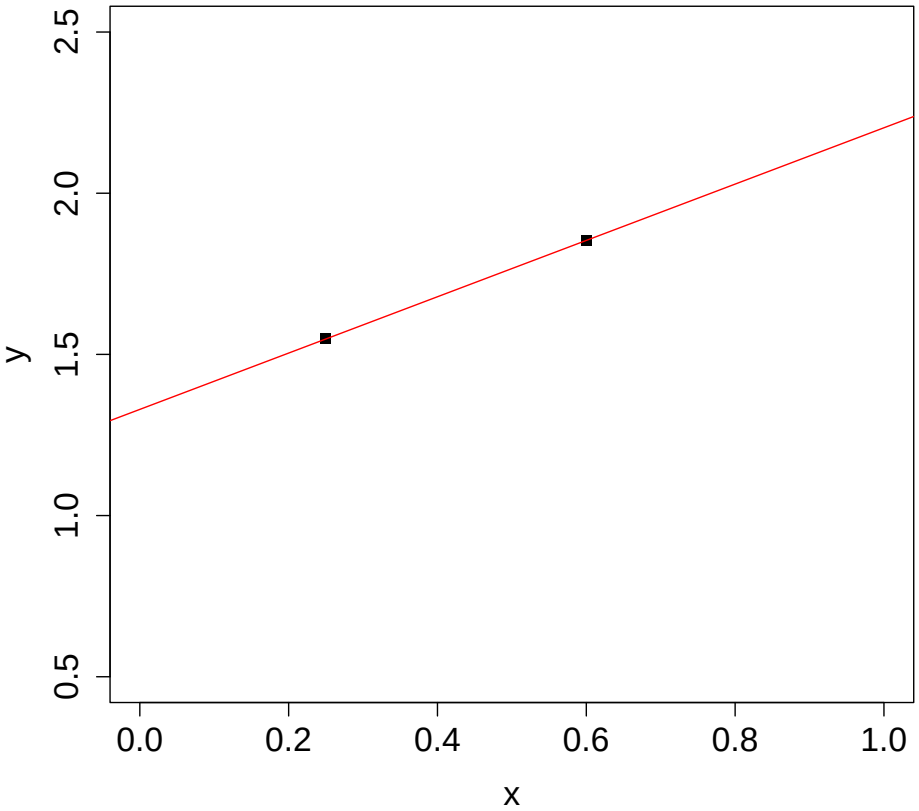
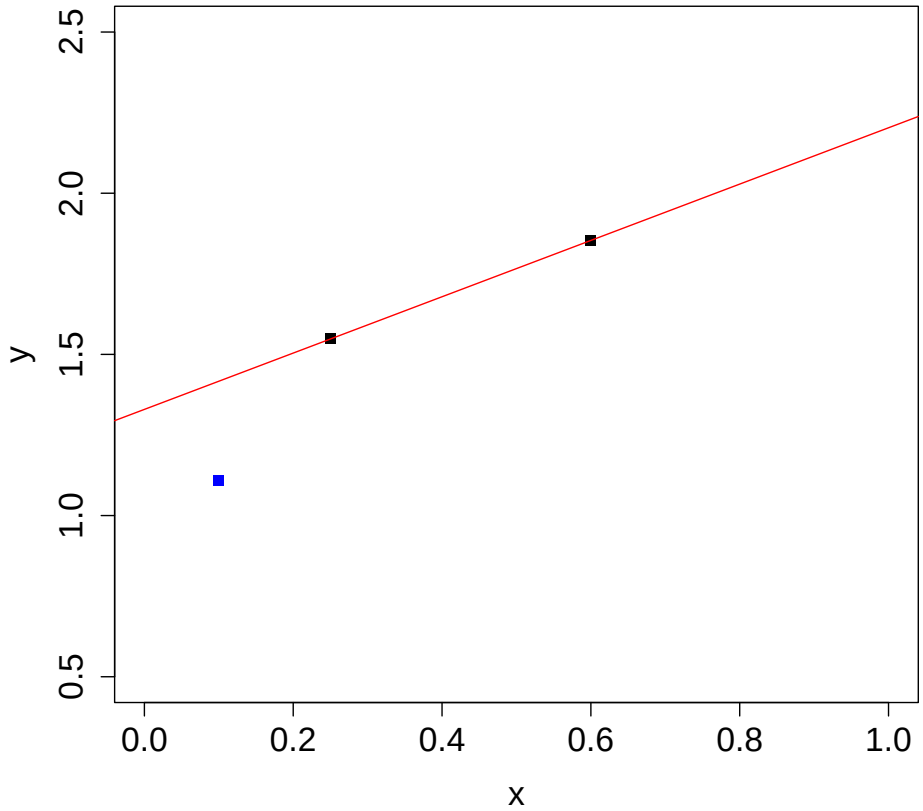
Regression Analysis
Chapter 1 and 2: Simple Linear Regression

Shaobo Jin

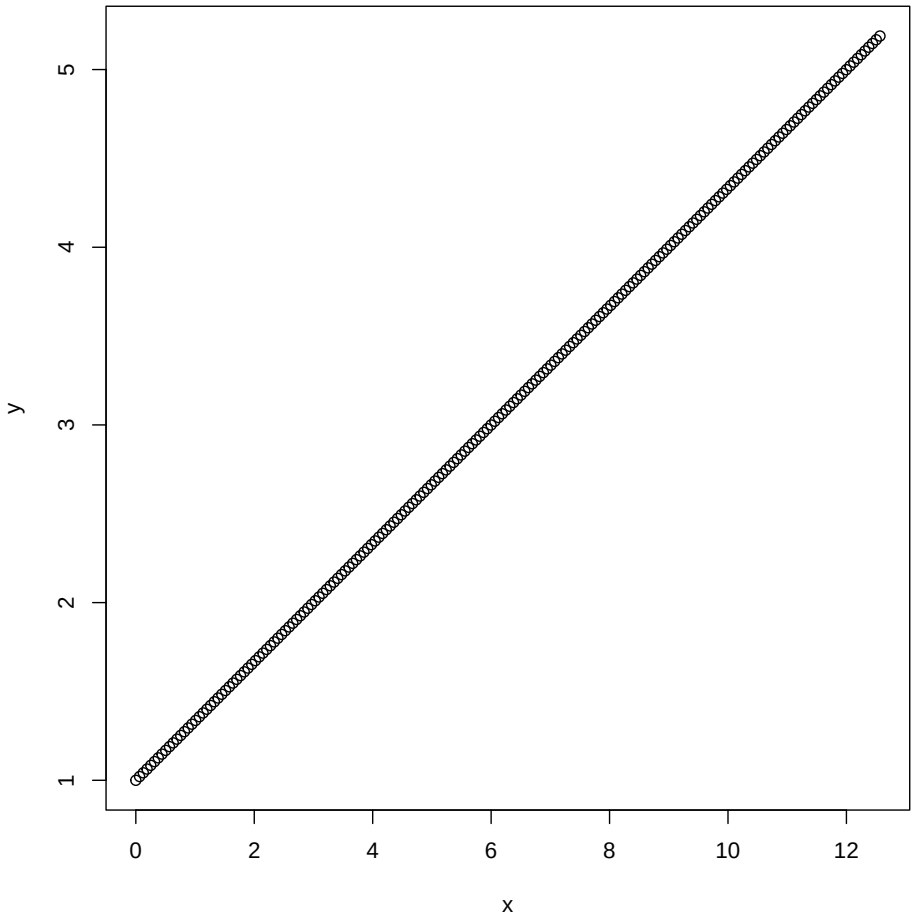
Department of Mathematics

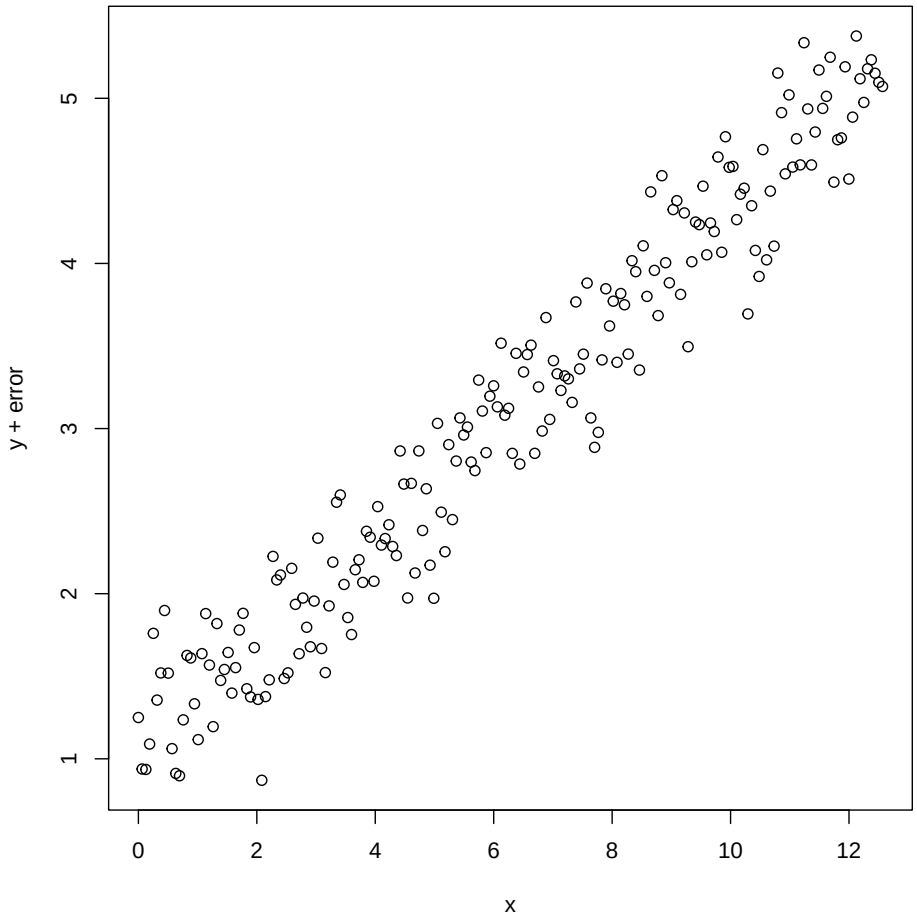
Two Points Determine A Line



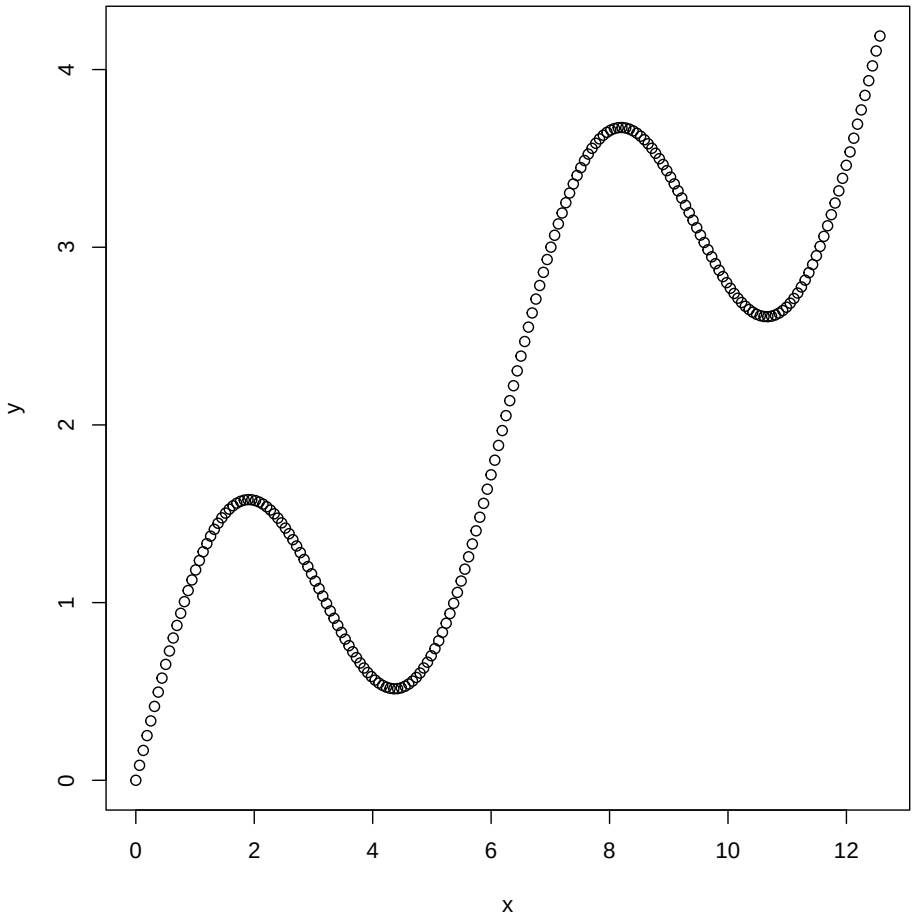


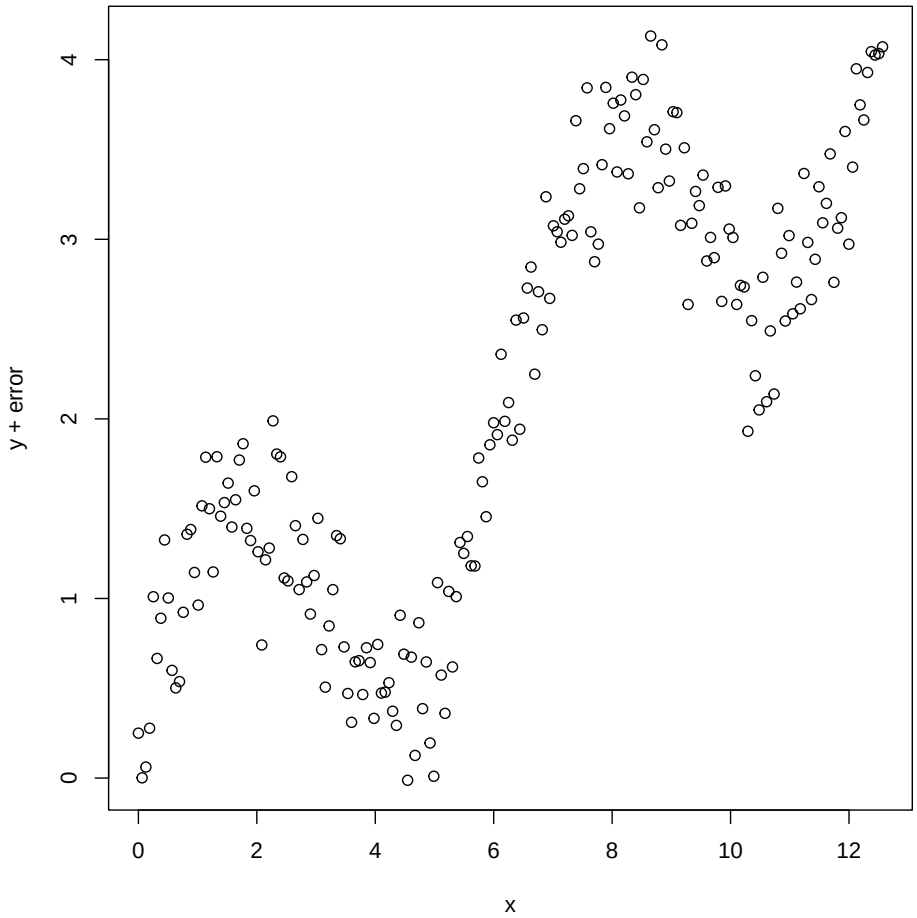
Draw A Line/Curve





Draw A Line/Curve





Goal of Regression

The goal of regression is

1. to summarize observed data as simply, usefully, and elegantly as possible,
2. to make predictions.

Mean Function

The conditional expectation of the response variable Y when the covariate/predictor/regressor is fixed at $\boldsymbol{X} = \boldsymbol{x}$ is

$$\mathrm{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) \quad = \quad f\left(\boldsymbol{x}\right),$$

which is called a mean function. For example

- $f\left(x\right) = \beta_0 + \beta_1 x$ is linear in x, where β_0 (intercept) and β_1 (slope) are the parameters.
- $f\left(\boldsymbol{x}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is linear in $\boldsymbol{x}$ and also linear in β's, where β_0 is the intercept and β_1 and β_2 are the slopes.
- $f\left(x\right) = \beta_0 + \beta_1 x + \beta_2 x^2$ is not linear in x, but still linear in the parameters.
- $f\left(x\right) = \beta_0 + \beta_1 / \left[1 - \exp\left(\beta_2 x\right)\right]$ is not linear.

Variance Function

The conditional variance of the response variable Y when the predictor/regressor is fixed at $\boldsymbol{X} = \boldsymbol{x}$ is

$$\operatorname{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x}),$$

which is called a variance function.

- A frequent assumption is that $\operatorname{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \sigma^2$ (homoscedasticity).
- If $\operatorname{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ depends on the value of $\boldsymbol{X} = \boldsymbol{x}$, then it is called heteroscedasticity.

Simple Linear Regression

We often write our model as

$$y = \mathrm{E}(Y \mid X = x) + e,$$

where e is the random error term. One assumptions that we often make is

$$\mathrm{E}(e \mid X = x) = 0.$$

The simple linear regression model consists of

$$\begin{aligned} \mathrm{E}(Y \mid X = x) &= \beta_0 + \beta_1 x, \\ \mathrm{Var}(Y \mid X = x) &= \sigma^2, \end{aligned}$$

where β_0 is the intercept and β_1 is the slope.

Estimation: Assumptions

The parameters are β_0, β_1 and possibly σ^2. Suppose that we observe a data set (x_i, y_i), $i = 1, 2, ..., n$.

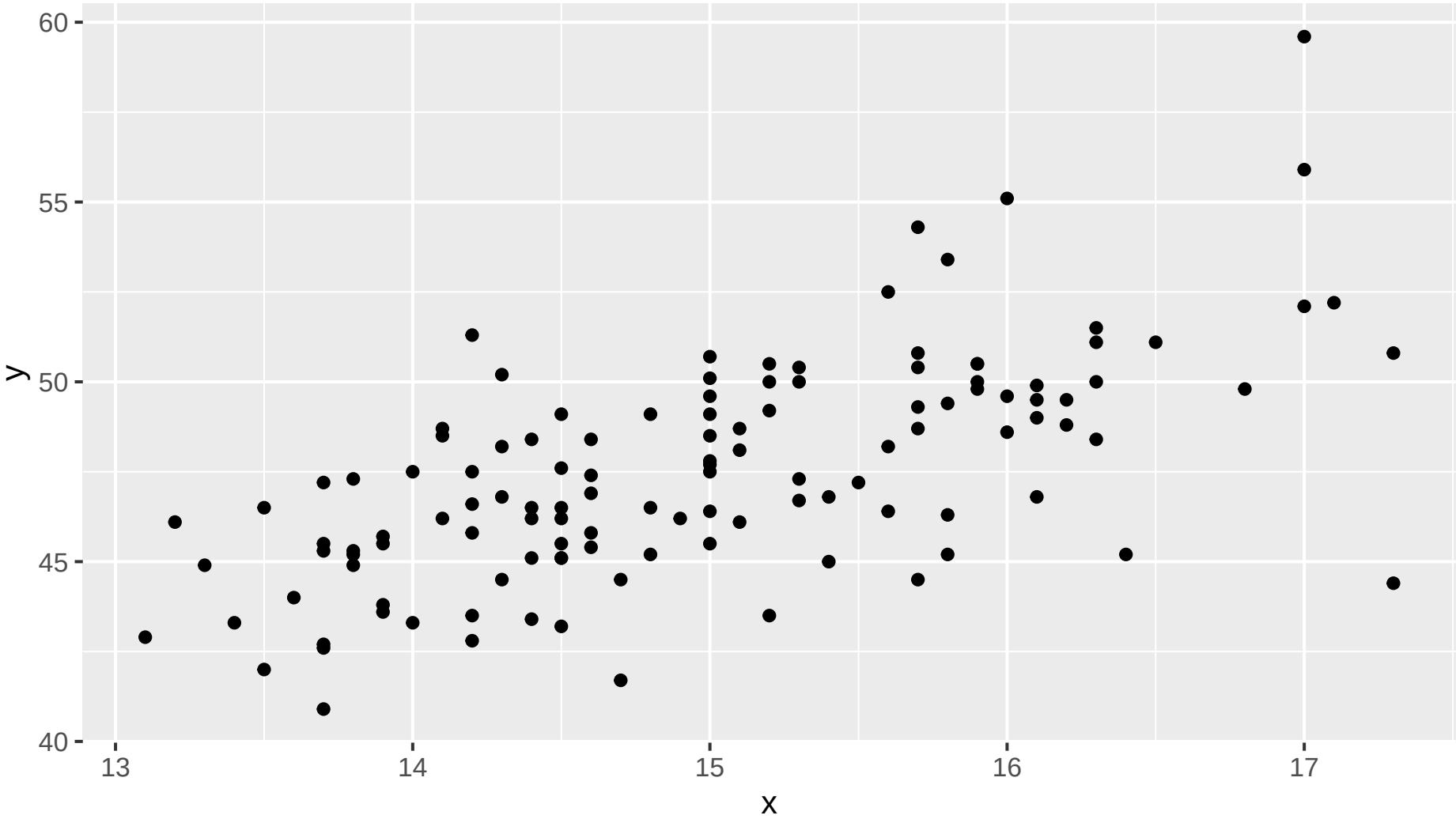
- We often write the simple linear regression model as

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, ...n.$$

- We often assume data are independent, i.e., given all $(X_1, ..., X_n)$, e_i is independent of e_j for all $i \neq j$.

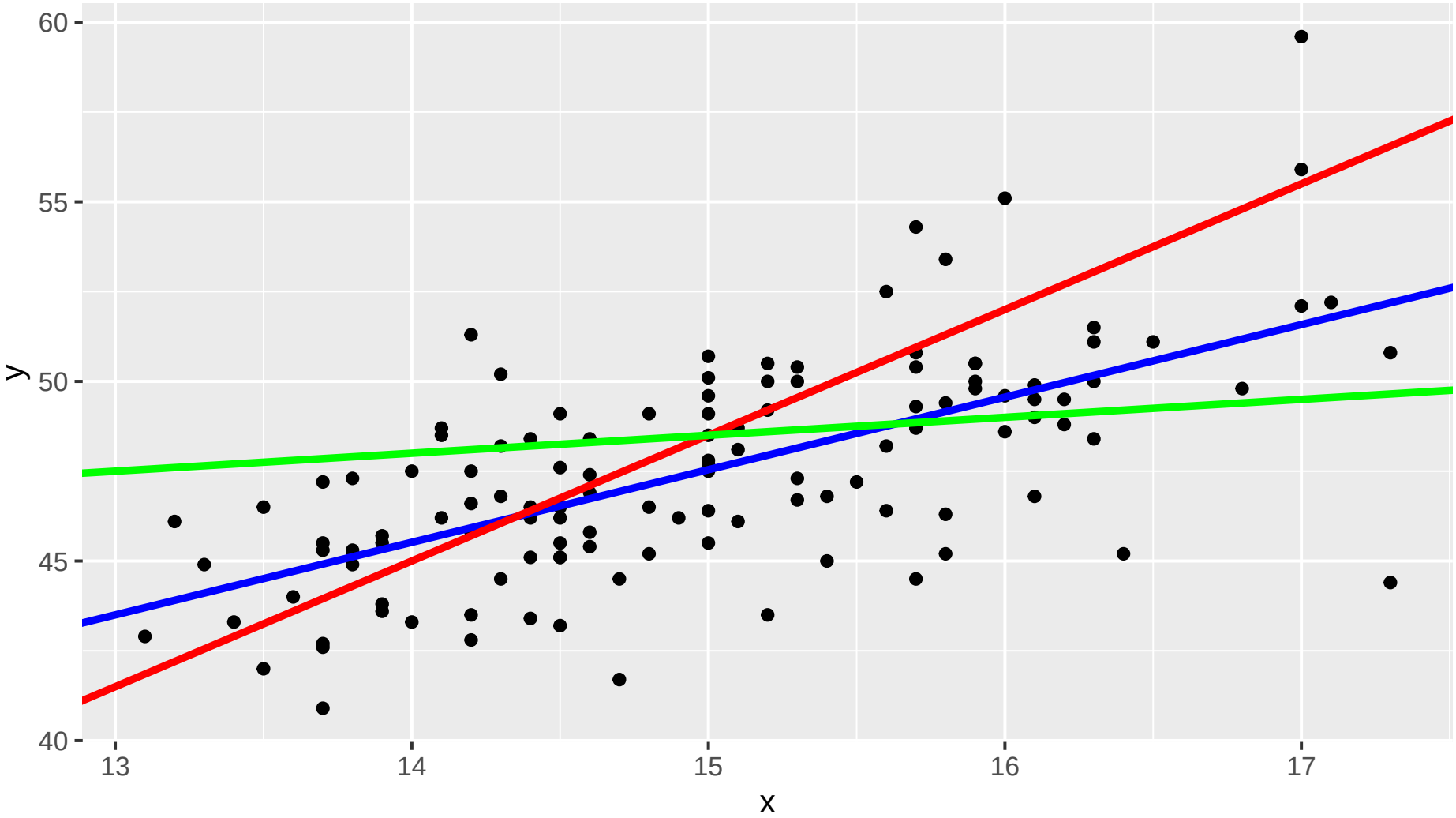
We need to estimate the unknown parameters using our data.

Scatter Plot



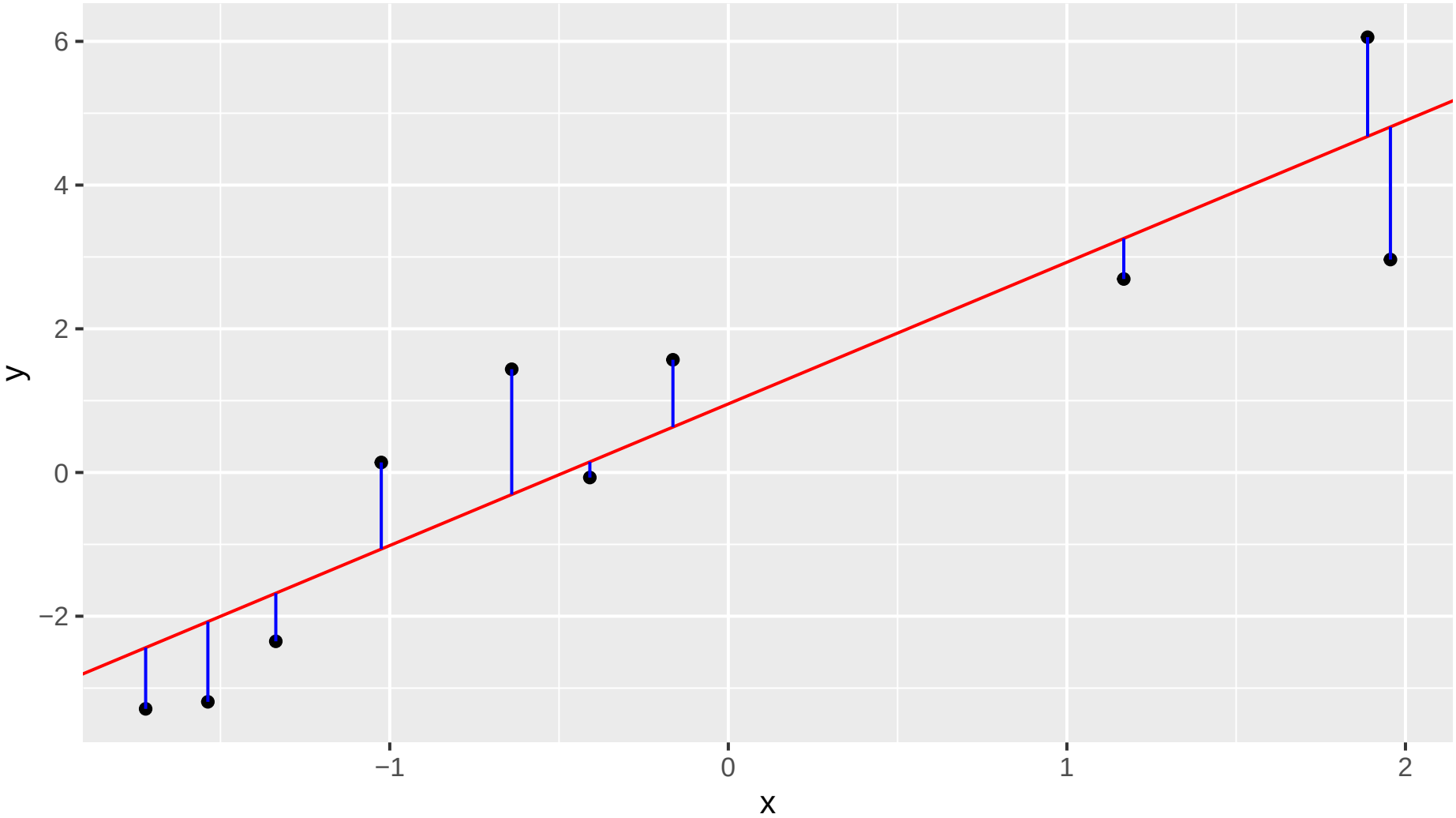
Scatter Plot

Which linear function describes the data the best?



Distance to Conditional Mean

For every straight line, we can compute the vertical distance of our data point to the line $y_i - (\beta_0 + \beta_1 x_i)$:



Ordinary Least Squares

The method of ordinary least squares (OLS) minimizes the sum of squares of $y_i - (\beta_0 + \beta_1 x_i)$. That is, the OLS estimator of β_0 and β_1 are the values such that

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

is minimized, where RSS stands for residual sum of squares.
The minimizer is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} = \sum_{i=1}^{n} y_i/n$ and $\bar{x} = \sum_{i=1}^{n} x_i/n$.

Property of OLS Estimator: Linear Combination

In fact,

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] y_i,$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\,\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] y_i.$$

Both are linear combinations of y_1, ..., y_n.

Property of OLS Estimator: Unbiased

Under the assumption that $\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x$ is correctly specified, the OLS estimators are unbiased as

$$\mathrm{E}\left(\hat{\beta}_0 \mid \boldsymbol{x}\right) = \beta_0, \qquad \mathrm{E}\left(\hat{\beta}_1 \mid \boldsymbol{x}\right) = \beta_1,$$

where $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$.

Property of OLS Estimator: Covariance

Under the assumptions that

1. data are independent conditional on $\boldsymbol{x}$,
2. $\text{Var}\left(Y \mid X = x\right) = \sigma^2$,

the OLS estimators satisfy

$$\begin{aligned}
\text{Var}\left(\hat{\beta}_0 \mid \boldsymbol{X}\right) &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\right], \\
\text{Var}\left(\hat{\beta}_1 \mid \boldsymbol{X}\right) &= \frac{\sigma^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}, \\
\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1 \mid \boldsymbol{X}\right) &= -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}.
\end{aligned}$$

Prediction/Fitted Value and Residual

Once β_0 and β_1 are estimated, the fitted/estimated regression line is

$$\hat{\mathrm{E}}\left(Y \mid X = x\right) \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 x.$$

The residual is

$$\hat{e}_i \quad = \quad y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right).$$

If we have a new observation with x_0 as covariate value, then the predicted value is

$$\hat{\mathrm{E}}\left(Y \mid X = x_0\right) \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Estimating σ^2

The estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}.$$

We call $\hat{\sigma}$ the standard error of regression.

Correlation and Linear Regression

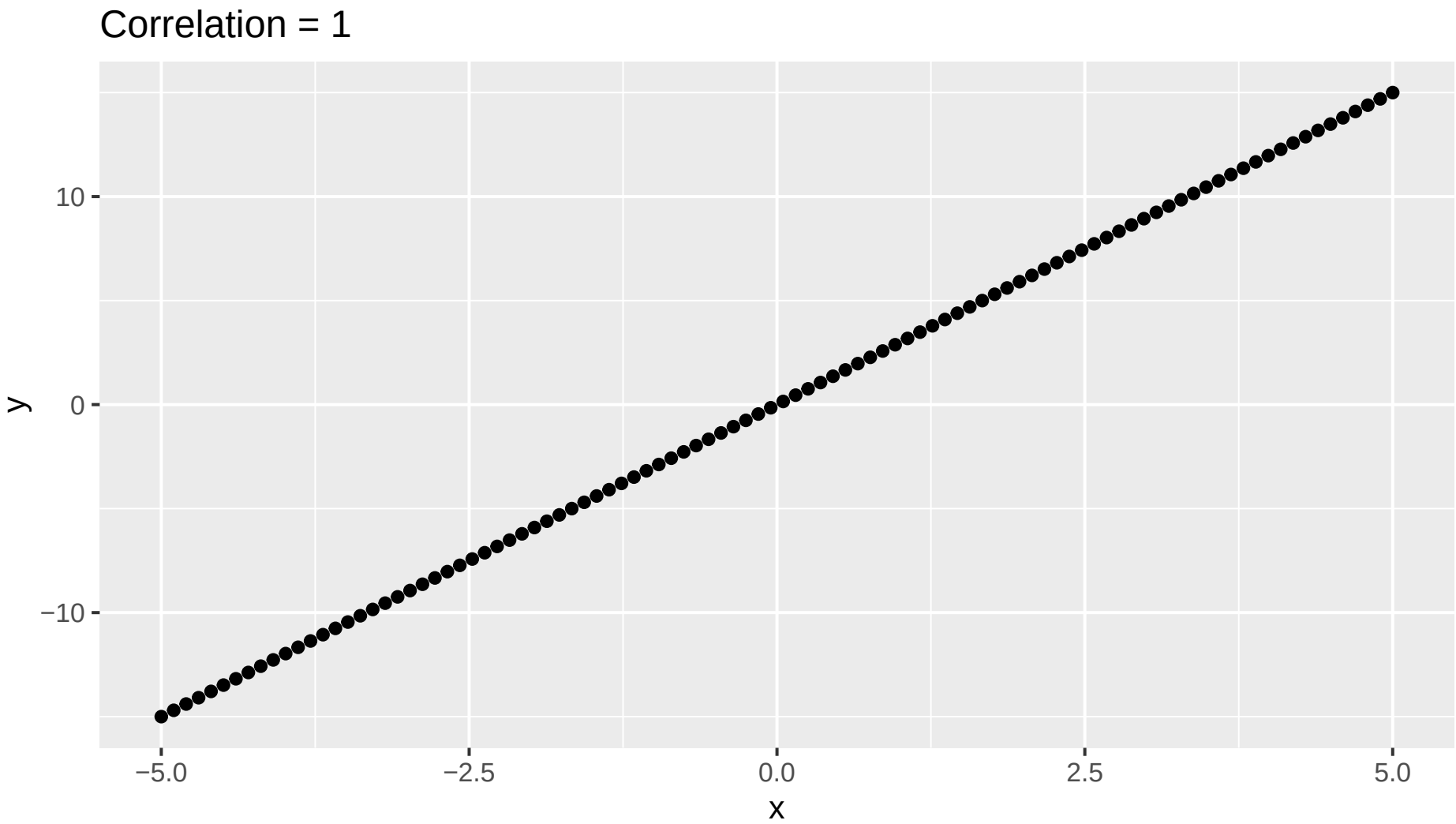
Note that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

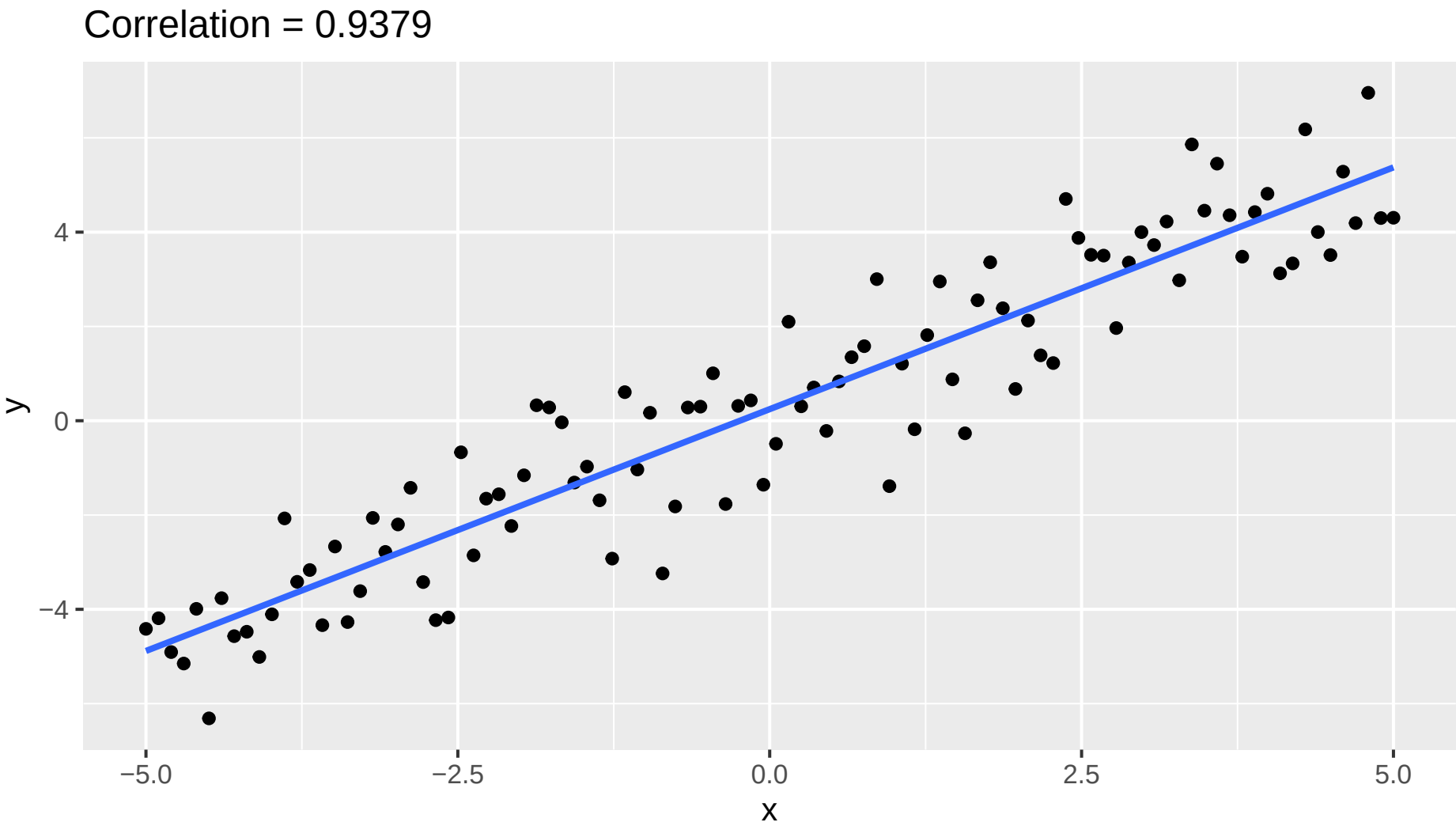
$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

If the Pearson correlation is zero, then we don't have any linear trend. That is, Pearson correlation measure the degree of the linear relation between y and x.

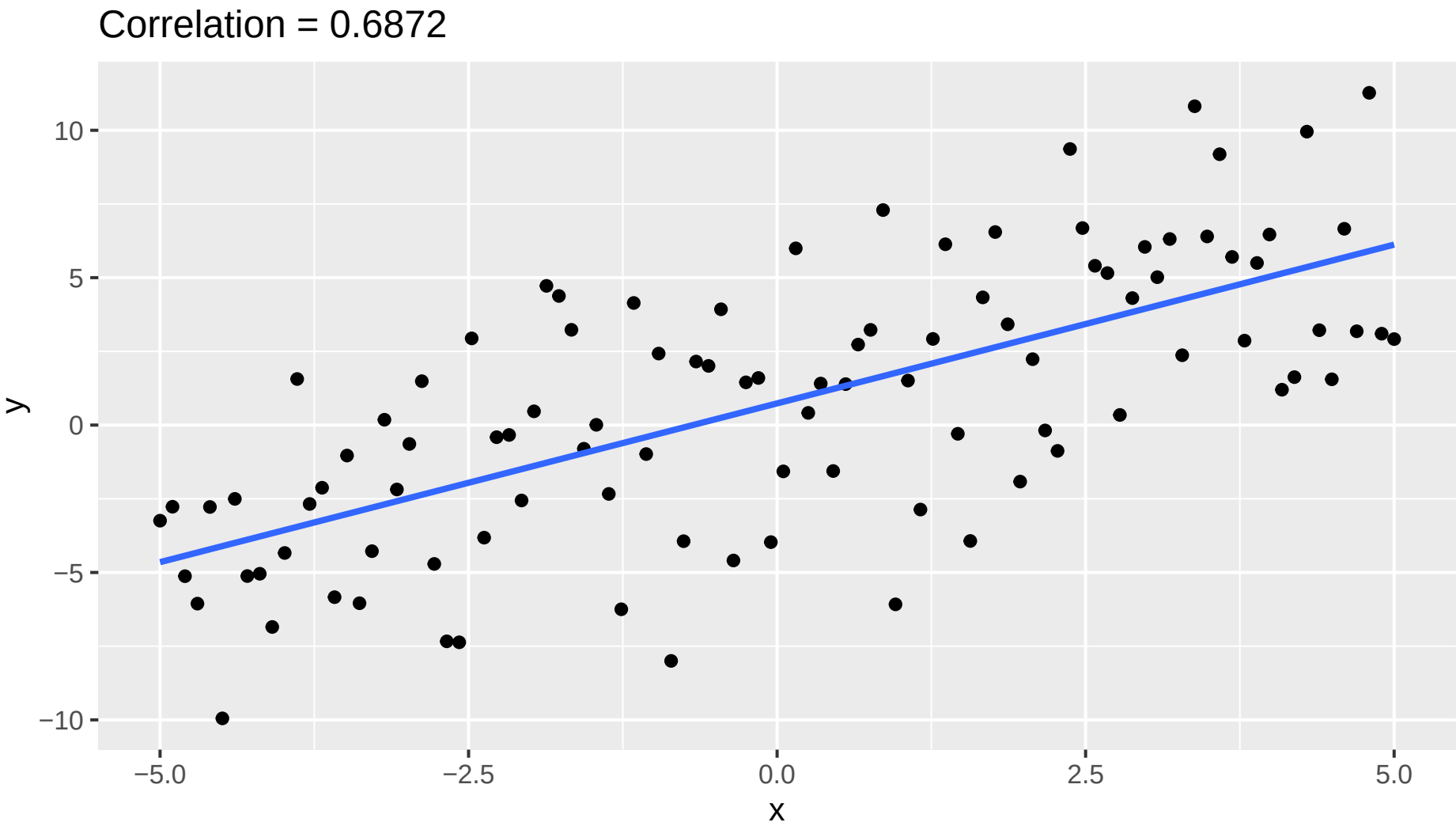
Correlation and Linear Regression



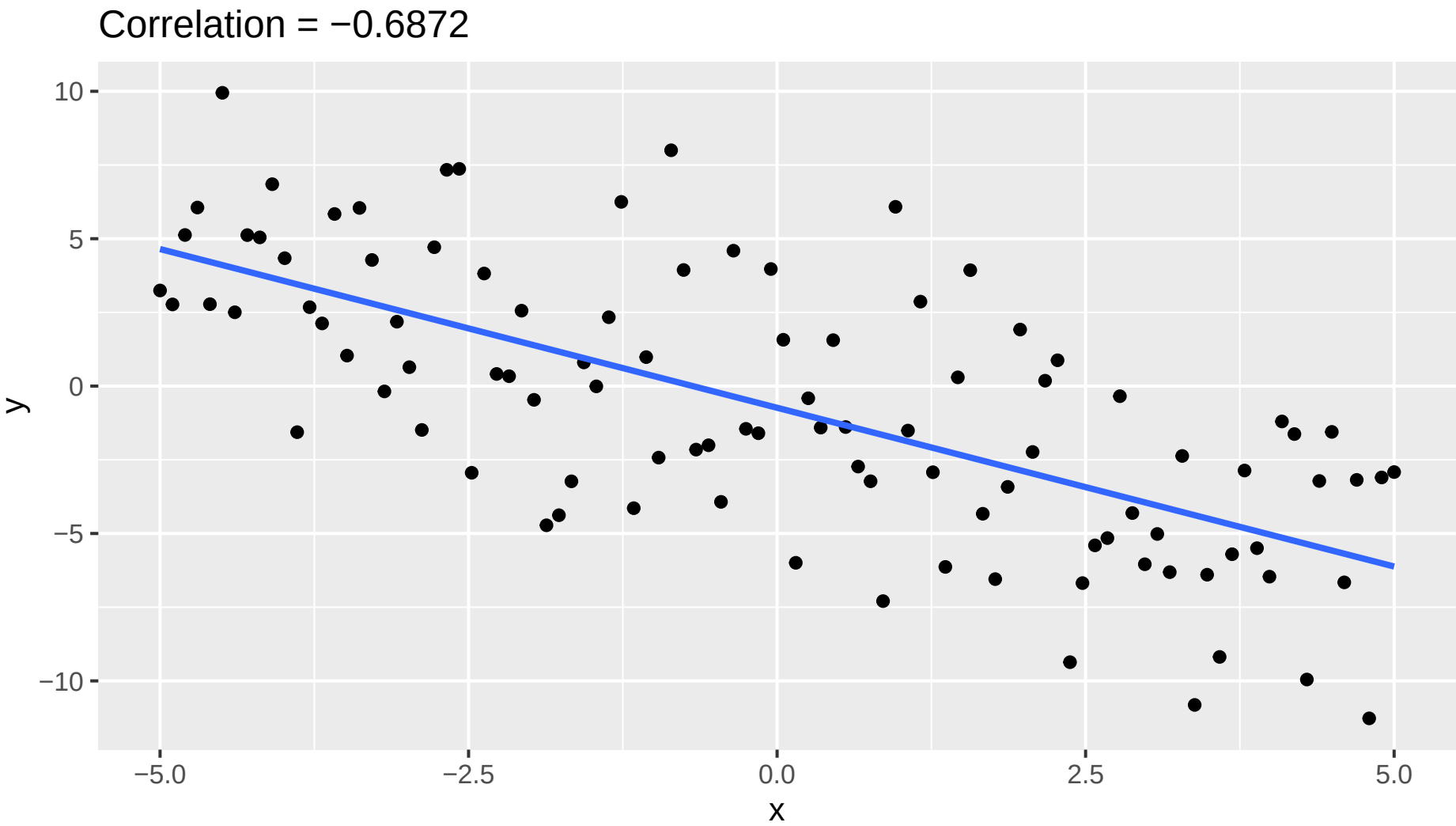
Correlation and Linear Regression



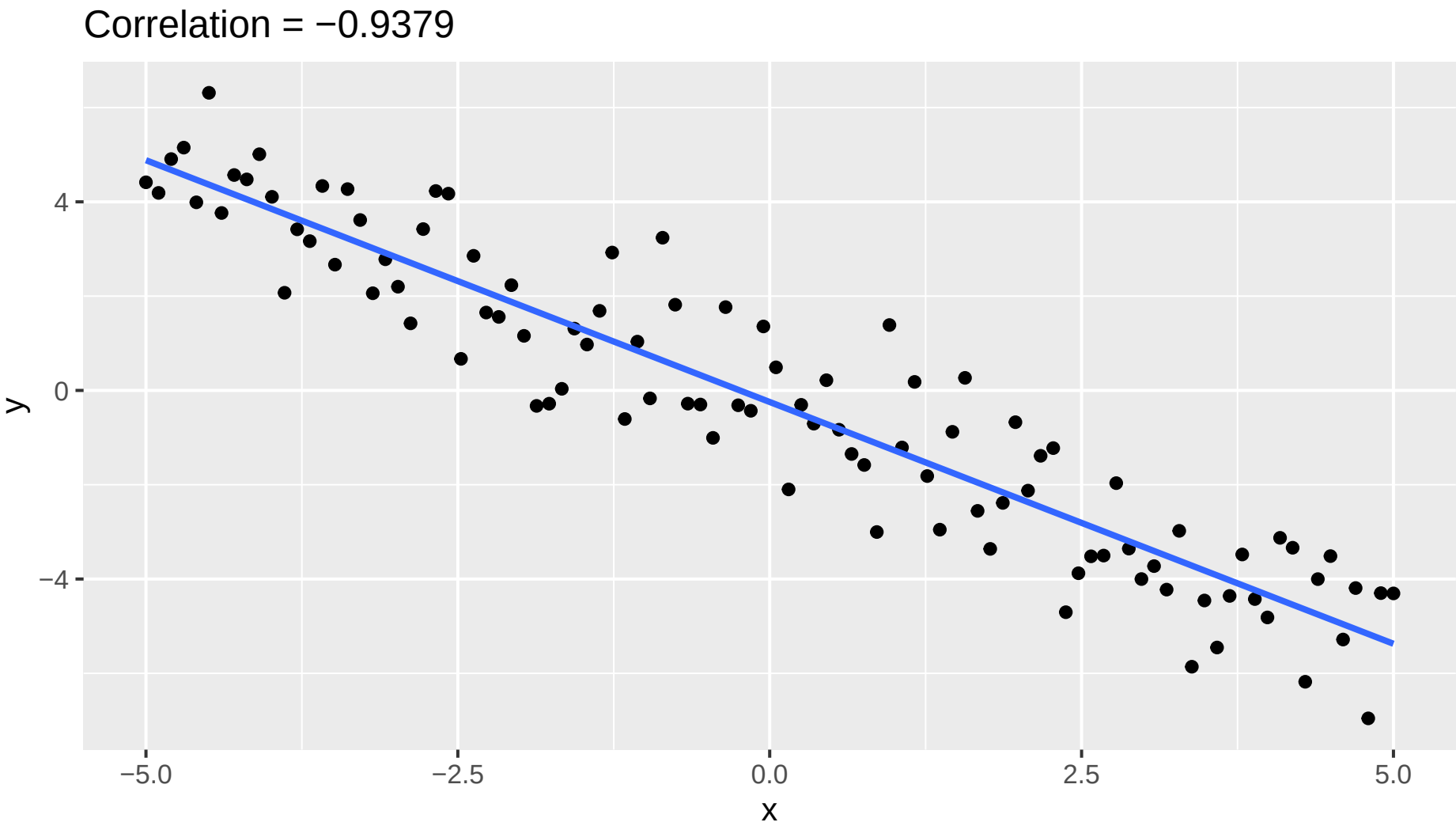
Correlation and Linear Regression



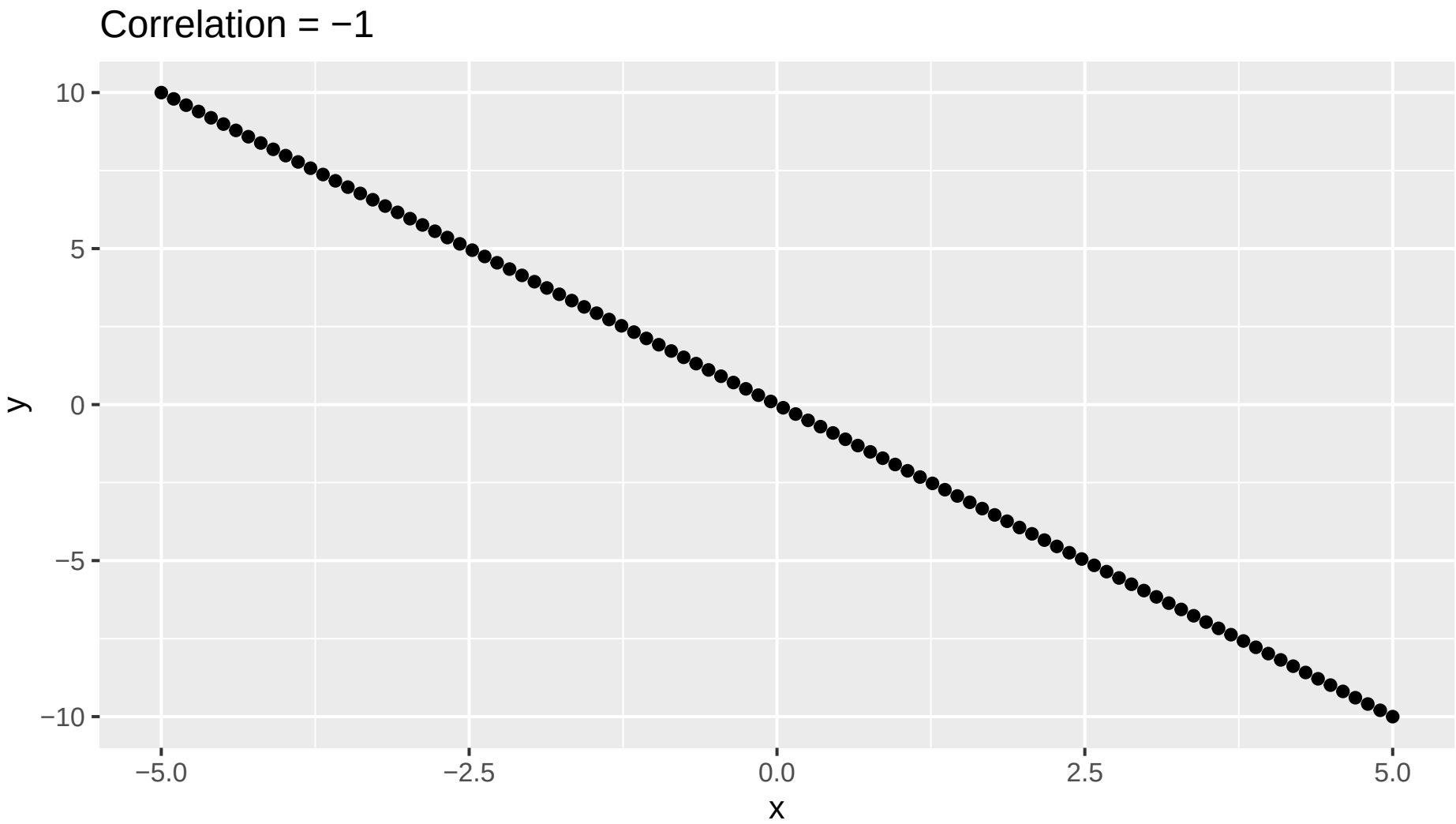
Correlation and Linear Regression



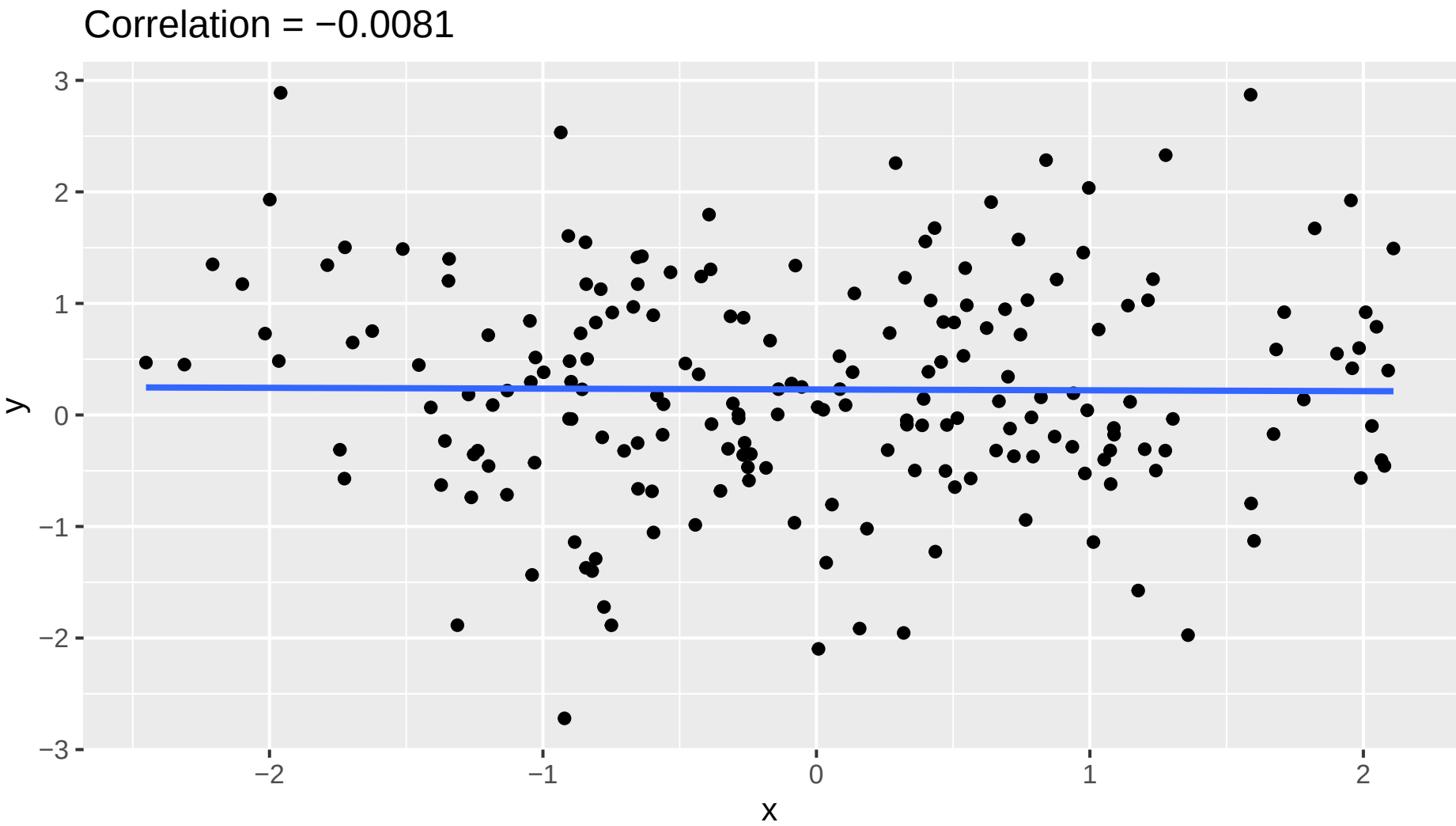
Correlation and Linear Regression



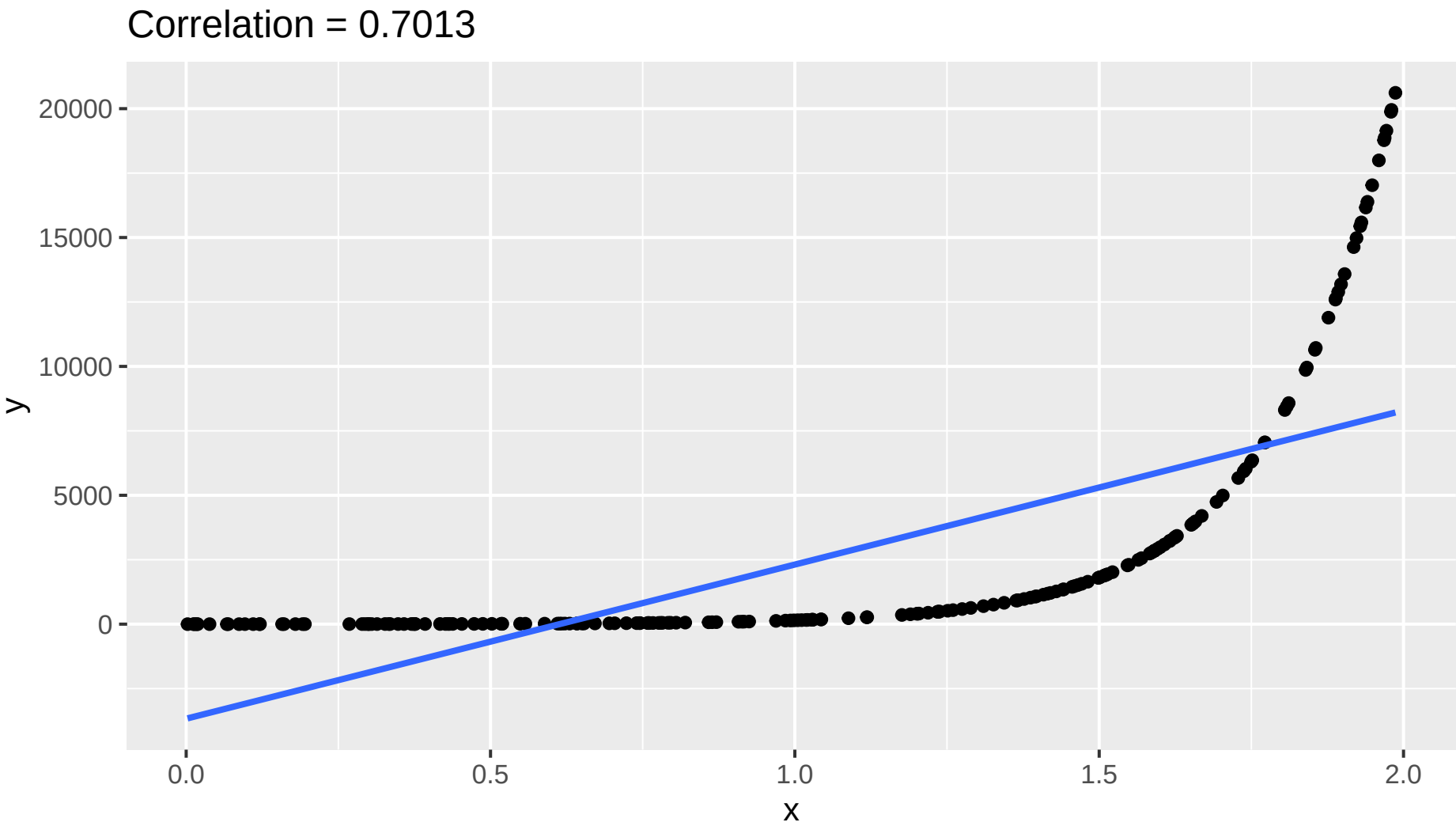
Correlation and Linear Regression



No Linear Relation

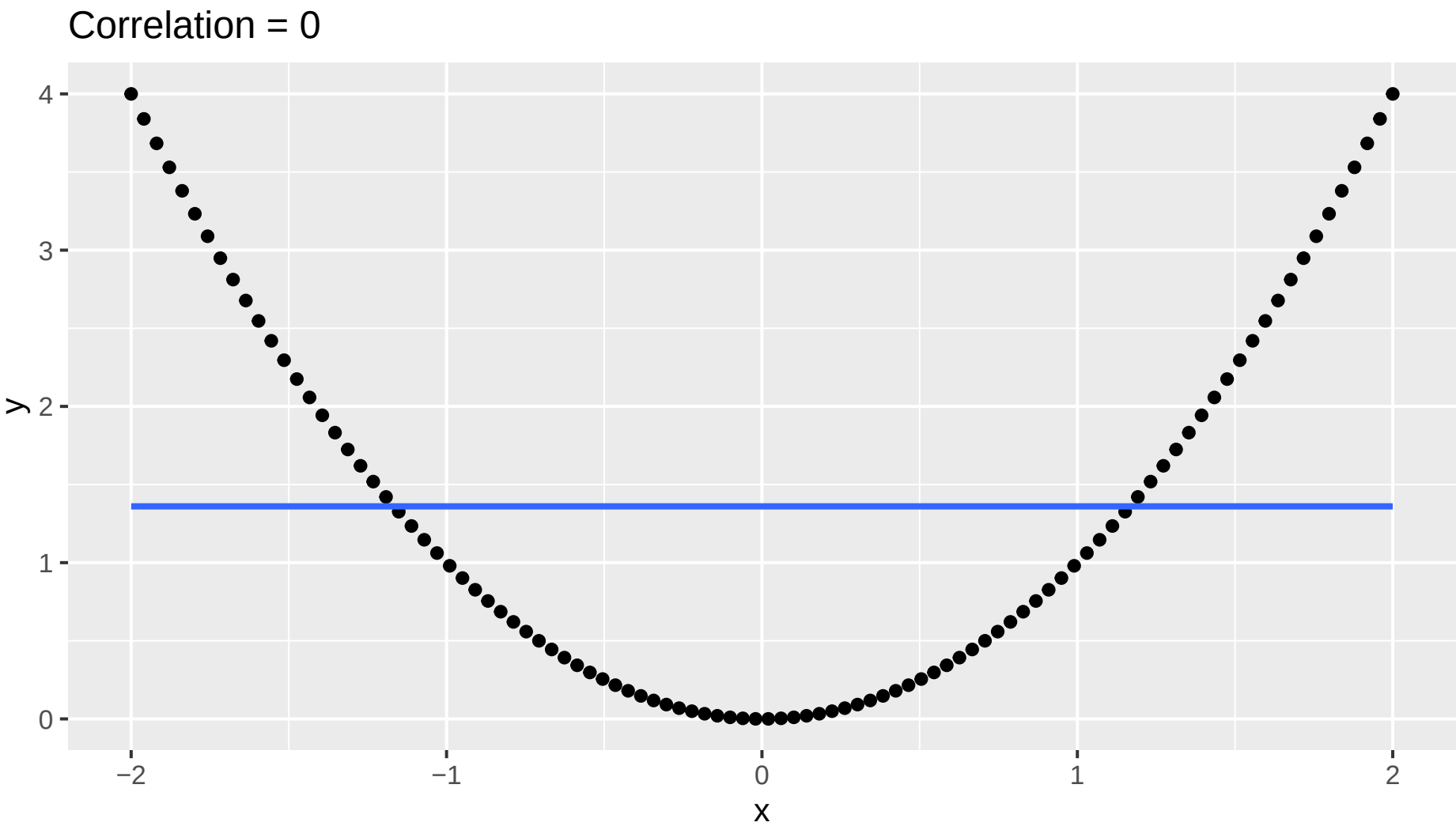


Nonlinear Relation



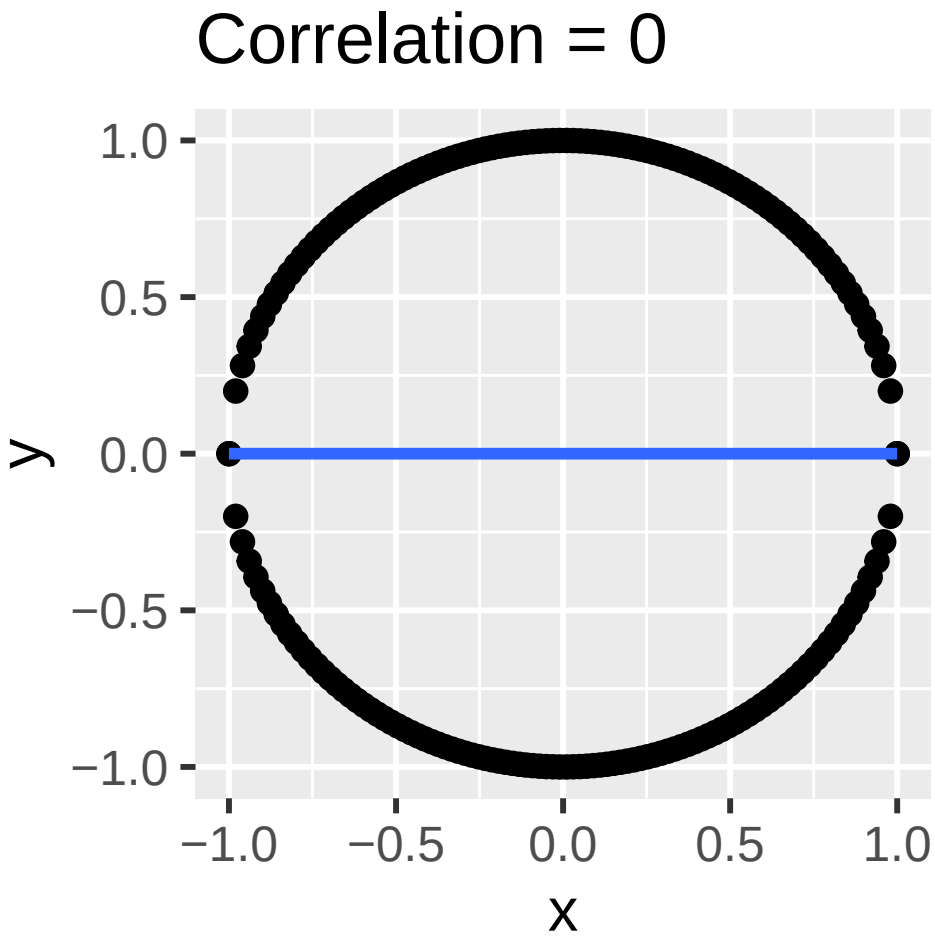
Problem with Pearson Correlation

Pearson correlation only measure the degree of the linear relation between y and x.



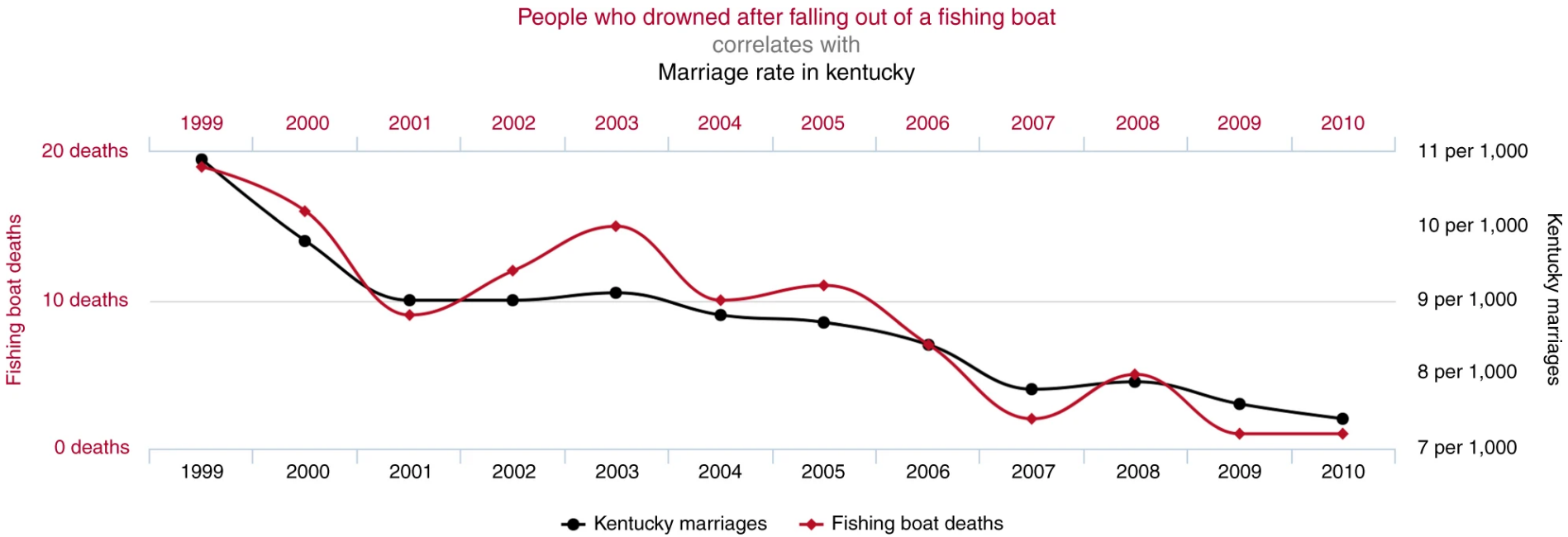
Problem with Pearson Correlation

Pearson correlation only measure the degree of the linear relation between y and x.



Correlation and Causation

Keep in mind that the regression model only describes correlation (association), not causation.



Normally Distributed e

We haven't assume the distribution of e yet. It is also common to assume

$$e \mid X = x \quad \sim \quad N\left(0, \sigma^2\right).$$

Under the independence assumption, the log-likelihood function of β_0, β_1, and σ^2 is

$$\begin{aligned} \ell\left(\beta_0, \beta_1, \sigma^2\right) \quad = \quad & \sum_{i=1}^{n} \left\{ -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\sigma^2\right) \right. \\ & \left. -\frac{1}{2\sigma^2}\left[y_i - \left(\beta_0 + \beta_1 x_i\right)\right]^2 \right\}. \end{aligned}$$

Maximum Likelihood Estimator

The maximum likelihood estimator (MLE) is given by

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\
\hat{\sigma}_n^2 &= \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^2.
\end{aligned}$$

The MLE of β_0 and β_1 are the same as their OLS estimator! Hence, they are still unbiased.

Random Vector and Random Matrix

A random vector/matrix is a vector/matrix of random variables. For example,

$$\boldsymbol{x} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix},$$

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}.$$

Mean and Covariance Matrix

1. The expected value/mean of a random vector/matrix is the vector/matrix consisting of the expected values of each of its elements:

$$\mathrm{E}\left(\boldsymbol{X}_{n\times p}\right) = \begin{bmatrix} \mathrm{E}\left(X_{11}\right) & \cdots & \mathrm{E}\left(X_{1p}\right) \\ \vdots & \ddots & \vdots \\ \mathrm{E}\left(X_{n1}\right) & \cdots & \mathrm{E}\left(X_{np}\right) \end{bmatrix}.$$

2. For a $p \times 1$ random vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu}_X$ and a $q \times 1$ random vector $\boldsymbol{y}$ with mean $\boldsymbol{\mu}_Y$, its covariance matrix is

$$\mathrm{Cov}\left(\boldsymbol{x}, \boldsymbol{y}\right) = \mathrm{E}\left[\left(\boldsymbol{x} - \boldsymbol{\mu}_X\right)\left(\boldsymbol{y} - \boldsymbol{\mu}_Y\right)^T\right] = \mathrm{E}\left(\boldsymbol{x}\boldsymbol{y}^T\right) - \boldsymbol{\mu}_X\boldsymbol{\mu}_Y^T,$$

where its (i,k)th element is $\mathrm{E}\left[\left(X_i - \mu_{X,i}\right)\left(Y_k - \mu_{Y,k}\right)\right]$.

- $\mathrm{Var}\left(\boldsymbol{x}\right) = \mathrm{Cov}\left(\boldsymbol{x}, \boldsymbol{x}\right)$, which is symmetric and positive semi-definite.

Basic Rules: Linear Combination

1. Random variable X, constants c and d:

$$\begin{aligned} \mathrm{E}\left(cX+d\right) &= c\mathrm{E}\left(X\right)+d, \\ \mathrm{Var}\left(cX+d\right) &= c^2\mathrm{Var}\left(X\right). \end{aligned}$$

2. Random vector $\boldsymbol{x}$, constant matrix $\boldsymbol{C}$, constant vector $\boldsymbol{d}$:

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{C}\boldsymbol{x}+\boldsymbol{d}\right) &= \boldsymbol{C}\mathrm{E}\left(\boldsymbol{x}\right)+\boldsymbol{d}, \\ \mathrm{Var}\left(\boldsymbol{C}\boldsymbol{x}+\boldsymbol{d}\right) &= \boldsymbol{C}\mathrm{Var}\left(\boldsymbol{x}\right)\boldsymbol{C}^T. \end{aligned}$$

3. Random matrix $\boldsymbol{X}$, and constant matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{D}$

$$\mathrm{E}\left(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}+\boldsymbol{D}\right) = \boldsymbol{A}\mathrm{E}\left(\boldsymbol{X}\right)\boldsymbol{B}+\boldsymbol{D}.$$

Basic Rules: Additive and Scaling Property

1. Random variables X, Y, and Z, constant a:

$$\begin{aligned} \mathrm{E}(X+Y) &= \mathrm{E}(X)+\mathrm{E}(Y), \\ \mathrm{Cov}(X+Y,Z) &= \mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z), \\ \mathrm{Cov}(aX,Z) &= a\mathrm{Cov}(X,Z). \end{aligned}$$

2. Random vectors $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$, constant matrices $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\begin{aligned} \mathrm{E}(\boldsymbol{x}+\boldsymbol{y}) &= \mathrm{E}(\boldsymbol{x})+\mathrm{E}(\boldsymbol{y}), \\ \mathrm{Cov}(\boldsymbol{x}+\boldsymbol{y},\boldsymbol{z}) &= \mathrm{Cov}(\boldsymbol{x},\boldsymbol{z})+\mathrm{Cov}(\boldsymbol{y},\boldsymbol{z}), \\ \mathrm{Cov}(\boldsymbol{A}\boldsymbol{x},\boldsymbol{B}\boldsymbol{y}) &= \boldsymbol{A}\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})\,\boldsymbol{B}^T. \end{aligned}$$

3. Random matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$\mathrm{E}(\boldsymbol{X}+\boldsymbol{Y}) = \mathrm{E}(\boldsymbol{X})+\mathrm{E}(\boldsymbol{Y}).$$

Multivariate Normal Distribution

Let $\boldsymbol{Z} = \begin{bmatrix} Z_1 & Z_2 & \cdots & Z_p \end{bmatrix}^T$ be a random vector, each $Z_j \sim N(0,1)$, and Z_j is independent of Z_k for any $j \neq k$. Let $\boldsymbol{A}$ be a constant matrix and $\boldsymbol{\mu}$ be a constant vector. Then,

$$\boldsymbol{X} \quad = \quad \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$$

follows a $p-$dimensional multivariate normal distribution. We often denote a multivariate normal random vector by $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the covariance matrix.

Using multivariate normal distribution, we can write the assumption $e \mid X = x \sim N\left(0, \sigma^2\right)$ as

$$\boldsymbol{e} \mid \boldsymbol{x} \quad \sim \quad N\left(0, \sigma^2 \boldsymbol{I}\right).$$

Properties Needed In Our Course

1. If $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{A}$ is a matrix of constants, and $\boldsymbol{d}$ is a vector of constants, then $\boldsymbol{AX} + \boldsymbol{d} \sim N\left(\boldsymbol{A\mu} + \boldsymbol{d}, \boldsymbol{A\Sigma A}^T\right)$.
2. Let $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}$ be distributed as $N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right)$.
   1. All subsets of $\boldsymbol{x}$ are normally distributed, i.e.,$\boldsymbol{X}_1$ is distributed as $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, and $\boldsymbol{X}_2$ is distributed as $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.
   2. If X_i is the ith element in $\boldsymbol{x}$, then $X_i \sim N(\mu_i, \Sigma_{ii})$.
   3. The conditional distribution of $\boldsymbol{x}_1$ given that $\boldsymbol{x}_2$, is

$$\boldsymbol{x}_1 \mid \boldsymbol{x}_2 \quad \sim \quad N\left\{\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2),\ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right\},$$

   provided that $\boldsymbol{\Sigma}_{22}$ is invertible.

Distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

Under the normality assumption, we can obtain

$$\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{bmatrix} \sim N\left(\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix}, \begin{bmatrix}\sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right] & -\frac{\sigma^2\bar{x}}{\sum_{i=1}^n(x_i-\bar{x})^2} \\ -\frac{\sigma^2\bar{x}}{\sum_{i=1}^n(x_i-\bar{x})^2} & \frac{\sigma^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\end{bmatrix}\right).$$

The marginal distributions are

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right]\right),$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right).$$

The standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\hat{\sigma}\sqrt{\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right]} \quad \text{and} \quad \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n(x_i-\bar{x})^2}}, \text{ respectively.}$$

Student t-Distribution

It can be shown that, conditional on $\boldsymbol{X}$,

$$\frac{\sum_{i=1}^{n} \hat{e}_i^2}{\sigma^2} \sim \chi^2 (n-2).$$

Then,

$$\frac{\left(\hat{\beta}_0 - \beta_0\right) / \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}{\sqrt{\sum_{i=1}^{n} \hat{e}_i^2 / (n-2)}} \sim t(n-2),$$

and

$$\frac{\left(\hat{\beta}_1 - \beta_1\right) / \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}{\sqrt{\sum_{i=1}^{n} \hat{e}_i^2 / (n-2)}} \sim t(n-2).$$

Confidence Interval

A $1-\alpha$ confidence interval for β_0 is

$$\hat{\beta}_0 \pm t_{1-\alpha/2}\,(n-2)\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)},$$

where $t_{1-\alpha/2}\,(n-2)$ satisfies

$$P\left(T \le t_{1-\alpha/2}\,(n-2)\right) \;=\; 1-\frac{\alpha}{2},$$

with $T \sim t\,(n-2)$.
A $1-\alpha$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{1-\alpha/2}\,(n-2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}.$$