STA302 - Lecture 3

Cedric Beaulac

May 14, 2019

Introduction

Today's plan

- Today we introduce the linear regression model
  - Introduction of linear relationship
  - Matrix notation for statistics
  - Least squares error formulation
  - Sum of squares revisited

Linear regression

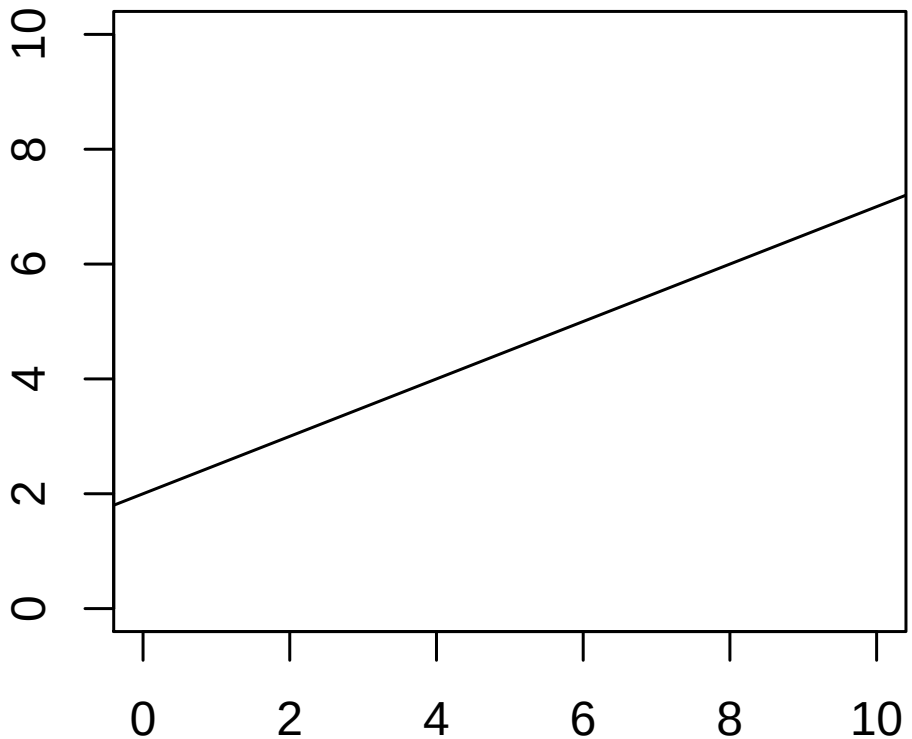
Regression is here!

- So far we have spend some time on test and models where the predictor is a categorical variable.
- We have compare 2 groups with t-test and multiples groups with ANOVA.
- When the predictor is continuous we cannot stratify the population into groups thus we cannot use these tests.
- The simplest model for continuous predictor is the linear regression.

Linear relationship

- In linear there is a line (straigth) $y = a + bx$

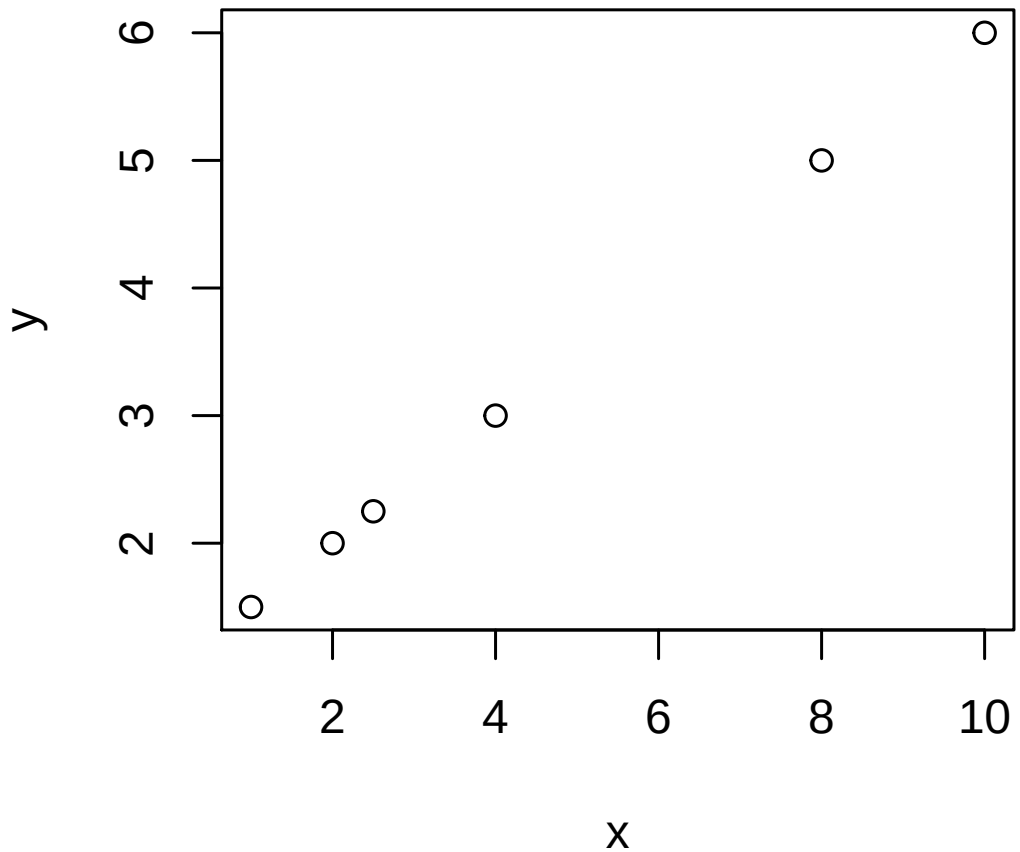
```
plot(x=NA,type='n',ylim=c(0,10),xlim=c(0,10),xlab="",ylab='')
abline(a=2,b=0.5)
```



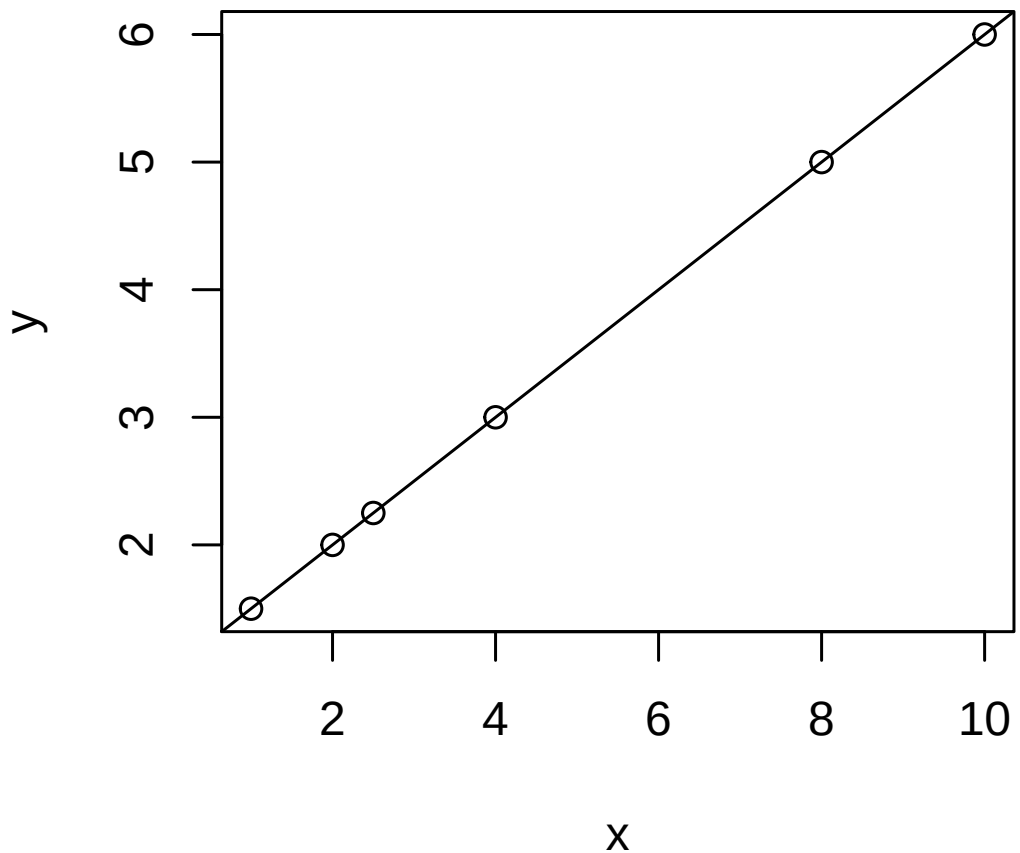
Linear relationship

- $y = \beta_0 + \beta_1 x$ is the type of relationship we are interested in.
- It's easy if the points are already aligned.
- Solve a linear equation system (high school maybe ?)

```
plot(x,y)
```

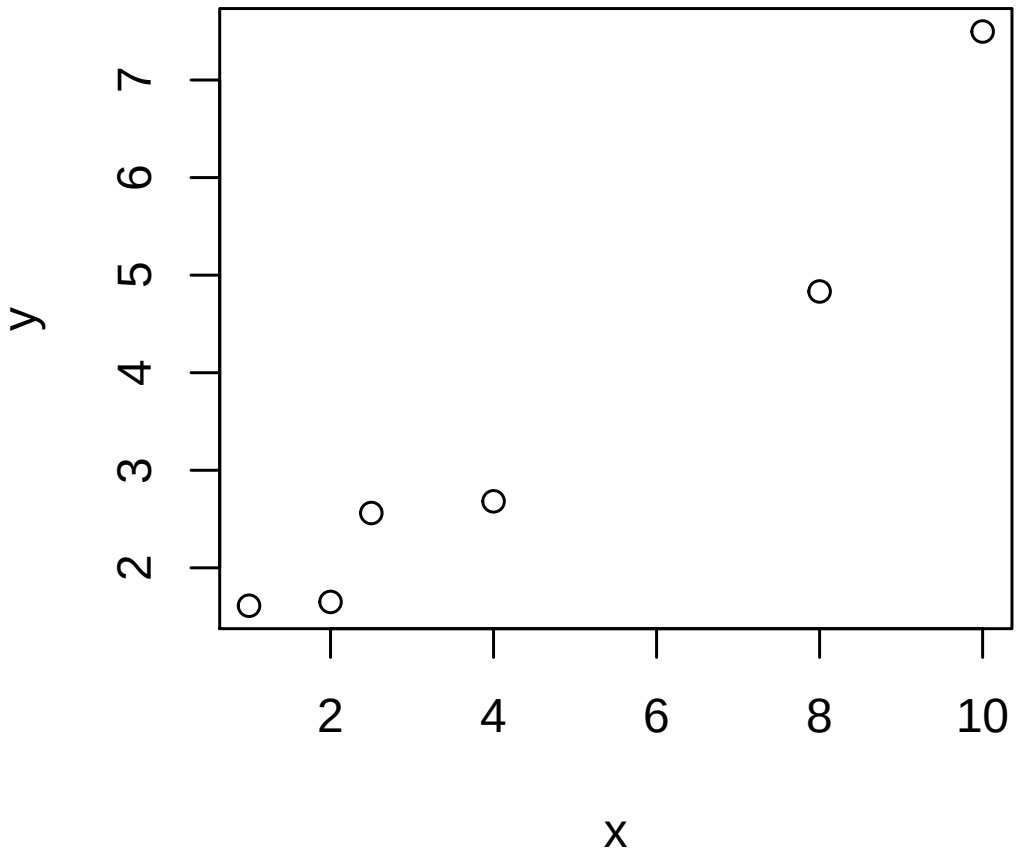


```
plot(x,y)
abline(1,0.5)
```



Linear relationship

- That's usually not what we observe with real-life data:
- You guessed it, we expecte some variability in the data!
- That's why we're here!


y
2
3
4
5
6
7
2
4
6
8
10
x

Linear relationship

- $y = \beta_0 + \beta_1 x$
- We will spend the next month on these type of models.
- How can we fit the best linear model to explain the relationship between y and x observed in the data ?

Matrix notation

Linear and matrix algebra

- Much more compact notation.
- Computationnaly efficient (matrix multiplication above all).
- Great generalzation properties.
- It's pretty.

Linear and matrix algebra

- From now on we will matrix notation.
- You will need to be somehow comfortable with matrix calculus.
- Google is your friend (so is Wikipedia).

Matrix notation for random variables

- We will introduce quickly how to compute Expectation, Variance of random vectors.
- It is important to respect the dimensions.
- Properties of Expectation and Variance are respected.

Matrix notation for random variables

$$\mathbf{A} = \begin{bmatrix} a_{11}, a_{12} \\ a_{21}, a_{22} \\ a_{31}, a_{32} \end{bmatrix}, \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

, where

$$E[\mathbf{x}] = \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, Var[\mathbf{x}] = \Sigma = \begin{bmatrix} \sigma_1, \sigma_{12} \\ \sigma_{12}, \sigma_2 \end{bmatrix}$$

Matrix notation for random variables

- As a small example, we will look at $\mathbf{Ax} + \mathbf{c}$

$$\mathbf{Ax} + \mathbf{c} = \begin{bmatrix} a_{11}, a_{12} \\ a_{21}, a_{22} \\ a_{31}, a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + c_1 \\ a_{21}x_1 + a_{22}x_2 + c_2 \\ a_{31}x_1 + a_{32}x_2 + c_3 \end{bmatrix}$$

Matrix notation for random variables : Expectation

- For univariate case $\mathbf{E}[ax + b] = a\mathbf{E}[x] + c = a\mu + c$
- Here we expect $\mathbf{E}[\mathbf{Ax} + \mathbf{c}] = \mathbf{A}\mu + \mathbf{c}$.

$$\mathbf{E}[\mathbf{Ax} + \mathbf{c}] = \mathbf{E}\begin{bmatrix} a_{11}x_1 + a_{12}x_2 + c_1 \\ a_{21}x_1 + a_{22}x_2 + c_2 \\ a_{31}x_1 + a_{32}x_2 + c_3 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{E}[a_{11}x_1 + a_{12}x_2 + c_1] \\ \mathbf{E}[a_{21}x_1 + a_{22}x_2 + c_2] \\ \mathbf{E}[a_{31}x_1 + a_{32}x_2 + c_3] \end{bmatrix} = \begin{bmatrix} a_{11}\mu_1 + a_{12}\mu_2 + c_1 \\ a_{21}\mu_1 + a_{22}\mu_2 + c_2 \\ a_{31}\mu_1 + a_{32}\mu_2 + c_3 \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}, a_{12} \\ a_{21}, a_{22} \\ a_{31}, a_{32} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \mathbf{A}\mu + \mathbf{c}$$

Matrix notation for random variables : Variance

- For univariate case $\mathbf{V}[ax + b] = a^2\mathbf{V}[x] = a^2\mu$
- Here we expect, it's a bit more complicated. We are not only interested in the vector of variances, but also all the covariances.
- So for: $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{c}$, intuitively we would have something like $\mathbf{A}^2\Sigma$, but $\mathbf{A}^2$ here is wrong!
- The correct equivalent is $\mathbf{A}\Sigma\mathbf{A}^T$. If you hesitate between $\mathbf{A}\Sigma\mathbf{A}^T$ and $\mathbf{A}^T\Sigma\mathbf{A}$ looks at the dimensions.

Matrix notation for random variables : Variance

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{c} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + c_1 \\ a_{21}x_1 + a_{22}x_2 + c_2 \\ a_{31}x_1 + a_{32}x_2 + c_3 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

$$\mathbf{V}[\mathbf{z}] = \begin{bmatrix} \mathbf{V}[z_1], \mathbf{Cov}[z_1, z_2], \mathbf{Cov}[z_1, z_3] \\ \mathbf{Cov}[z_1, z_2], \mathbf{V}[z_2], \mathbf{Cov}[z_2, z_3] \\ \mathbf{Cov}[z_1, z_3], \mathbf{Cov}[z_2, z_3], \mathbf{V}[z_3] \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}^2\sigma_1 + a_{12}^2\sigma_2 + a_{11}a_{22}\sigma_{12}, a_{11}a_{21}\sigma_1 + a_{11}a_{22}\sigma_{12} + a_{12}a_{21} \\ \dots \\ \dots \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}, a_{12} \\ a_{21}, a_{22} \\ a_{31}, a_{32} \end{bmatrix} \begin{bmatrix} \sigma_1, \sigma_{12} \\ \sigma_{12}, \sigma_2 \end{bmatrix} \begin{bmatrix} a_{11}, a_{21}, a_{3,1} \\ a_{12}, a_{22}, a_{32} \end{bmatrix} = \mathbf{A}\Sigma\mathbf{A}^T$$

- Want to try it out ?

Matrix notation for random variables : Derivatives

- Many different scenarios (https://en.wikipedia.org/wiki/Matrix_calculus)
- Might go deeper in futur lectures.
- For now we will keep it simple and introduce it as needed.

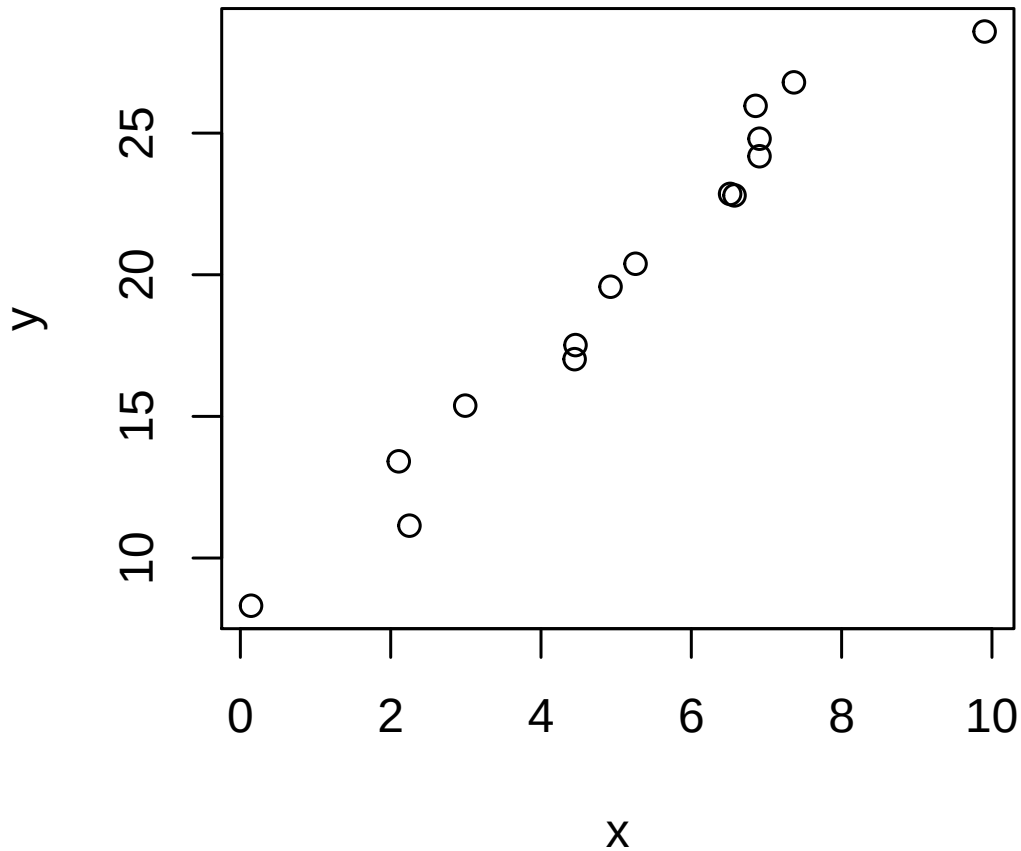
Linear Regression

The set up

- We have a vector of n predictors $\mathbf{x} = [x_1, ..., x_n]$
- We also have n associated reponse variables $\mathbf{y} = [y_1, ..., y_n]$
- We believe a linear relationship exist between x and y : $y = \beta_0 + \beta_1 x$.
- We want to select (estimate) the parameters β_0 and β_1 that better reflects the data.

```
##  [1] 2.2497472 6.5162055 0.1419573 7.3653283 4.4473090 6.5747553 9.9
##  [8] 2.9918667 6.8550410 4.9250690 2.1072009 4.4588271 5.2576737 6.9
## [15] 6.9072560
```

```
##  [1] 11.143938 22.854463  8.316969 26.792725 17.020152 22.801908 28.
##  [8] 15.380902 25.963459 19.578404 13.413609 17.517250 20.388415 24.
## [15] 24.184177
```


y
10
15
20
25
0
2
4
6
8
10
x

Least squares

The least squares line of best fit

- One way to decide on our estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ would be to settle on values that make it so $\hat{\beta}_0 + \hat{\beta}_1 x_i$ is close to $y_i \ \forall i$.
- Usually we say β is the real parameter and $\hat{\beta}$ is its estimate.
- $\hat{\beta}_0 + \hat{\beta}_1 x = \hat{y}$ is our predicted response given the predictor x .
- We have can produce a prediction for all predictor observed : $\hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{y}_i \ \forall i$. These are the *fitted* values.

The least squares line of best fit

- With a good model we expect the prediction error $e_i = \hat{y}_i - y_i$ to be small. (Well that's one way to define a good model!)
- Since we only care about distance and not direction and because absolute value are inconvenient we will be looking at the squared error : $e_i^2 = (\hat{y}_i - y_i)^2$

The least squares line of best fit

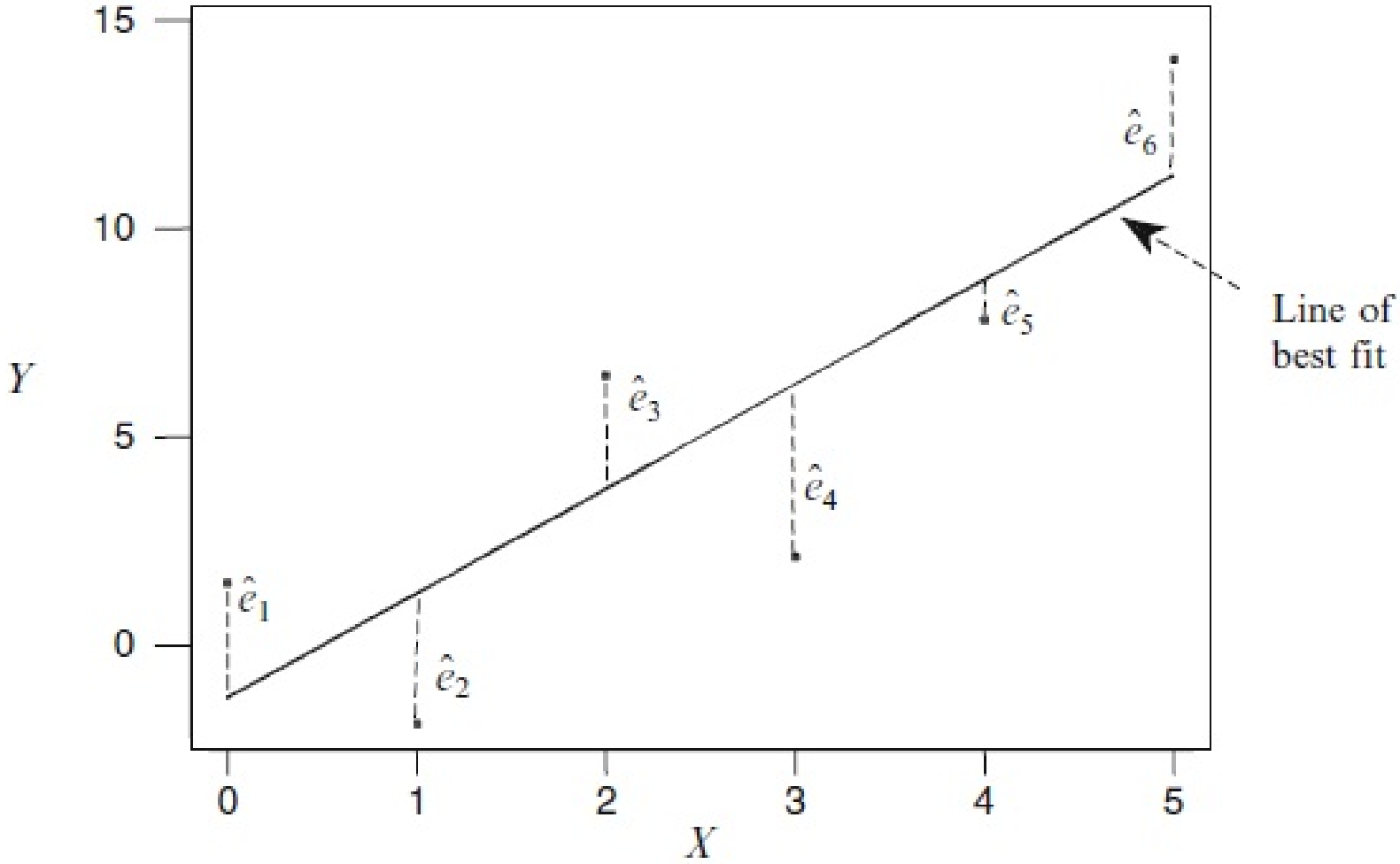


Figure 1: Stolen without shame in the Textbook

The least squares line of best fit

- Our first proposed linear model is to select $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimizes the sum of squared errors : $\sum_{i=1}^{n} \hat{e}_i^2$.
- Let's write everything in matrix notation and solve the optimization problem!

The least squares line of best fit

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ . \\ . \\ \hat{y}_n \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ . \\ . \\ x_n \end{bmatrix}$$

- $\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$ which can be written in an even more compact fashion.

The least squares line of best fit

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ . \\ . \\ \hat{y}_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}, \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

- $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$
- $\mathbf{e}_i = \mathbf{y} - \mathbf{X}\hat{\beta}$

The least squares line of best fit

- Remember we want to minimize $\sum_{i=1}^{n} e_i^2 = \mathbf{e}^t\mathbf{e}$ with respect to β.
- Here we are take the derivative of a scale with respect to a vector. This should span a vector of partial derivative.

$$\frac{d}{d\hat{\beta}}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$$

- RECALL : $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

The least squares line of best fit

$$\frac{d}{d\hat{\beta}}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$$

$$\frac{d}{d\hat{\beta}}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\hat{\beta} - (\mathbf{X}\hat{\beta})^T\mathbf{y} + (\mathbf{X}\hat{\beta})^T\mathbf{X}\hat{\beta}) = 0$$

$$\frac{d}{d\hat{\beta}}(\mathbf{y}^T\mathbf{y} - 2\hat{\beta}^T\mathbf{X}^T\mathbf{y} + \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta}) = 0$$

$$(0 - 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\beta}) = 0$$

$$2\mathbf{X}^T\mathbf{y} = 2\mathbf{X}^T\mathbf{X}\hat{\beta}$$

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\hat{\beta}$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \hat{\beta}$$

- Ouf!

The least squares line of best fit

- So, the fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$.
- $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ this is often referred as the hat matrix (it puts the hat on y).
- This matrix as some great properties you will discover in practice problems today.

The least squares line of best fit

- How do we handle variability ?
- How good are our estimate of β_0 and β_1 ?
- Hummmmmmm. . . (next lecture)

How's our model doing ?

- What is a *good* a model ?
- It would be interesting to check if your model is good at *explaining* the response.
- Let's revisit the concept of sum of squares and apply it to this simple linear regression model.

Sum of squares revisited

Sum of squares revisited

- Sums of squares are important metrics of variability.
- They are the fundamental blocs of ANOVA and are usefull in the context of regression.
- We introduced them (too quickly maybe ?) last Thursday. So let's speak about them today again.

Sum of squares revisited

- We expect that data is variable.
- If a predictor is usefull, we would like the predictor to *explained* parts of the variability.
- If a predictor is usefull, we should be able to use it to produces *better* predictions.

Sum of squares revisited

- We'll think of $\bar{y}$ as the *base* prediction.
- For ANOVA, given the group/treatement t our prediction for an observation in group t is $\bar{y}_t$.
- For regression, the prediction for observation x_i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Sum of squares revisited

- SST $= \sum_{i=1}^{n}(y_i - \hat{y})^2$ is the sum across all observations of squarred distance between the observation and the base prediction (the mean)
- SSG $= \sum_{i=1}^{n}(\hat{y}_i - \hat{y})^2$ is the sum across all observations of squarred distance between the regression prediction and the base prediction.
- It is the *explained* variation. How much of our explanation takes us away from the base prediction.
- SSE $= \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ is the sum across all observations of squarred distance between the observation and the regression prediction.
- It is the *unexplained* variation. How much our model prediction is away from the true observation. This distance is unexplained by the model.

Sum of squares revisited

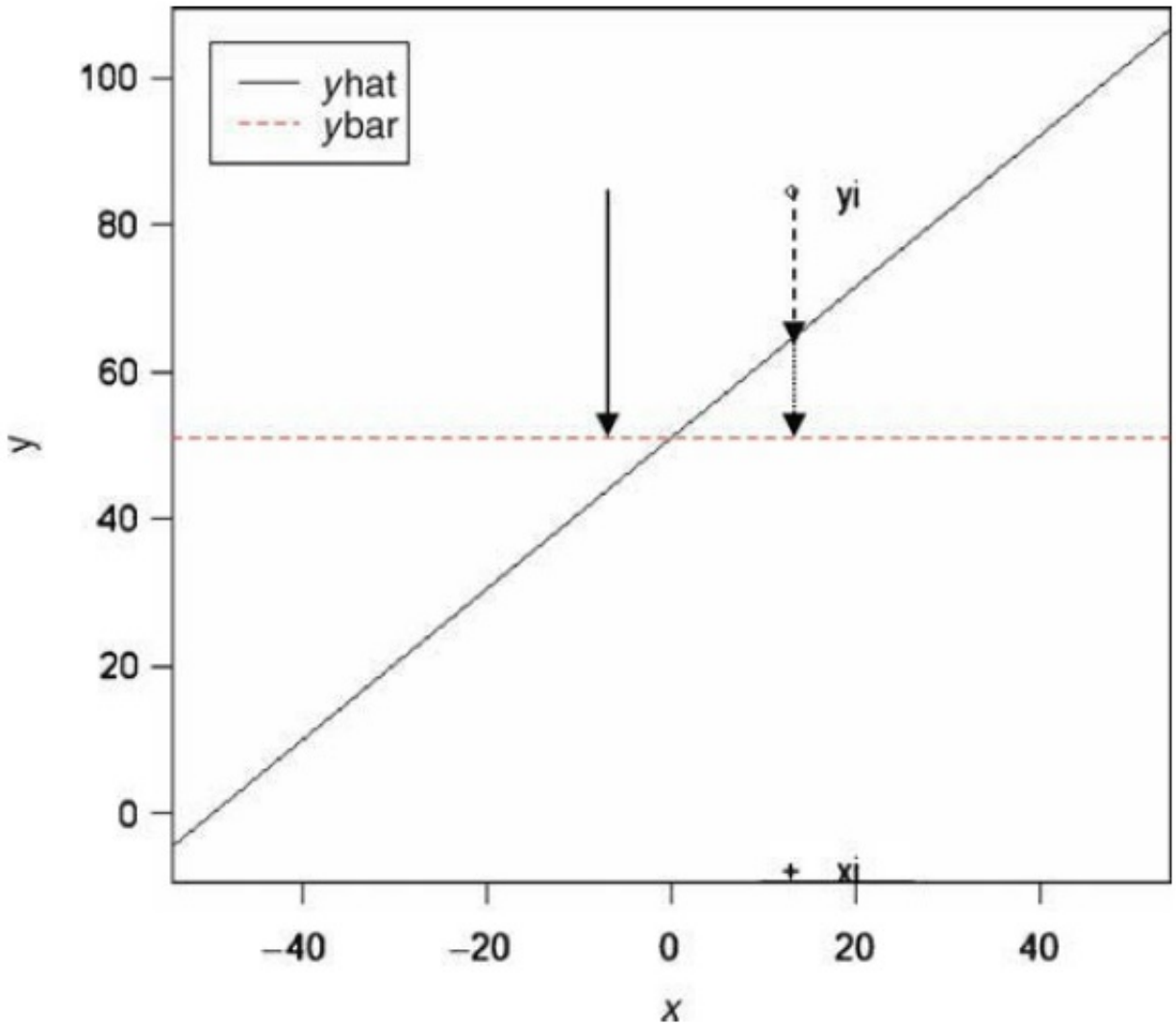


Figure 2: Stolen without shame in the Textbook

r-squared

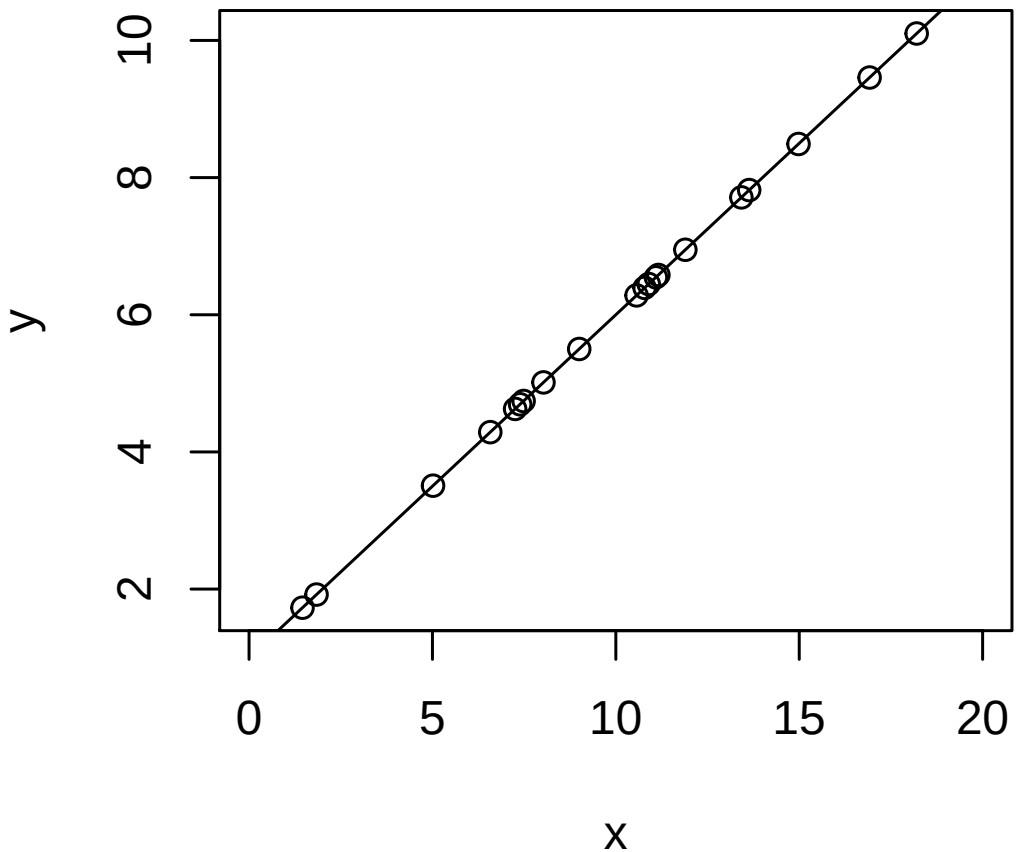
- R^2 (r-squared) is the coeffeficent of determination.
- It is one diagnostic tools to check how *good* our model is.

$$R^2 = \frac{SSG}{SST} = 1 - \frac{SSE}{SST},\ 0 \leq R^2 \leq 1$$

- The closer R^2 is from 1, the better the fit is, the closer it is to 0 the worst the fit is.

r-squared : sanity check

- If the observations are sitting directly on the fitted line, this is the best possible fit.



r-squared : sanity check

- If the observation are sitting directly on the fitte line, this is the best possible fit.
- In this case $y_i = \hat{y}_i$ and thus SST = $= \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \hat{y})^2 =$ SSG which implies SSG/SST = 1!
- Similarly if $y_i = \hat{y}_i$, then SSE $= \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = 0$, whici implies 1 - SSE/SST = 1 - 0 = 1!

Conlusion

- The least squares line is the simplest formulation of linear regression.
- It requires no assumption and lead to easy to obtain estimate.
- We can also asses how *good* is the model at predicting using R^2.

Practice problem

- Optimize for β_0 and β_1 in non-matrix formulation i.e:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

- You can get the solution in A Modern Approach to Regression with R ch.2
- With $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, the hat matrix.
- Show that $H^T = H$, $HH = H$ and that $H\mathbf{X} = \mathbf{X}$.
- For $M = H - I$, show that $M^T = M$.

External ressources

- Wikipedia (https://en.wikipedia.org/wiki/Multivariate_random_variable)
- Linear Models with R ch.2
- A Modern Approach to Regression with R ch.2
- A Modern Approach to Regression with R ch.5

Break time

- Congratulation!
- Now let's take a break and do R coding!