

STAC67: Regression Analysis

Lecture 4

Thibault Randrianarisoa

September 18, 2026

University of Toronto Scarborough

Least Square Estimators

Other criteria. Why square the errors?

We could use least absolute deviations estimates, minimizing

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

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).

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.

↳ linear estimators

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: $Cov(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$

} no Gaussian assumption

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.

Fitted values and Residuals

Definition

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

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**:

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

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

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

The residuals can be thought of as predicted values of the unknown errors

$$\underbrace{\epsilon_1, \dots, \epsilon_n}_{\text{unknown errors}}$$

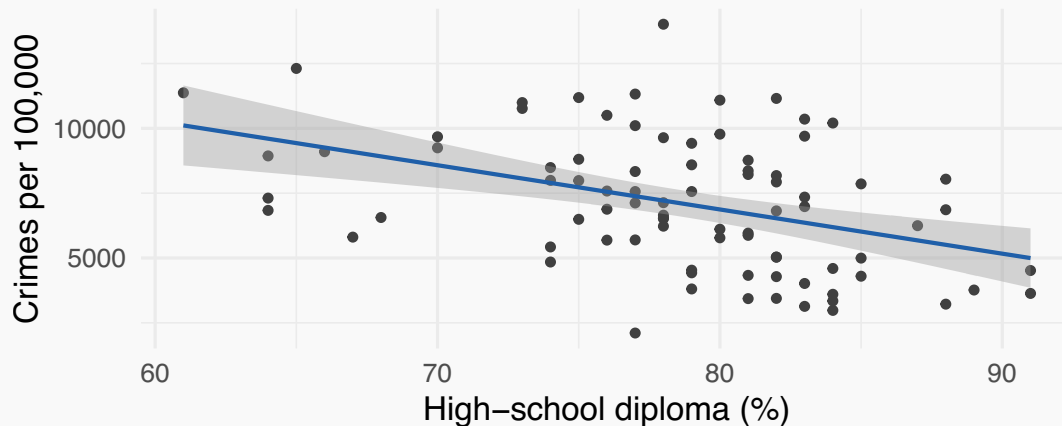
Example: Crime rate

A criminologists studying the relationship between level of education and crime rate in medium-sized U.S. counties collected from a random sample of 84 counties; X is the percentage of individuals having at least a high-school diploma, and Y is the crime rate (crimes reported per 100,000 residents) last year.

```
library(ggplot2)
```

```
Crime <- read.table("CrimeRate.txt")  
names(Crime) <- c("Y", "X")
```

```
ggplot(Crime, aes(X, Y)) +  
  geom_point(size = 0.9, colour = "grey25") +  
  geom_smooth(method = "lm", colour = "#1F5FA9", linewidth = 0.6) +  
  labs(x = "High-school diploma (%)", y = "Crimes per 100,000") +  
  theme_minimal(base_size = 10)
```



Example (Continued)

$$Y \sim X \text{ means } Y = \beta_0 + \beta_1 X + \epsilon$$

```
attach(Crime)
fit = lm(Y~X, data=Crime)
coefficients(fit)
```

```
## (Intercept)          X
## 20517.5999    -170.5752
```

```
c(mean(Y), mean(X))
```

```
## [1] 7111.20238    78.59524
c(sum((Y-mean(Y))*(X-mean(X))), sum((X-mean(X))^2))
```

```
## [1] -547928.119    3212.238
```

```
c(Y[10], X[10])
```

```
## [1] 7932    82
```

```
Crime[10, ]
```

```
##      Y  X
```

```
## 10 7932 82
```

```
fit$residuals[10]
```

```
##      10
```

```
## 1401.566
```

if $\text{lm}(Y \sim \text{edu}, \text{Crime})$
edu
-170,57

Exercise

1. Let's compute the estimated regression coefficients by hands:

$$\hat{b}_0:$$

$$\hat{b}_1: \frac{S_{xy}}{S_{xx}} = \frac{-547928}{3212} = -170,46$$

2. Obtain the point estimates of the following:

- the difference in the mean crime rate for two counties whose high-school graduation rates differ by one percentage points.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X; \hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1 (X+1) \Rightarrow \hat{Y}_2 - \hat{Y} = \hat{\beta}_1$$

- the mean crime rate last year in counties with high school graduation percentage $X = 80$.

$$\hat{Y}_{80} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 80 =$$

• ϵ_{10}

$$(X_{10}, Y_{10}) \quad e_{10} = Y_{10} - \hat{Y}_{10} = 7932 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 82)$$

Properties of the fitted regression line

- The least squares line always **passes through the point** $(\bar{X}, \bar{Y})$.
- The estimated slope $\hat{\beta}_1$ always has the same sign as the sample correlation between X and Y , $\widehat{Cov}(X, Y)$.
- The sum of residuals are 0.

(i) $\sum e_i = 0$

$$\sum e_i = \sum (Y_i - \hat{Y}_i) = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

(by the 1st normal equation)

(ii) $\sum e_i X_i = 0$

$$\sum X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

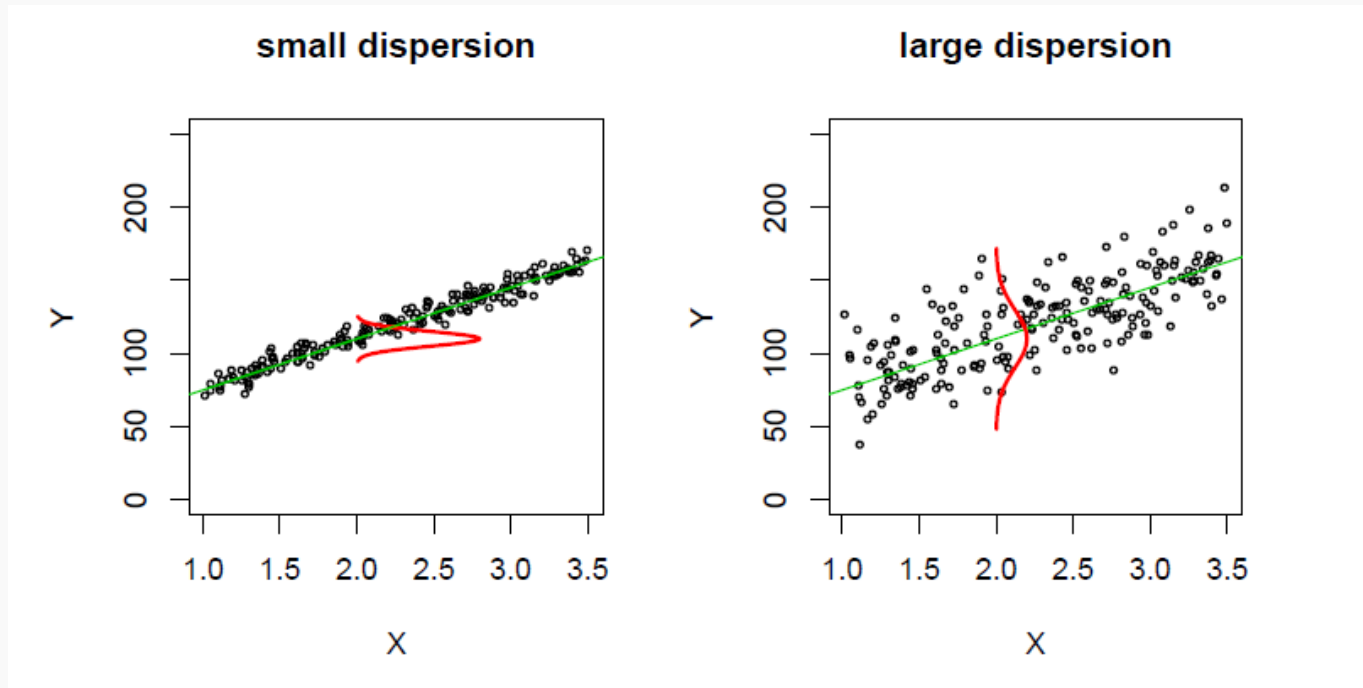
(by the 2nd normal equation)

- The sum of squares of the e_i 's is called: **Residual sum of Squares (RSS)** or **Sum of Squared Errors (SSE)**.

$$\sum_i (Y_i - \hat{Y}_i)^2 \approx \sum_i \varepsilon_i^2 \approx n\sigma^2$$

The interpretation of σ^2

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



Properties of the fitted regression line

- An unbiased estimator of σ^2 is: $\hat{\sigma}^2 = \frac{SSE}{(n-2)} = \frac{\sum_i e_i^2}{(n-2)}$
- Result:

$$\textcircled{1} \sum_i (x_i - \bar{x}) = 0 \qquad = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

$$\textcircled{2} \sum_i y_i = \sum_i \hat{y}_i \Leftrightarrow \sum_i \underbrace{(y_i - \hat{y}_i)}_{=e_i} = 0$$

$$\textcircled{3} \sum_i e_i \hat{y}_i = \underbrace{\sum_i e_i}_{=0} \hat{\beta}_0 + \underbrace{\sum_i e_i x_i}_{=0} \hat{\beta}_1 = 0$$

Properties of Least Squares Estimates

Since $\sum(X_i - \bar{X}) = 0$, the numerator of $\hat{\beta}_1$ simplifies:

$$\sum_{xy} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i - \bar{X})Y_i - \bar{Y} \underbrace{\sum(X_i - \bar{X})}_{=0} = \sum(X_i - \bar{X})Y_i$$

so, with $S_{xx} = \sum(X_i - \bar{X})^2$,

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})Y_i}{S_{xx}} = \sum_{i=1}^n k_i Y_i, \quad \text{where } k_i = \frac{X_i - \bar{X}}{S_{xx}}.$$

The constants k_i have several interesting properties:

1. $\sum_{i=1}^n k_i = \frac{\sum(X_i - \bar{X})}{S_{xx}} = 0$
2. $\sum_{i=1}^n k_i X_i = \frac{\sum(X_i - \bar{X})X_i}{S_{xx}} = \frac{\sum(X_i - \bar{X})(X_i - \bar{X})}{S_{xx}} = \frac{S_{xx}}{S_{xx}} = 1$
3. $\sum_{i=1}^n k_i^2 = \frac{\sum(X_i - \bar{X})^2}{S_{xx}^2} = \frac{S_{xx}}{S_{xx}^2} = \frac{1}{S_{xx}}$

In 2, $\sum(X_i - \bar{X})X_i = \sum(X_i - \bar{X})(X_i - \bar{X})$ because $\sum(X_i - \bar{X})\bar{X} = 0$.

Chapter 2: Inferences in Regression

Sampling Distribution of Least Square Estimators

Expected value of LSE

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$1. E(\hat{\beta}_1) = E\left[\sum_{i=1}^n h_i Y_i\right] = \sum_{i=1}^n h_i E[Y_i]$$

$$= \sum_{i=1}^n h_i (\beta_0 + \beta_1 X_i)$$

$$= \beta_0 \underbrace{\sum_i h_i}_{=0} + \beta_1 \underbrace{\sum_{i=1}^n h_i X_i}_{=1} = \beta_1$$

$$2. E(\hat{\beta}_0) = E[\bar{Y} - \hat{\beta}_1 \bar{X}]$$

$$= E[\bar{Y}] - E[\hat{\beta}_1] \bar{X}$$

$$= \frac{1}{n} \sum_{i=1}^n E[Y_i] - \beta_1 \frac{1}{n} \sum_{i=1}^n X_i = \beta_0 + \cancel{\beta_1 \frac{1}{n} \sum_{i=1}^n X_i} - \beta_1 \frac{1}{n} \sum_{i=1}^n X_i$$

$$3. E(\hat{\sigma}^2) = \sigma^2 \quad \text{where } \sigma^2 = \beta_0 + \beta_1 X_i$$

$$\frac{SSE}{n-2}$$

$$= \beta_0$$

Sampling Distribution of Least Square Estimators

Variance of least square estimators

$$\begin{aligned} 1. \text{Var}(\hat{\beta}_1) &= \text{Var}\left(\sum_{i=1}^m \beta_i y_i\right) \\ &= \sum_{i=1}^m \beta_i^2 \text{Var}(y_i) + \sum_{i=1}^m \sum_{j \neq i} \beta_i \beta_j \underbrace{\text{Cov}(y_i, y_j)}_{=0} \\ &= \sum_{i=1}^m \beta_i^2 \sigma^2 = \sigma^2 / S_{xx} \end{aligned}$$

$\underbrace{\sum_{i=1}^m \beta_i^2}_{=1/S_{xx}} \sigma^2 = \sigma^2 / S_{xx}$

$$\begin{aligned} 2. \text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= \underbrace{\text{Var}(\bar{y})}_{= \frac{1}{m^2} \sum_{i,j} \text{Cov}(y_i, y_j)} + \bar{x}^2 \text{Var}(\hat{\beta}_1) - 2 \underbrace{\text{Cov}(\bar{y}, \hat{\beta}_1)}_{=0} \\ &= \frac{1}{m^2} \sum_{i,j} \text{Cov}(y_i, y_j) = \sigma^2 \left(\frac{1}{m} + \frac{\bar{x}^2}{S_{xx}} \right) \\ &= \sigma^2 / m \end{aligned}$$

Sampling Distribution of Least Square Estimators

Theorem 2

1.

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$$

2.

$$Var(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} \right) \sigma^2, \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

The variances in Theorem 2, depends on the unknown value of σ^2 . If we replace σ^2 with the unbiased estimator $\hat{\sigma}^2$ and take the square root, we get the standard error of the estimators:

Standard error $\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{SSE}{(n-2)S_{xx}}} = SE(\hat{\beta}_1)$

$$SE(\hat{\beta}_0) = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \frac{SSE}{n-2}}$$

Inference about Regression Parameters

Until now, we did not use any property of the noise ϵ_i beyond the first and second moments.

Let's now focus on the simple Linear Model with **Normal Errors**

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

where

- β_0 and β_1 are parameters
- ϵ_i are i.i.d with **normal distribution** with mean 0 and variance, σ^2 .

$$\epsilon_i \sim N(0, \sigma^2)$$

- Under this model,

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Inference about Regression Parameters

- Inference about β_1
- σ is unknown, and must be estimated.

$$T = \frac{\hat{\beta}_1 - \beta_1}{\left(\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \right)} \sim t(n - 2)$$

- Review: the **Student's t-distribution with k degrees of freedom** can be defined as the random variable T

$$T = \frac{Z}{\sqrt{V/k}} \sim t_{(k)},$$

where

Sampling distribution of $\hat{\beta}_1$

- How to prove

$$\frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n - 2)?$$

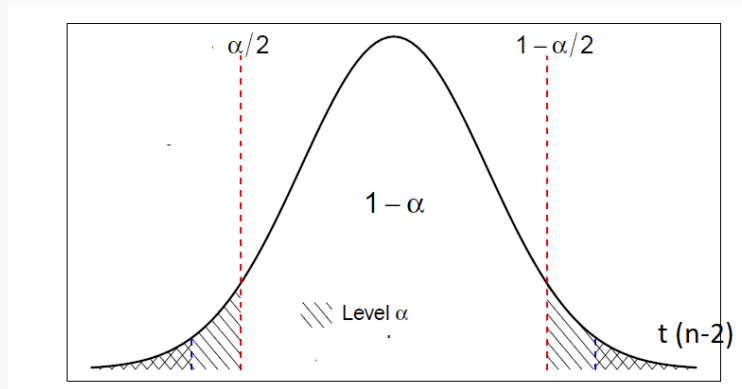
Inference about Regression Parameters

Theorem 3: Sampling distributions

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim t(n-2), \quad \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t(n-2)$$

Confidence Interval for β_0 and β_1

Since $\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t(n-2)$, $j = 0, 1$



Confidence Intervals

$$\hat{\beta}_j \pm t(1 - \alpha/2; n - 2)SE(\hat{\beta}_j), j = 0, 1$$

Why should we care about confidence intervals?

The confidence interval **completely** captures the information in the data about the parameter.

