Linear Regression with One Predictor Variable

Zhenisbek Assylbekov

Department of Mathematics

Regression Analysis

Introduction

Simple regression is about modeling one variable as a function of another variable

$$y = f(x)$$

given some data $(x_1, y_1), \ldots, (x_n, y_n)$.

Introduction

Simple regression is about modeling one variable as a function of another variable

$$y = f(x)$$

given some data $(x_1, y_1), \ldots, (x_n, y_n)$.

Usually $f(x)$ is parameterized:

$$y = f(x; \theta)$$

Introduction

Simple regression is about modeling one variable as a function of another variable

$$y = f(x)$$

given some data $(x_1, y_1), \ldots, (x_n, y_n)$.

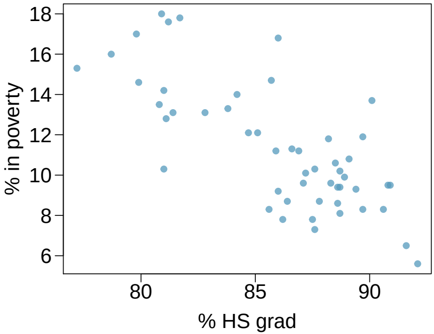
Usually $f(x)$ is parameterized:

$$y = f(x; \theta)$$

The goal is to tweak θ so that $y = f(x; \theta)$ fits the data in the best possible way.

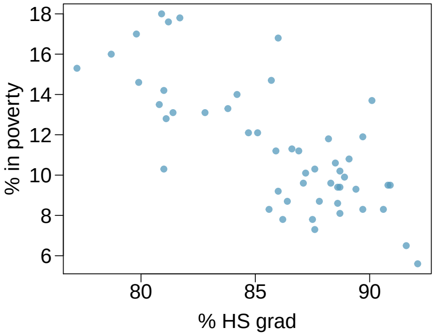
Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Poverty vs. HS graduate rate

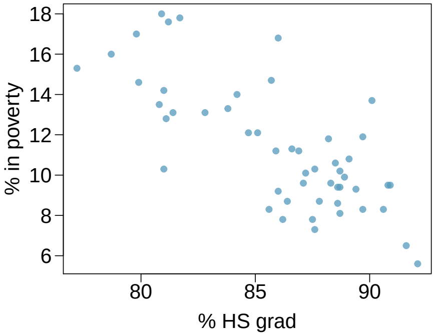
The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable (y)

Poverty vs. HS graduate rate

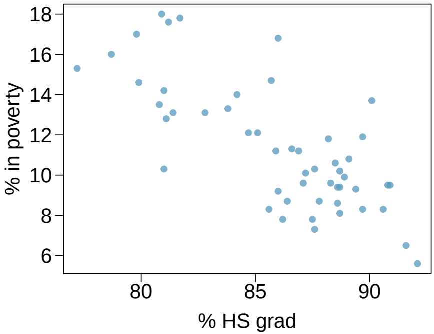
The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable (y)
% in poverty

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).

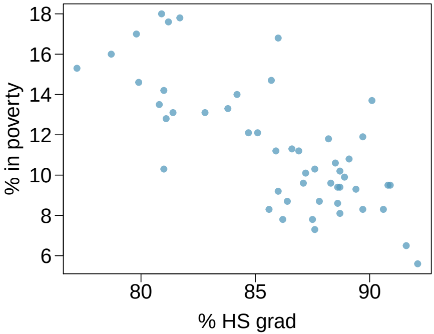


Response variable (y)
% in poverty

Explanatory variable (x)

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).

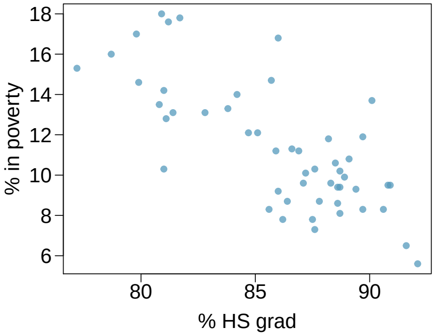


Response variable (y)
% in poverty

Explanatory variable (x)
% HS grad

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable (y)

% in poverty

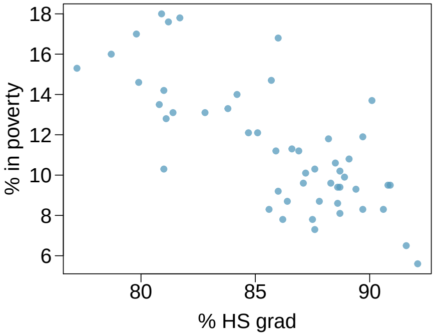
Explanatory variable (x)

% HS grad

Relationship

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable (y)
% in poverty

Explanatory variable (x)
% HS grad

Relationship
linear, negative, moderately strong

Simple linear regression model

- The simplest relationship between two variables x and y is linear:
$$y = \beta_0 + \beta_1 x$$

Simple linear regression model

- The simplest relationship between two variables x and y is linear:
$$y = \beta_0 + \beta_1 x$$
- To allow variability around the line we add random noise:
$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministic}} + \epsilon_i, \quad i = 1, \ldots, n,$$

  where

Simple linear regression model

- The simplest relationship between two variables x and y is linear:
$$y = \beta_0 + \beta_1 x$$
- To allow variability around the line we add random noise:
$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministic}} + \epsilon_i, \quad i = 1, \ldots, n,$$

  where
  - Y_i is the value of the **response variable** in the i^{th} example,

Simple linear regression model

- The simplest relationship between two variables x and y is linear:

$$y = \beta_0 + \beta_1 x$$

- To allow variability around the line we add random noise:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministic}} + \epsilon_i, \quad i = 1, \ldots, n,$$

  where

  - Y_i is the value of the **response variable** in the i^{th} example,
  - x_i is the value of the **predictor variable** in the i^{th} example.

Simple linear regression model

- The simplest relationship between two variables x and y is linear:

$$y = \beta_0 + \beta_1 x$$

- To allow variability around the line we add random noise:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministic}} + \epsilon_i, \quad i = 1, \ldots, n,$$

where

  - Y_i is the value of the **response variable** in the i^{th} example,
  - x_i is the value of the **predictor variable** in the i^{th} example.

Simple linear regression model

- The simplest relationship between two variables x and y is linear:

$$y = \beta_0 + \beta_1 x$$

- To allow variability around the line we add random noise:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministic}} + \epsilon_i, \quad i = 1, \ldots, n,$$

  where
  - Y_i is the value of the **response variable** in the i^{th} example,
  - x_i is the value of the **predictor variable** in the i^{th} example.
- Assumptions on ϵ_i's:
  - $\mathrm{E}[\epsilon_i] = 0$

Simple linear regression model

- The simplest relationship between two variables x and y is linear:

$$y = \beta_0 + \beta_1 x$$

- To allow variability around the line we add random noise:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministic}} + \epsilon_i, \quad i = 1, \dots, n,$$

  where
  - Y_i is the value of the **response variable** in the i^{th} example,
  - x_i is the value of the **predictor variable** in the i^{th} example.
- Assumptions on ϵ_i's:
  - $\mathrm{E}[\epsilon_i] = 0$
  - $\mathrm{Var}[\epsilon_i] = \sigma^2$

Simple linear regression model

- The simplest relationship between two variables x and y is linear:
$$y = \beta_0 + \beta_1 x$$
- To allow variability around the line we add random noise:
$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministic}} + \epsilon_i, \quad i = 1, \ldots, n,$$
where
  - Y_i is the value of the **response variable** in the i^{th} example,
  - x_i is the value of the **predictor variable** in the i^{th} example.
- Assumptions on ϵ_i's:
  - $\mathrm{E}[\epsilon_i] = 0$
  - $\mathrm{Var}[\epsilon_i] = \sigma^2$
  - $\mathrm{Cov}[\epsilon_i, \epsilon_j] = 0$ for $i \neq j$

Mean and variance of each Y_i

Given the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
$$\mathrm{E}[\epsilon_i] = 0, \quad \mathrm{Var}[\epsilon_i] = \sigma^2, \quad \mathrm{Cov}[\epsilon_i, \epsilon_j] = 0 \text{ for } i \neq j$$

show that

Mean and variance of each Y_i

Given the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
$$\mathrm{E}[\epsilon_i] = 0, \quad \mathrm{Var}[\epsilon_i] = \sigma^2, \quad \mathrm{Cov}[\epsilon_i, \epsilon_j] = 0 \text{ for } i \neq j$$

show that

- $\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i,$

Mean and variance of each Y_i

Given the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
$$\mathrm{E}[\epsilon_i] = 0, \quad \mathrm{Var}[\epsilon_i] = \sigma^2, \quad \mathrm{Cov}[\epsilon_i, \epsilon_j] = 0 \text{ for } i \neq j$$

show that

- $\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i$,
- $\mathrm{Var}[Y_i] = \sigma^2$,

Mean and variance of each Y_i

Given the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
$$\mathrm{E}[\epsilon_i] = 0, \quad \mathrm{Var}[\epsilon_i] = \sigma^2, \quad \mathrm{Cov}[\epsilon_i, \epsilon_j] = 0 \text{ for } i \neq j$$

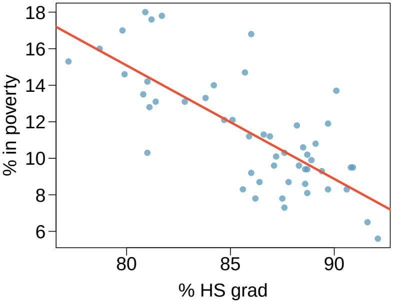
show that

- $\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i$,
- $\mathrm{Var}[Y_i] = \sigma^2$,
- $\mathrm{Cov}[Y_i, Y_j] = 0$.

Interpretation of β_0 and β_1

Suppose we somehow estimated β_0 and β_1 in the Poverty vs HS grads example:

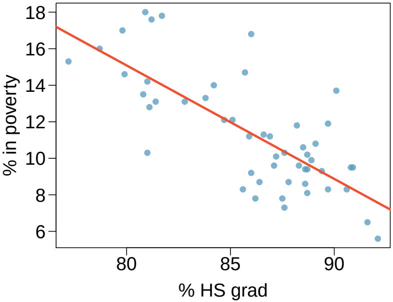
$$Y_i = 64.68 - 0.62 \cdot x_i + \epsilon_i$$



Interpretation of β_0 and β_1

Suppose we somehow estimated β_0 and β_1 in the Poverty vs HS grads example:

$$Y_i = 64.68 - 0.62 \cdot x_i + \epsilon_i$$

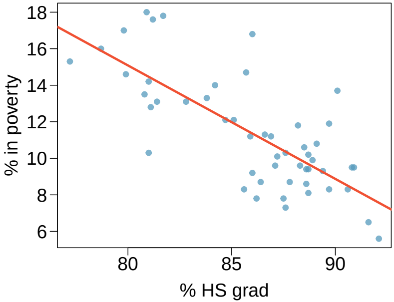


What does 64.68 represent here?

Interpretation of β_0 and β_1

Suppose we somehow estimated β_0 and β_1 in the Poverty vs HS grads example:

$$Y_i = 64.68 - 0.62 \cdot x_i + \epsilon_i$$

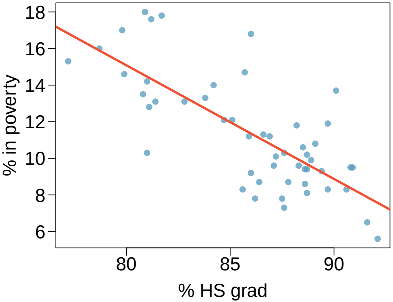


What does 64.68 represent here? Poverty rate for states with no high-school grads. Not sensible/interpretable.

Interpretation of β_0 and β_1

Suppose we somehow estimated β_0 and β_1 in the Poverty vs HS grads example:

$$Y_i = 64.68 - 0.62 \cdot x_i + \epsilon_i$$



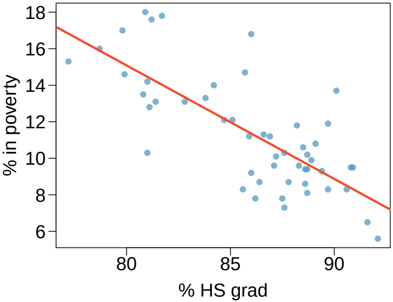
What does 64.68 represent here? Poverty rate for states with no high-school grads. Not sensible/interpretable.

How is 0.62 interpreted here?

Interpretation of β_0 and β_1

Suppose we somehow estimated β_0 and β_1 in the Poverty vs HS grads example:

$$Y_i = 64.68 - 0.62 \cdot x_i + \epsilon_i$$



What does 64.68 represent here? Poverty rate for states with no high-school grads. Not sensible/interpretable.

How is 0.62 interpreted here? Increasing % of high-school grads by 1% is associated with 0.62% decrease in poverty rate *on average*.

Simple linear regression using matrices

Note the simple linear regression model for all examples

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1, \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 x_n + \epsilon_n
\end{aligned}
$$

can be written in matrix terms as

$$
\underbrace{\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\boldsymbol{\epsilon}},
$$

or equivalently

$$
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}
$$

Remark on $\mathbf{X}^\top \mathbf{X}$

Notice that

$$\begin{aligned}\mathbf{X}^\top \mathbf{X} &= \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \\ &= \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}\end{aligned}$$

Convex function of one variable

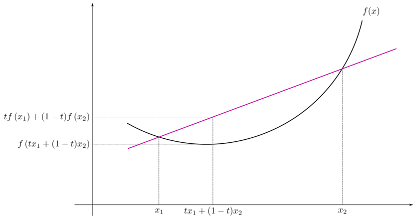
Definition. A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** (**concave up**) if

Convex function of one variable

Definition. A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** (**concave up**) if

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

for all $x_1, x_2 \in \mathrm{dom} f$ and all $t \in [0,1]$.

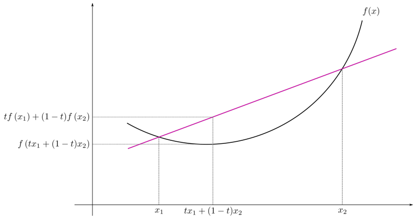


Convex function of one variable

Definition. A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** (**concave up**) if

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

for all $x_1, x_2 \in \mathrm{dom} f$ and all $t \in [0, 1]$.



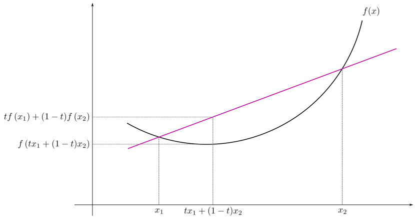
How do we usually check convexity for a twice-differentiable function?

Convex function of one variable

Definition. A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** (**concave up**) if

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

for all $x_1, x_2 \in \mathrm{dom} f$ and all $t \in [0,1]$.



How do we usually check convexity for a twice-differentiable function?

Theorem. If $f'' \geq 0$, then f is convex.

Hessian

The **Hessian** matrix of $f : \mathbb{R}^d \to \mathbb{R}$ is a matrix of second-order partial derivatives:

$$\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \qquad \text{i.e.} \qquad \mathbf{H}_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

Hessian

The **Hessian** matrix of $f : \mathbb{R}^d \to \mathbb{R}$ is a matrix of second-order partial derivatives:

$$\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \qquad \text{i.e.} \qquad \mathbf{H}_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

If the partial derivatives are continuous, the order of differentiation can be interchanged, so the Hessian matrix will be symmetric.

Convex function of multiple variables

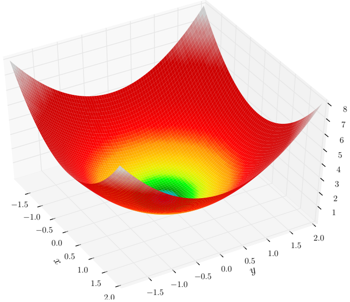
Assume $f : \mathbb{R}^d \to \mathbb{R}$. A function f is **convex** if

Convex function of multiple variables

Assume $f : \mathbb{R}^d \to \mathbb{R}$. A function f is **convex** if

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$ and all $t \in [0, 1]$.

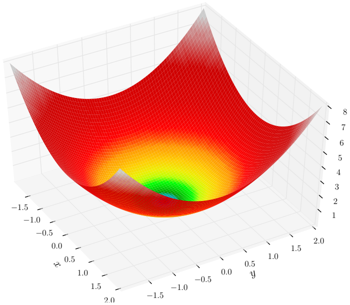


Convex function of multiple variables

Assume $f : \mathbb{R}^d \to \mathbb{R}$. A function f is **convex** if

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$ and all $t \in [0, 1]$.



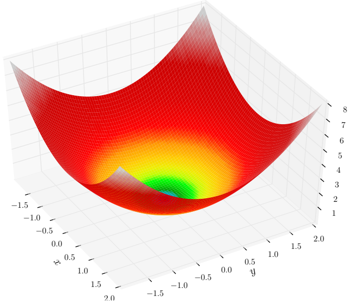
Theorem. If $\mathbf{H} \succeq \mathbf{0}$, then f is convex.

Convex function of multiple variables

Assume $f : \mathbb{R}^d \to \mathbb{R}$. A function f is **convex** if

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$ and all $t \in [0, 1]$.



Theorem. If $\mathbf{H} \succeq \mathbf{0}$, then f is convex.

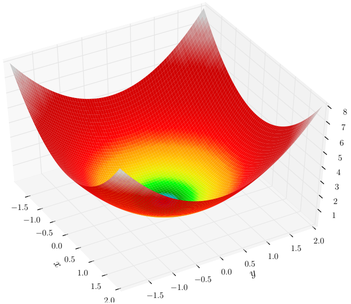
$\mathbf{H} \succeq \mathbf{0}$ denotes that $\mathbf{H}$ is a **positive semi-definite** matrix. What does this mean?

Convex function of multiple variables

Assume $f : \mathbb{R}^d \to \mathbb{R}$. A function f is **convex** if

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \\ \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$ and all $t \in [0, 1]$.



Theorem. If $\mathbf{H} \succeq \mathbf{0}$, then f is convex.

$\mathbf{H} \succeq \mathbf{0}$ denotes that $\mathbf{H}$ is a **positive semi-definite** matrix. What does this mean? $\mathbf{a}^\top \mathbf{H} \mathbf{a} \geq 0$ for any $\mathbf{a} \in \mathbb{R}^d$.

Why are we interested in convex functions?

Why are we interested in convex functions?

Convex functions do not have saddle points or local minima.

Why are we interested in convex functions?

Convex functions do not have saddle points or local minima.

Theorem. If f is convex, then any local minimum of f is also a *global* minimum.

Why are we interested in convex functions?

Convex functions do not have saddle points or local minima.

Theorem. If f is convex, then any local minimum of f is also a *global* minimum.

$\Rightarrow$ We can find any stationary point and guarantee that it is the global minimum.

Why are we interested in convex functions?

Convex functions do not have saddle points or local minima.

Theorem. If f is convex, then any local minimum of f is also a *global* minimum.

$\Rightarrow$ We can find any stationary point and guarantee that it is the global minimum.

What is a stationary point? A point $\mathbf{x}_0$ is called **stationary** if $\nabla f(\mathbf{x}_0) = 0$.

Model

Convexity

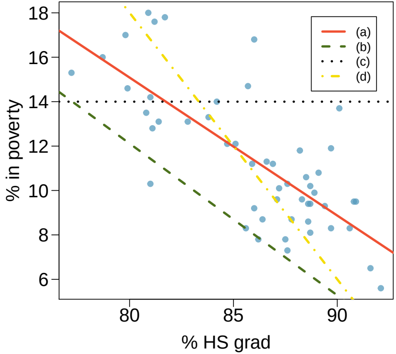
Parameter Estimation

Least Squares Estimation (LSE)

Maximum Likelihood Estimation (MLE)

Eyeballing the line

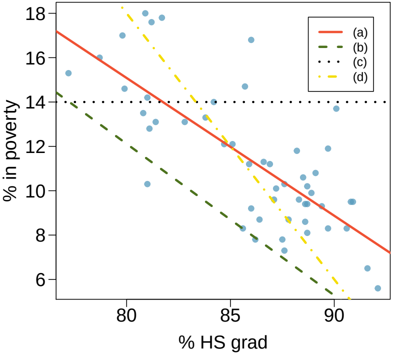
Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad?



Eyeballing the line

Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad?

(a)

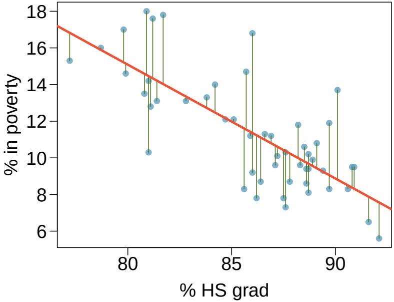


Residuals

Errors are simply $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$,

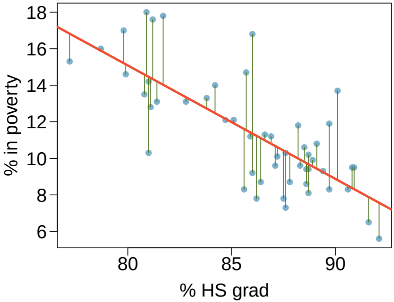
Residuals

Errors are simply $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$,
Residuals are *estimated* errors: $e_i = Y_i - (b_0 + b_1 x_i)$ once β_0 and β_1 have been replaced by their estimates b_0 and b_1.



Residuals

Errors are simply $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$,
Residuals are *estimated* errors: $e_i = Y_i - (b_0 + b_1 x_i)$ once β_0 and β_1 have been replaced by their estimates b_0 and b_1.

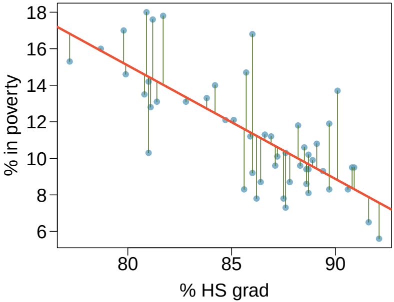


We want a line that has smallest possible residuals:

$$\min_{b_0, b_1} \sum_{i=1}^{n} e_i^2$$

Residuals

Errors are simply $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$,
Residuals are *estimated* errors: $e_i = Y_i - (b_0 + b_1 x_i)$ once β_0 and β_1 have been replaced by their estimates b_0 and b_1.



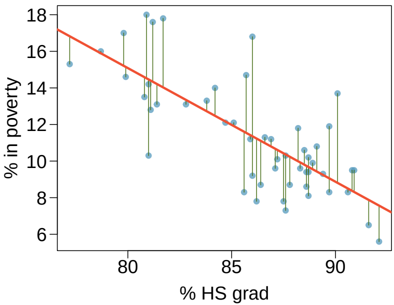
We want a line that has smallest possible residuals:

$$\min_{b_0, b_1} \sum_{i=1}^{n} e_i^2$$

This is called **least squares estimation (LSE)**.

Residuals

Errors are simply $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$,
Residuals are *estimated* errors: $e_i = Y_i - (b_0 + b_1 x_i)$ once β_0 and β_1 have been replaced by their estimates b_0 and b_1.



We want a line that has smallest possible residuals:

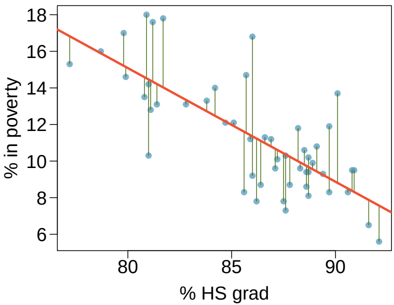
$$\min_{b_0, b_1} \sum_{i=1}^{n} e_i^2$$

This is called **least squares estimation (LSE)**.

Why don't we minimize the sum $\sum_{i=1}^{n} e_i$ instead?

Residuals

Errors are simply $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$,
Residuals are *estimated* errors: $e_i = Y_i - (b_0 + b_1 x_i)$ once β_0 and β_1 have been replaced by their estimates b_0 and b_1.



We want a line that has smallest possible residuals:

$$\min_{b_0, b_1} \sum_{i=1}^{n} e_i^2$$

This is called **least squares estimation (LSE)**.

Why don't we minimize the sum $\sum_{i=1}^{n} e_i$ instead? Positive and negative residuals will compensate each other $\Rightarrow$ we won't be sure that the magnitudes $|e_i|$ are small.

Least squares estimation of β_0 and β_1

Theorem. The function $Q(\beta_0, \beta_1) = \sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)]^2$ has the global minimum at

$$b_1 := \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 := \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Least squares estimation of β_0 and β_1

Theorem. The function $Q(\beta_0, \beta_1) = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 x_i)]^2$ has the global minimum at

$$b_1 := \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$b_0 := \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Proof. First derivatives of Q:

$$\frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^n 2(Y_i - \beta_0 - \beta_1 x_i)(-x_i) = -2 \left[\sum_{i=1}^n x_i Y_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 \right]$$

Least squares estimation of β_0 and β_1

Theorem. The function $Q(\beta_0, \beta_1) = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 x_i)]^2$ has the global minimum at

$$b_1 := \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b_0 := \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Proof. First derivatives of Q:

$$\frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^n 2(Y_i - \beta_0 - \beta_1 x_i)(-x_i) = -2\left[\sum_{i=1}^n x_i Y_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2\right]$$

$$\frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^n 2(Y_i - \beta_0 - \beta_1 x_i)(-1) = -2\left[\sum_{i=1}^n Y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i\right].$$

Two equations in two unknowns

Setting these equal to zero, we have

$$\sum x_i Y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2 \tag{1}$$

$$\sum Y_i = n\beta_0 + \beta_1 \sum x_i \tag{2}$$

Two equations in two unknowns

Setting these equal to zero, we have

$$\sum x_i Y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2 \tag{1}$$

$$\sum Y_i = n\beta_0 + \beta_1 \sum x_i \tag{2}$$

Multiply (1) by n and multiply (2) by $\sum x_i$ and subtract yielding

$$n \sum x_i Y_i - \sum x_i \sum Y_i = \beta_1 \left[n \sum x_i^2 - \left(\sum x_i \right)^2 \right].$$

Two equations in two unknowns

Setting these equal to zero, we have

$$\sum x_i Y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2 \tag{1}$$

$$\sum Y_i = n\beta_0 + \beta_1 \sum x_i \tag{2}$$

Multiply (1) by n and multiply (2) by $\sum x_i$ and subtract yielding

$$n \sum x_i Y_i - \sum x_i \sum Y_i = \beta_1 \left[n \sum x_i^2 - \left(\sum x_i \right)^2 \right].$$

Solving for β_1 we get

$$\hat{\beta}_1 = \frac{n \sum x_i Y_i - \sum x_i \sum Y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i Y_i - n\bar{Y}\bar{x}}{\sum x_i^2 - n\bar{x}^2}.$$

Plugging this into (2), we have $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$.

Convexity of Q

How can we show that $Q(\beta_0, \beta_1)$ is convex?

Convexity of Q

How can we show that $Q(\beta_0, \beta_1)$ is convex? Its Hessian should be positive semi-definite.

Convexity of Q

How can we show that $Q(\beta_0, \beta_1)$ is convex? Its Hessian should be positive semi-definite.

Second-order partial derivatives of Q:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \quad \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2, \quad \frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} = 2\sum x_i$$

Convexity of Q

How can we show that $Q(\beta_0, \beta_1)$ is convex? Its Hessian should be positive semi-definite.

Second-order partial derivatives of Q:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \quad \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2, \quad \frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} = 2\sum x_i$$

Hessian of Q:

$$\mathbf{H} = 2 \cdot \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = 2\mathbf{X}^\top \mathbf{X}$$

Convexity of Q

How can we show that $Q(\beta_0, \beta_1)$ is convex? Its Hessian should be positive semi-definite.

Second-order partial derivatives of Q:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \quad \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2, \quad \frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} = 2\sum x_i$$

Hessian of Q:

$$\mathbf{H} = 2 \cdot \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = 2\mathbf{X}^\top \mathbf{X}$$

For an arbitrary $\mathbf{a} = (a_1, a_2)$ we have

$$\mathbf{a}^\top \mathbf{H} \mathbf{a} = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \cdot 2 \cdot \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \mathbf{a}^\top 2\mathbf{X}^\top \mathbf{X}\mathbf{a}$$
$$= 2(\mathbf{X}\mathbf{a})^\top(\mathbf{X}\mathbf{a}) = 2\|\mathbf{X}\mathbf{a}\|^2 \geq 0 \quad \Rightarrow \quad \mathbf{H} \succeq \mathbf{0}$$

Probabilistic Setup

Assume

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

Probabilistic Setup

Assume

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Probabilistic Setup

Assume

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

This is equivalent to

$$Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \qquad \text{(Why?)}$$

Probabilistic Setup

Assume

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

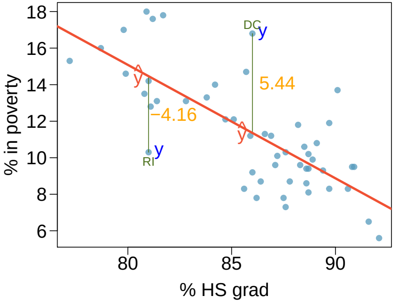
This is equivalent to

$$Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \qquad \text{(Why?)}$$

This means that p.d.f. of Y_i is

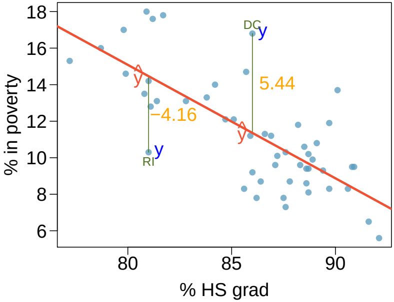
$$f_{Y_i}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[y - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \right)$$

Properties of the residuals



Denote $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Properties of the residuals



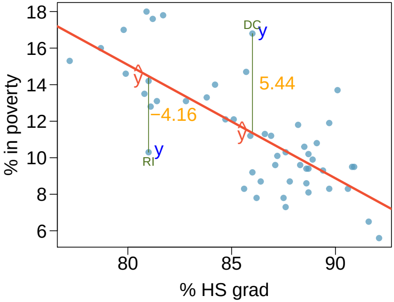
Denote $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Residuals are differences between observed and predicted responses:

$$e_i = Y_i - \hat{Y}_i$$

Exercise. Show that

Properties of the residuals



Denote $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

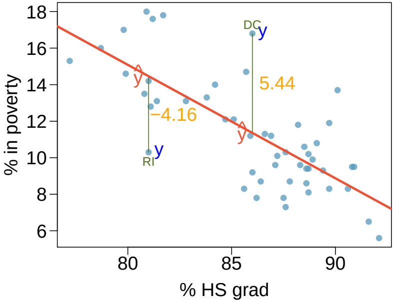
Residuals are differences between observed and predicted responses:

$$e_i = Y_i - \hat{Y}_i$$

Exercise. Show that

- $\sum_{i=1}^{n} e_i = 0$

Properties of the residuals



Denote $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

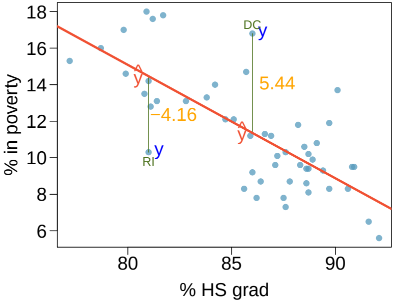
Residuals are differences between observed and predicted responses:

$$e_i = Y_i - \hat{Y}_i$$

Exercise. Show that

- $\sum_{i=1}^{n} e_i = 0$ (follows from (2)),

Properties of the residuals



Denote $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

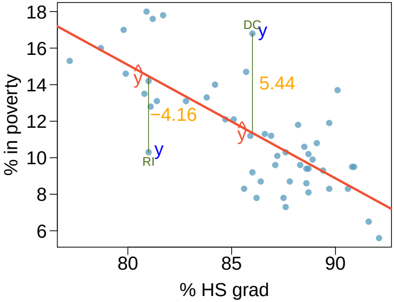
Residuals are differences between observed and predicted responses:

$$e_i = Y_i - \hat{Y}_i$$

Exercise. Show that

- $\sum_{i=1}^{n} e_i = 0$ (follows from (2)),
- $\sum_{i=1}^{n} x_i e_i = 0$ (follows from (1))

Properties of the residuals



Denote $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

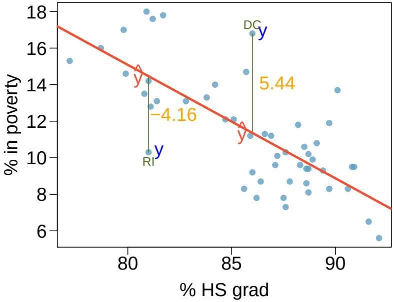
Residuals are differences between observed and predicted responses:

$$e_i = Y_i - \hat{Y}_i$$

Exercise. Show that

- $\sum_{i=1}^{n} e_i = 0$ (follows from (2)),
- $\sum_{i=1}^{n} x_i e_i = 0$ (follows from (1))
- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$ (from the previous two),

Properties of the residuals



Denote $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Residuals are differences between observed and predicted responses:

$$e_i = Y_i - \hat{Y}_i$$

Exercise. Show that

- $\sum_{i=1}^{n} e_i = 0$ (follows from (2)),
- $\sum_{i=1}^{n} x_i e_i = 0$ (follows from (1))
- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$ (from the previous two),
- Least squares line always goes through $(\bar{x}, \bar{Y})$.

Maximum Likelihood Estimation (MLE)

The *likelihood function* for the sample $Y_1, \ldots, Y_n$ given parameters β_0, β_1, σ is

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = \prod_i f_{Y_i}(Y_i) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[Y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2}\right)$$

Maximum Likelihood Estimation (MLE)

The *likelihood function* for the sample $Y_1, \ldots, Y_n$ given parameters β_0, β_1, σ is

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = \prod_i f_{Y_i}(Y_i) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[Y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2}\right)$$

Exercise. Find the MLE for β_0 and β_1:

$$(\hat{\beta}_{0,\text{MLE}}, \hat{\beta}_{1,\text{MLE}}) = \arg\max_{\beta_0, \beta_1} \mathcal{L}(\beta_0, \beta_1, \sigma^2)$$

Maximum Likelihood Estimation (MLE)

The *likelihood function* for the sample $Y_1, \ldots, Y_n$ given parameters β_0, β_1, σ is

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = \prod_i f_{Y_i}(Y_i) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[Y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2}\right)$$

Exercise. Find the MLE for β_0 and β_1:

$$(\hat{\beta}_{0,\text{MLE}}, \hat{\beta}_{1,\text{MLE}}) = \arg\max_{\beta_0,\beta_1} \mathcal{L}(\beta_0, \beta_1, \sigma^2)$$

Show that

$$\max_{\beta_0,\beta_1} \mathcal{L}(\beta_0, \beta_1, \sigma^2) \qquad \Leftrightarrow \qquad \min_{\beta_0,\beta_1} Q(\beta_0, \beta_1)$$

Maximum Likelihood Estimation (MLE)

The *likelihood function* for the sample $Y_1, \ldots, Y_n$ given parameters β_0, β_1, σ is

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = \prod_i f_{Y_i}(Y_i) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[Y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2}\right)$$

Exercise. Find the MLE for β_0 and β_1:

$$(\hat{\beta}_{0,\mathrm{MLE}}, \hat{\beta}_{1,\mathrm{MLE}}) = \arg\max_{\beta_0,\beta_1} \mathcal{L}(\beta_0, \beta_1, \sigma^2)$$

Show that

$$\max_{\beta_0,\beta_1} \mathcal{L}(\beta_0, \beta_1, \sigma^2) \qquad \Leftrightarrow \qquad \min_{\beta_0,\beta_1} Q(\beta_0, \beta_1)$$

$\Rightarrow$ MLE is a probabilistic justification for the LSE.

Estimating σ^2

Recall that $\sigma^2 = \mathrm{Var}[\epsilon_i]$.

Estimating σ^2

Recall that $\sigma^2 = \mathrm{Var}[\epsilon_i]$.

A natural estimator of σ^2 is

Estimating σ^2

Recall that $\sigma^2 = \mathrm{Var}[\epsilon_i]$.

A natural estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$

Estimating σ^2

Recall that $\sigma^2 = \mathrm{Var}[\epsilon_i]$.

A natural estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$ (why $\bar{e} = 0$?

Estimating σ^2

Recall that $\sigma^2 = \mathrm{Var}[\epsilon_i]$.

A natural estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$
(why $\bar{e} = 0$? because $\sum e_i = 0$)

Estimating σ^2

Recall that $\sigma^2 = \mathrm{Var}[\epsilon_i]$.

A natural estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$ (why $\bar{e} = 0$? because $\sum e_i = 0$)

In Chapter 5, we will show that $\frac{\sum e_i^2}{\sigma^2} \sim \chi^2_{n-2}$. This implies $\mathrm{E}[\hat{\sigma}^2] = \frac{n-2}{n}\sigma^2$.

Estimating σ^2

Recall that $\sigma^2 = \mathrm{Var}[\epsilon_i]$.

A natural estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (e_i - \bar{e})^2 = \frac{1}{n}\sum_{i=1}^n e_i^2$ (why $\bar{e} = 0$? because $\sum e_i = 0$)

In Chapter 5, we will show that $\frac{\sum e_i^2}{\sigma^2} \sim \chi^2_{n-2}$. This implies $\mathrm{E}[\hat{\sigma}^2] = \frac{n-2}{n}\sigma^2$. I.e., $\hat{\sigma}^2$ is a biased estimator of σ^2, and we prefer the unbiased version:

$$\frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n e_i^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - b_0 - b_1 x_i)^2 := \mathrm{MSE}.$$

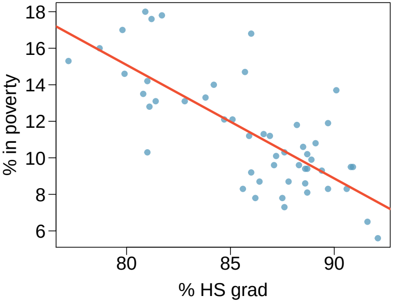
which is referred to as **mean squared error** (**MSE**).

Parameter estimation in R: Poverty vs HS grad rate

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp01/poverty.R

```
Coefficients:
             Estimate
(Intercept) 64.78097
Graduates   -0.62122

Residual standard error: 2.082
```

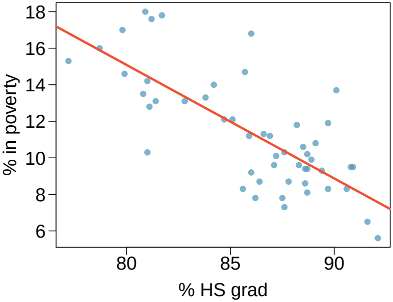


Parameter estimation in R: Poverty vs HS grad rate

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp01/poverty.R

```
Coefficients:
             Estimate
(Intercept) 64.78097
Graduates   -0.62122

Residual standard error: 2.082
```



The regression line is $y = \underbrace{64.78}_{b_0} - \underbrace{0.62}_{b_1} x$.

$\sqrt{\mathrm{MSE}} = 2.08$ is an estimate of σ.