $\hat{Y}$ is the estimated mean of the response variable at level X of the explanatory.

- For each observation, we can compute the fitted value:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n$$

- The vertical distance from the observed y_i to the fitted value is called the residual

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i), \quad i = 1, \ldots, n$$

The residuals can be thought of as predicted values of the unknown errors $\epsilon_1, \ldots, \epsilon_n$.