Regression

AU STAT-615

Emil Hvitfeldt

2021-1-20

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Linear Regression with one Predictor Variable

Definition

> Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that a response variable can be predicted from the others

Examples

> Sales of a product predicted by the amount of advertising spent
>
> Amount of rain predicted by hours of rain

Relationos between variables

- Functional relation
- Statistical Relation

Functional Relation

Is expression by a mathematical formula

$$Y = f(X)$$

where f is a function mapping X to Y.

Functional Relation

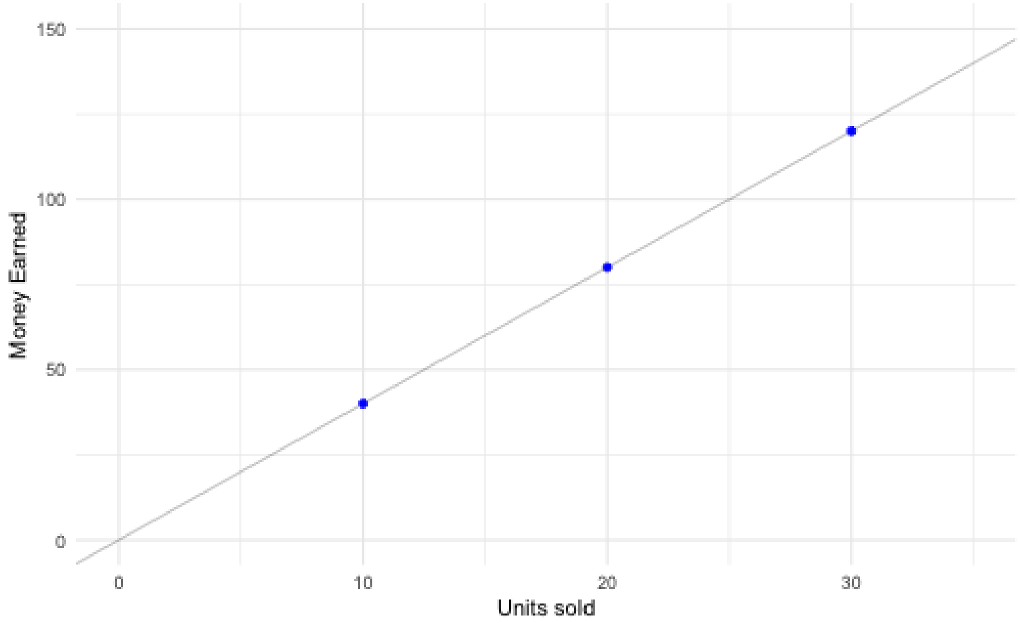
Case:

Money made (Y) of a product sold at a fixed price and the number of units sold (X)

Price of unit: 2

The functional relation will be

$$Y = 4X$$



Statistical Relation

We don't have a perfect relation between the variables

In other words, the points will not always fall on the line

The relationship between the response and predictors can strong or weak depending on the case

Statistical Relation

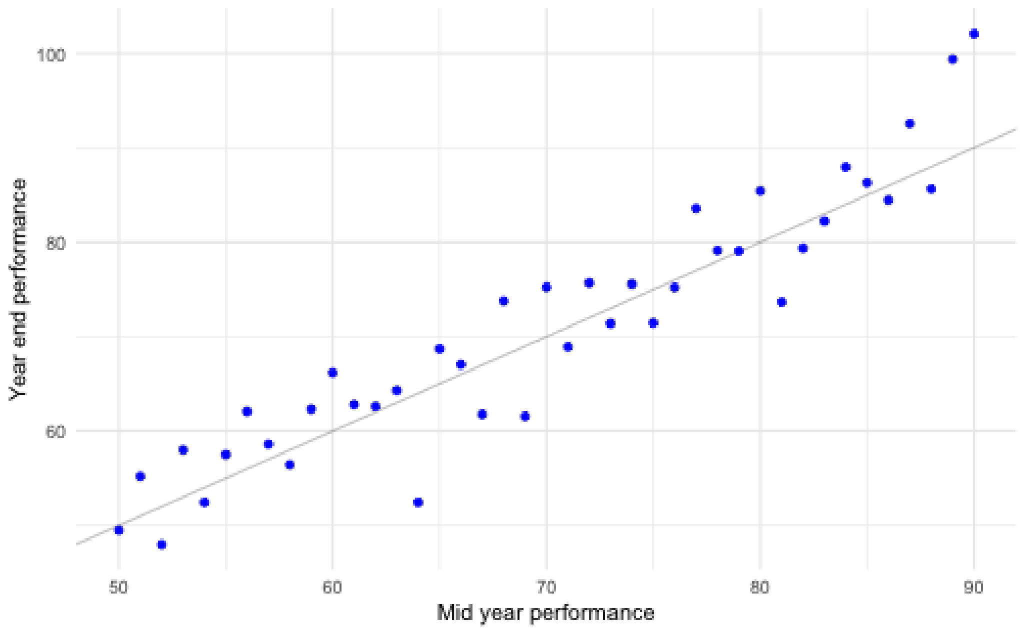
Case:

mid-year and year-end performance for employees

The Statistical relation will be

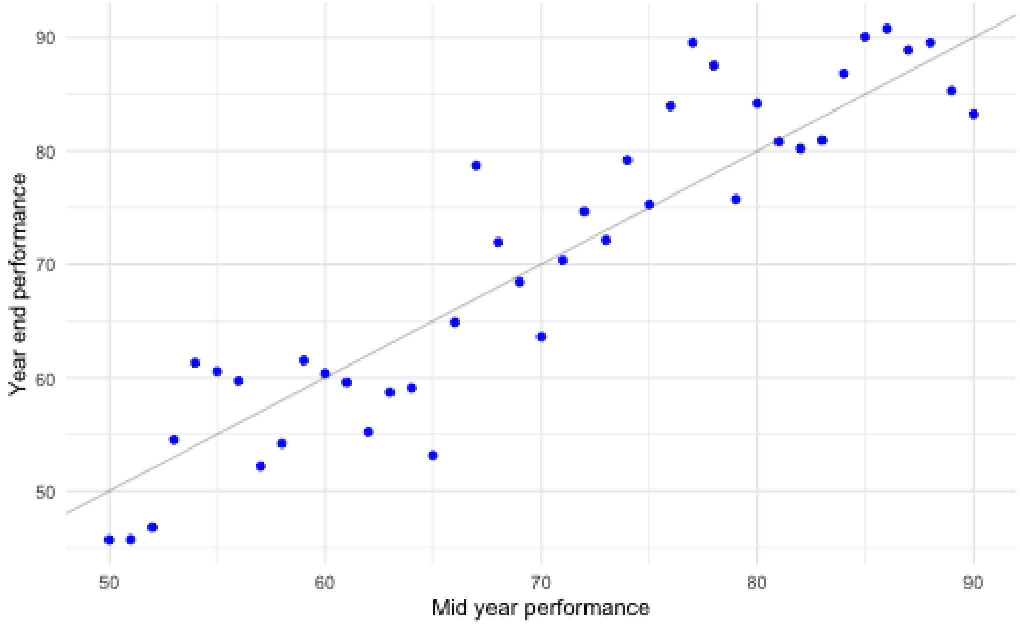
$$Y = X + \varepsilon$$

Notice how the relationship is not perfect



Statistical Relation

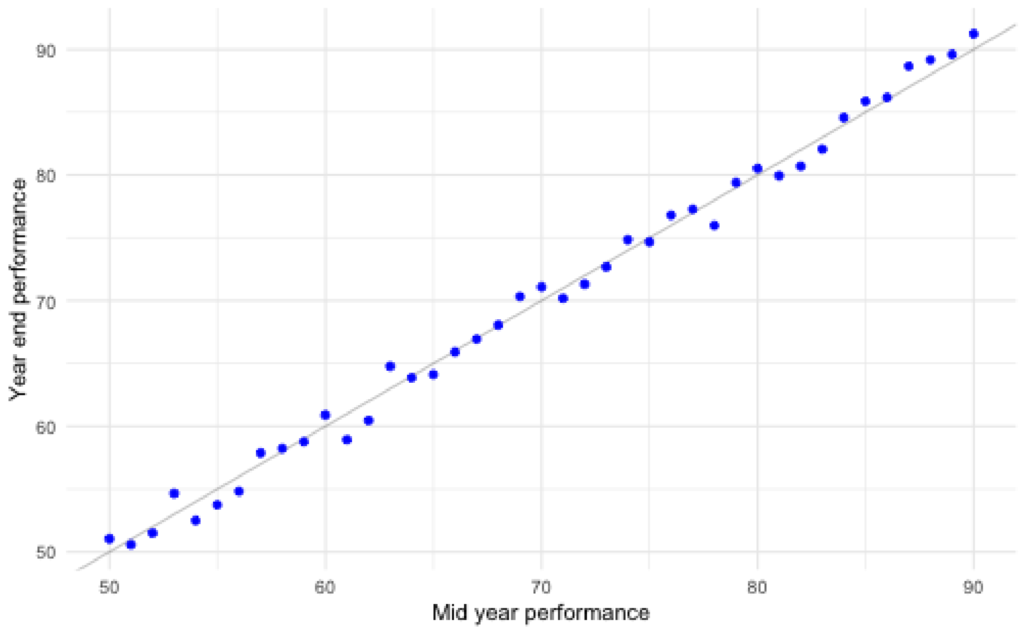
The scattering of the points represents **variation** in the year-end performance that is not associated with the mid-year performance



Statistical Relation

The scattering of the points represents **variation** in the year-end performance that is not associated with the mid-year performance

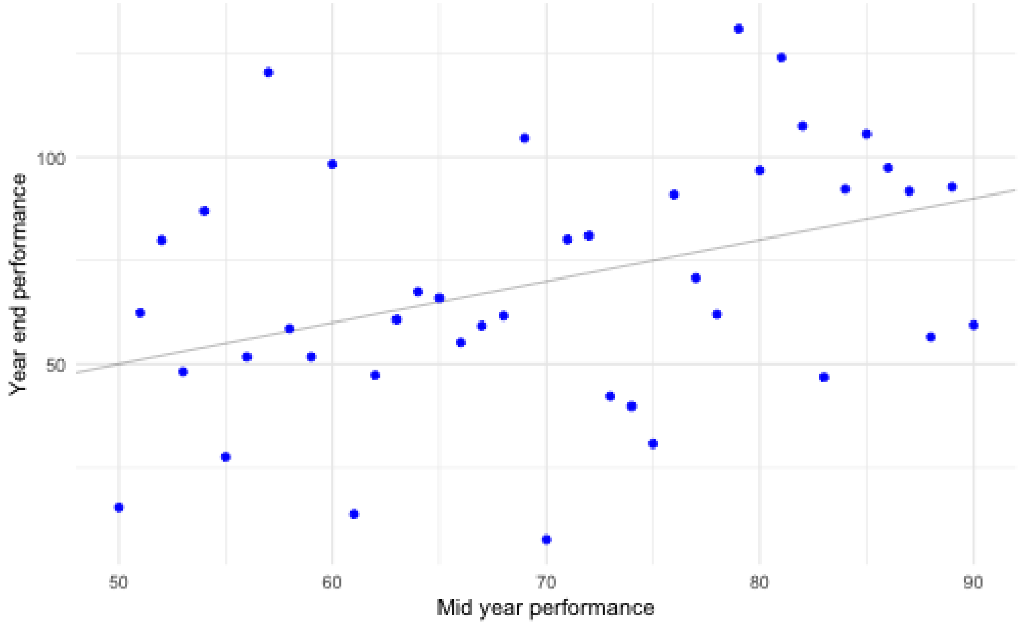
And it can be a **small** amount of variation



Statistical Relation

The scattering of the points represents **variation** in the year-end performance that is not associated with the mid-year performance

And it can be a **large** amount of variation



Sneak peek: More than one predictor variable

Example: Study of short children

X_1: Age
X_2: Gender X_3: Height X_4: Weight

Y: Peak plasma growth hormone level

There is a probability distribution of Y for each level of X_i

Construction of Regression models

- Selection of predictor variables (more about this in chapter 9)
- Functional form of regression relation
- Scope of Model, must be generalizable

Use of regression Analysis

- Description/Inference
- Control
- Prediction

Regression and Causality

The existence of a statistical relation between the response variable Y and the predictor X does not imply that Y depends on X

Example:

X: Size of vocabulary Y: Writing speed of children

Will show a positive statistical relation

This does not imply that an increase in vocabulary causes a faster writing speed

What is more likely is that a 3rd variable such as "age of the child" positively affects both

Regression and Causality

This should not mean that statistical relations never have a causal link, but that we need to spend a little more time with the problem to infer that there is one

Notation

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

Y_i Value of response variable at the ith observation

β_0, β_1 parameters (1 value each)

X_i Value of response variable at the ith observation

ε_i Random error at ith observation

ε is the random error term with mean $E\{\varepsilon_i\} = 0$ and $V\{\varepsilon_i\} = \sigma^2$

The different error terms are uncorrelated

Notation

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

We say that Y denotes a random variable and y denotes a potential value of that random variable

Some observations

- Y_i is the sum of $\beta_0 + \beta_1 X_i$ which is constant and the random term ε, hence Y_i is a random variable.

- $E\{\varepsilon\} = 0$ then $E\{Y_i\} = E\{\beta_0 + \beta_1 X_i + \varepsilon\} = \beta_0 + \beta_1 X_i$

- $V\{Y_i\} = V\{\beta_0 + \beta_1 X_i + \varepsilon\} = V\{\varepsilon_i\} = \sigma^2$

- Y_i and Y_j are uncorrelated since the error terms are uncorrelated

Example

> Relationship between the number of bids requested for contractors during a week and the time required to prepare the bids

Let the regression model be

$$Y_i = 9.5 + 2.1X_i + \varepsilon_i$$

for i representing different weeks

Y: Number of hours required to prepare bids

X: Number of bids prepared in a week

Example

> Relationship between the number of bids requested for contractors during a week and the time required to prepare the bids

The regression function for this model is

$$EY = 9.5 + 2.1X$$

If we suppose that the ith week, $X_i = 45$ then we would expect the number of hours spent preparing to be 104. But if the actual number of hours $Y_i = 108$ then the error is $\varepsilon_i = 4$

ε_i is the deviation or Y_i from its mean value $E\{Y_i\}$

Meaning of Regression Parameters

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

β_1: Slope of the regression line. It indicates the change in the mean of the probability distribution of Y per unit increase in X.

β_0: Y intercept of the regression line. When $X = 0$ gives mean of probability distribution of Y.

Estimation of Regression Function

The data will be used to estimate the parameters of the regression function.

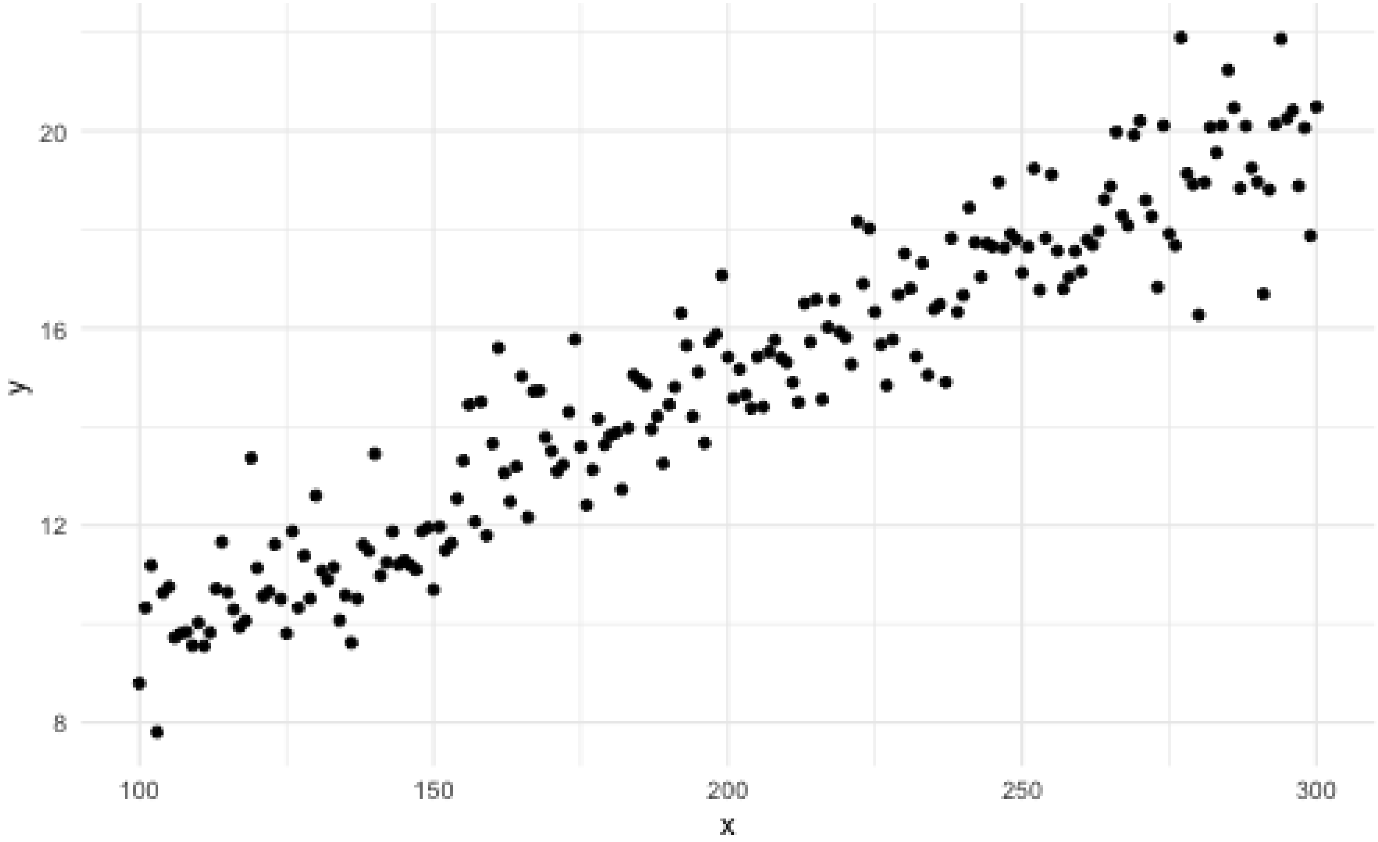
We will think of the observations (X, Y) as consisting of the pair of numbers

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

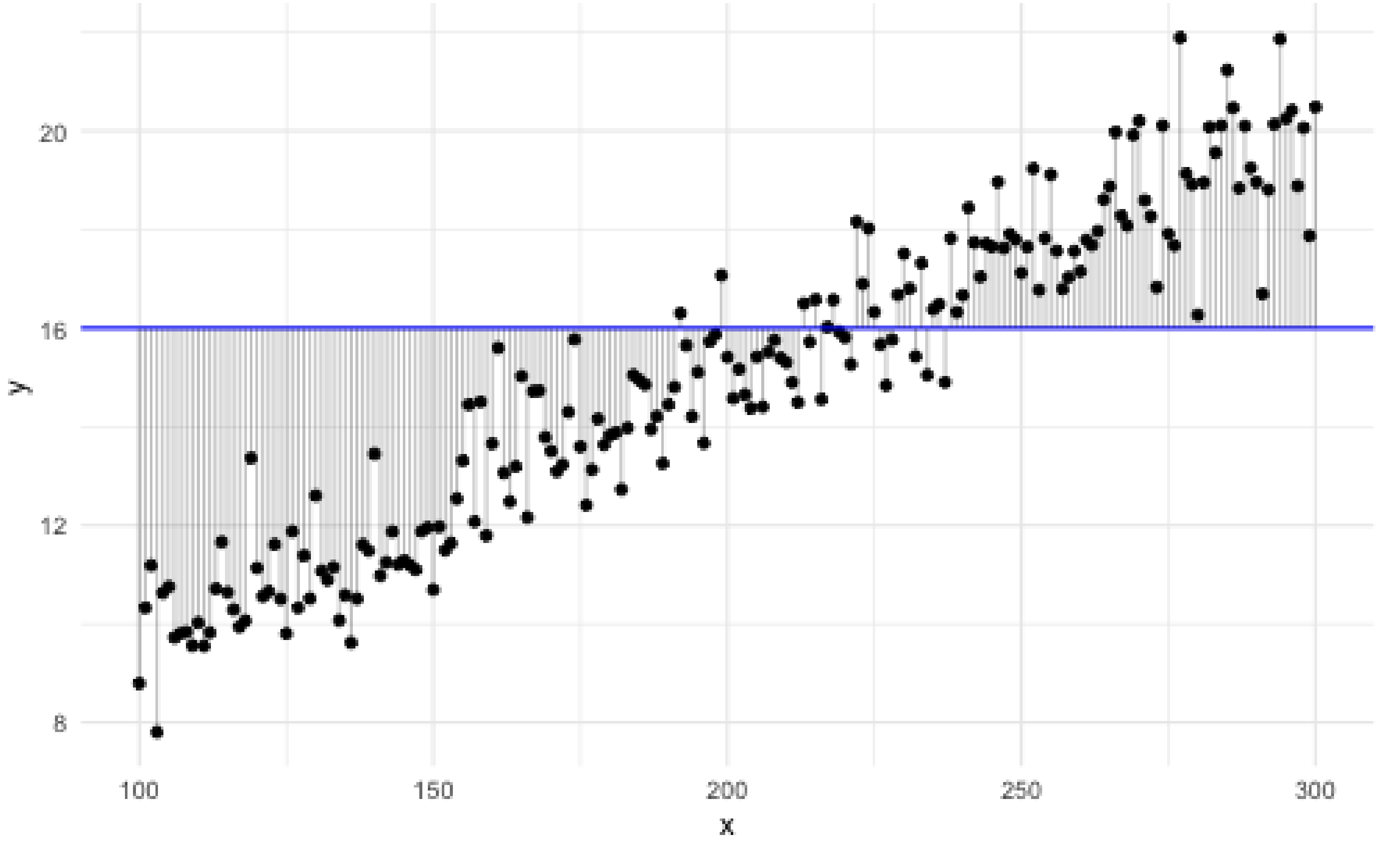
Method of least squares

We use the method of least squares (MLS) to effectively estimate β_0 and β_1

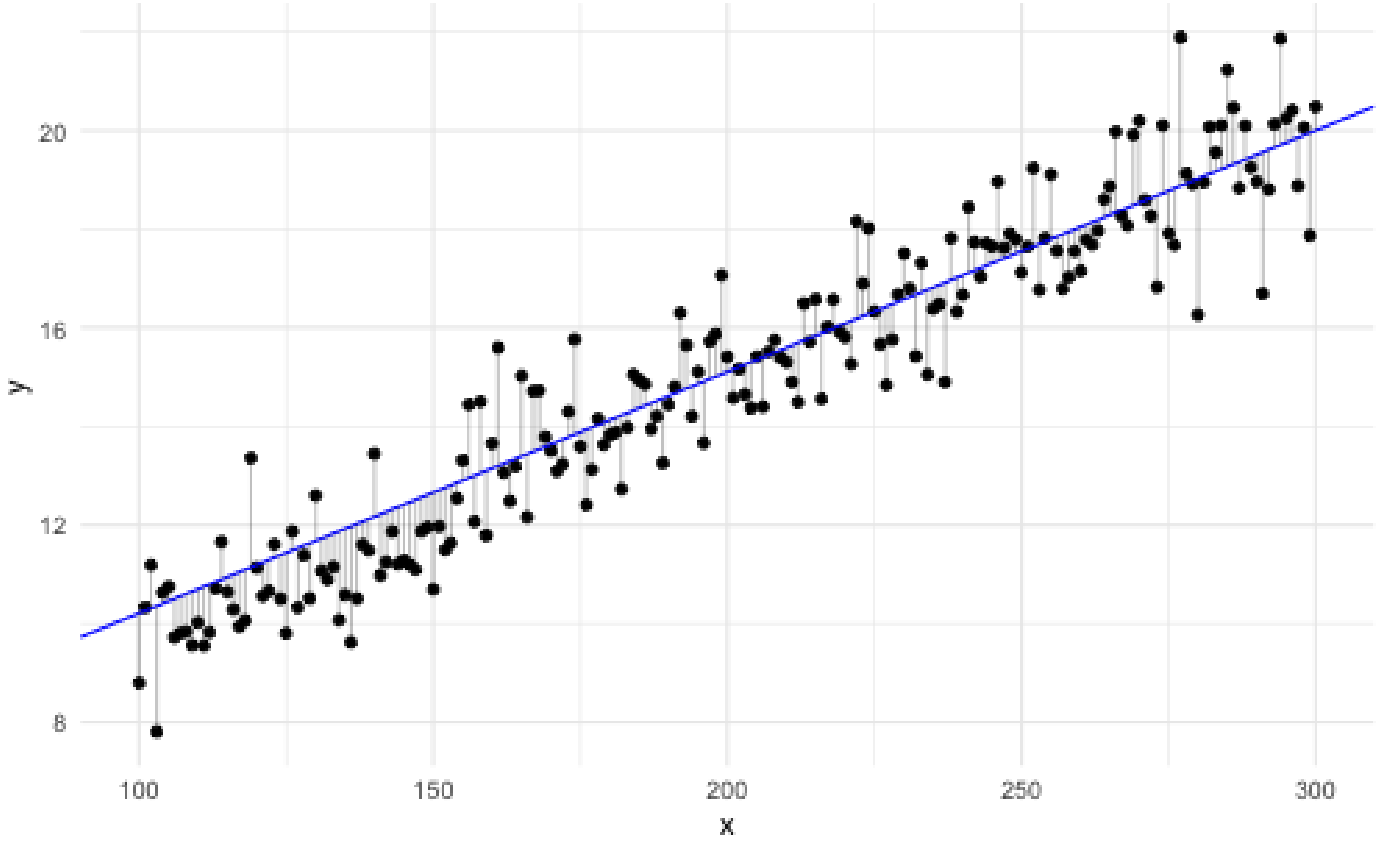
Method of least squares



Method of least squares



Method of least squares



Method of least squares

The error is

$$Y_i - (\beta_0 + \beta_1 X_i)$$

so we want to minimize

$$Q = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

This can be done in 2 ways:

- Numerical search procedure
- Analytical procedures

Method of least squares

Since we have

$$Q = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

that means that Q is a function of β_0 and β_1.

To estimate β_0 and β_1 we can take the partial derivatives of Q with respect to β_0 and β_1
.

Method of least squares

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \rightarrow \sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} X_i = 0 \quad (1)$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i = 0 \rightarrow \sum_{i=1}^{n} X_i Y_i - \beta_0 \sum_{i=1}^{n} X_i - \beta_1 \sum_{i=1}^{n} X_i^2 = 0 \quad (2)$$

Method of least squares

$$(1) \rightarrow n\beta_0 = \sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i \rightarrow \beta_0 = \sum_{i=1}^{n} \frac{Y_i}{n} - \beta_1 \sum_{i=1}^{n} \frac{X_i}{n}$$

So $b_0 = \bar{Y} = \beta_1 \bar{X}$.

Method of least squares

And we will see in chapter 2 that

we can rearrange the terms in (2) that

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Properties of Least Squares Estimators

- **Unbiased property** $E\{b_1\} = \beta_1$ and $E\{b_0\} = \beta_0$

- **Variance** it can be shown that b_0, b_1 gives the minimum variance in the group of linear and unbiased estimators.

These two points are part of the Gauss-Markov theorem

Point estimator of Mean Response

Given b_0 and b_1 of the parameters in the regression function

$$E\{Y_i\} = \beta_0 + \beta_1 X_i$$

We estimate the regression as follows

$$\hat{Y} = b_0 + b_1 X$$

Also $\hat{Y}$ is unbiased with minimum variance

Which means that

$$\hat{Y}_i = b_0 + b_1 X_i, \quad i = 1, \ldots, n$$

Residuals

The ith residual is denoted by e_i and is defined as

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

For the regression model

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

the residual e_i is defined as

$$e_i = Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i$$

Residuals

$\varepsilon_i = Y_i - E\{Y_i\}$ is the vertical deviation of $Y_i% from the unknown true regression line

$e_i = Y_i - \hat{Y}_i$ is the vertical deviation of Y_i from the fitted value $\hat{Y}_i$ on the estimated regression line and is thus known

Residuals are useful for studying whether a given regression model is appropriate for the data at hand (chapter 3)

Properties of the fitted regression line

- Sum of residuals is zero

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

- The sum of the squred residuals is minimum.

- $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

Properties of the fitted regression line

- $\sum_{i=1}^{n} X_i e_i = 0$
- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$
- The regression line goes through $(\hat{X}, \hat{Y})$

Estimationn of Error Terms Variance

The variance σ^2 of ε_i needs to be estimated to obtain an indication of the variability of the probability distribution of Y

The variance of a single population is estimated by sample variance s^2

$$s^2 = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{(n-1)}$$

Estimationn of Error Terms Variance

Similar to estimators for σ^2 for the regression model we have

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

since Y_i come fromo different distributions the sum of squares SSE is

$$SSE = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} e_i^2$$

Now we know that SSE has $n-2$ degrees of freedom.

Estimationn of Error Terms Variance

Two degrees of freedom are lost because both β_0 and β_1 need to be estimated in obtaining the estimated means $\hat{Y}_i$, hence

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

where MSE = error mean square

It can be shown that $E\{s^2\} = E\{MSE\} = \sigma^2$ of the regression model

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

Confidence intervals

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of n observations from a normal population with mean μ and standard deviation σ.

The sample mean is: $\bar{Y} = \dfrac{\sum_{i=1}^{n} Y}{n}$

The sample sd is:

$$s = \left[\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} \right]^{1/2}$$

We then get that

$$\bar{Y} \sim \mu_{\bar{Y}} = \mu \quad \& \quad \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

Confidence intervals

Thus the estimates standard deviation is $s\{\bar{Y}\} = \frac{\sigma}{\sqrt{n}}$

We define $\dfrac{\bar{Y} - \mu}{s\{\bar{Y}\}}$ t-distributed with $n - 1$ degrees of freedom

The confidence interval for μ are

$$\bar{Y} \pm t(1 - \alpha/2, n - 1)s\{\bar{Y}\}$$