Case Study: R for Statistics

Intro to R for Researchers

[1]Quantitative Science Pillar
Stanley Manne Children's Research Institute
Ann & Robert H. Lurie Children's Hospital of Chicago

Outline

Getting started checklist

- Can open RStudio
- Have access to the class folder with class material
- Can specify a location with the class folder

Case for cleaning data and pointers

- Ensure relevant data is accessible to software
- Detect and correct errors
- Portability
- Assess limitations to analysis plan

A few pointers

- Study units e.g. patient, encounter, hospital. uniquely identifiable
- Variable in columns, study units information in rows
- Rows in dataset are unique and identifiable
- Columns contain unique information on the study unit
- Consistent and speciifc method of denoting missing data
- Consistent data format, dates denoted mm/dd/yyyy (and not dd/mm/yy or dd/mm/yy)
- No empty rows/no empty columns
- Always have each dataset having a data dictionary

Exercise in data cleaning (1)

From the dataset pick out three issues

StudyID	1786	1953	1431	1650	1814
DOB	4/24/21	7-15-21	Mar/9/2021	9/8/20	7/11/21
Race	White	Other	white	Black or African American	Other
Insurance	Missing	Private	Public		Public

R functionality for cleaning

- Basic object information -`dim()`, `attributes()`, `summary()`, `nrow`, `ncol()`, `names()`, `head()`, `tail()`
- cross-tabulation –`ftable()`, `table()`
- Reshaping data – `t()`, `reshape()`, `attach()`, `reattach()`
- joining datasets – `merge()`
- Subsetting datasets `[]`, `subset()`
- Recoding `recode()`, `cut()`
- Sorting `order()`

Exercise in data cleaning NCSCH Data (1)

The National Survey of Childhood Health (NSCH) is an annual survey to examines the physical and emotional health of children ages 0-17 years of age. The survey employs a two stage cluster sample, (1) A household is randomly selected then (2) a randomly selected child in the household has their physical and emotional recorded. The 2020 data is available in two tables (a) Screener table with information on the household (b) topical table with information on a randomly selected child. We are interested in creating a database to study childhood depression and its association with the absence of one or more parents. Variables of interest in the data table are:

Variable Name	Description	Response Code
`ACE3`	Parent/Guardian Divorced?	1=Yes, 2=No
`ACE4`	Death of a parent/Guardian?	1=Yes, 2=No
`ACE5`	Parent/Guardian in jail or prison	1=Yes, 2=No
`K2Q32A`	Has had diagnosis of depression?	1=Yes, 2=No

The child identifier is `HHID`

Exercise in data cleaning NCSCH Data (2)

- The topical dataframe is named `ncsh.t`, we can learn more about it with `summary, dim, table`
- Create a subset of data with adverse childhood events (`HHID, ACE3, ACE4, ACE5, K2Q32A`)

```
ncsch.tab1 <- ncsch.t[ , c("ACE3","ACE4", "ACE5", "HHID", "K2Q32A")]
```

- Further refine this data to exclude any child with missing data on depression:

```
ncsch.tab2 <- ncsch.tab1[ !is.na(ncsch.tab1$K2Q32A), ]
```

- Create a new variable `PACE` such that `PACE` = 1 if a parent is missing due divorce/death/jail and zero if otherwise using the function `ifelse()`:

```
PACE <- ifelse((ncsch.tab1$ACE3==1) &
   (ncsch.tab1$ACE4==1) &
   (ncsch.tab1$ACE3==1), 1, 0)

 ncsch.tab2$PACE <- PACE
```

Recoding variables

- Can create/recode variables with `ifelse(condition met?, option 1, option 2)`
- Recode with function `recode()` in package `car`. In the NCSCH 2020 dataset `FPL_I6` is a measure of poverty and takes on value 50-400, we would like to recode it to

$$\texttt{FPL.new} = \begin{cases} 1 & \text{if } \textit{FPL}\_6 < 100 \\ 2 & \text{if } 100 \leq \textit{FPL}\_6 \leq 199 \\ 3 & \text{if } \textit{FPL}\_6 \geq 200 \end{cases}$$

```
#load library car
library(car)

ncsch.tab4 <- within(ncsch.t,{
   FPL.new <- recode(FPL_I6, "50:100 = '<100'; 100:199 = '100-199';  200:400 = '>200'")
})
```

Randomization – Example

You plan to perform a study to investigate the effect of 4 types of feed additives (A, B, C, D) in the diet of mice. The study plans on using 32 lab mice, 16 males and 16 females as shown. Treatment assignment is shown on the right. Class discussion: what is the problem with this treatment assignment

ID	Sex	Weight	Tmt
1	M	276	A
2	M	292	A
3	M	297	A
4	M	345	A
5	M	360	A
6	M	369	A
7	M	270	A
8	M	284	A
9	M	303	B
10	M	332	B
11	M	301	B
12	M	250	B
13	M	263	B
14	M	285	B
15	M	285	B
16	M	180	B
17	F	278	C
18	F	273	C
19	F	278	C
20	F	276	C
21	F	278	C
22	F	270	C
23	F	276	C
24	F	276	C
25	F	276	D
26	F	273	D
27	F	275	D
28	F	271	D
29	F	273	D
30	F	282	D
31	F	269	D
32	F	268	D

Randomization – Example continued

- A completely random assignment, we can use `sample()` to randomly assign the treatment

```
Tmt <-  c("A","B","C","D")
mice$Tmt.Rand1.a <- sample(Tmt, 32, replace=TRUE)
```

- What are the issues with this randomization approach?

Stratified Randomization – Example continued

- A random assignment within sex (stratified by sex),

```
mice$Tmt.Str.Rand <- c(sample(Tmt, 16, replace = TRUE),
   sample(Tmt, 16, replace = TRUE))
```

Still have issue with balance (tabulate assignment with sex)

Randomized block (within sex)– Example continued

- The function `runif()` generates random numbers in the interval 0-1,
- Randomization blocks of size four (stratified by sex) can help achieve balance:

```
Blocks <- rep(1:8, each=4 ) #create blocks of size four
r.unif <- runif(32)
Tmt.all <- rep(c("A","B","C","D"), 8)
mice$Tmt.Blk.rand <- Tmt.all[order(Blocks, r.unif)]
```

Check for balance (tabulate assignment with sex)

Bivariate Association

- Scatter plot (scatter plot matrix) – This is plot of observations of Y versus those of X. Use `plot(X, Y )`
- A scatterplot matrix is a matrix of all possible pairwise scatter plots of a group of variables. Use `pairs()` or `splom` in library `lattice`
- Correlation coefficient (Pearson, Kendall, Spearman rank) Pearson measures linear association, Spearman and Kendall are rank correlation, use `cor(x, y, method = c('pearson', 'kendall', spearman'))` and `cor.test(x, y, method = c('pearson', 'kendall', spearman'))`

Regression Analysis

- Regression – Finds the functional relationship between a response variable Y and a set of predictors $X_1, X_2, \ldots, X_p$ via a mathematical function

$$Y = f(X_1, X_2, \ldots, X_p, \beta_0, \beta_1, \ldots, \beta_q) + \epsilon$$

  ϵ is the residual/random error
- The process of estimation of a response by a function is a multi-step process: Data collection $\Rightarrow$ Numerical and visual exploration $\Rightarrow$ Model selection $\Rightarrow$ Model estimation $\Rightarrow$ Model validation $\Rightarrow$ Inference

Definition of Linear and Non-linear Models

- Linear Model – Linear models are linear on the set of regression coefficients

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \\ Y &= \beta_0 + \beta_1 X_1^2 + \epsilon \\ Y &= \beta_0 + \beta_1 \log(X_1) + \beta_2 \exp(X_1^2) + \epsilon \end{aligned}$$

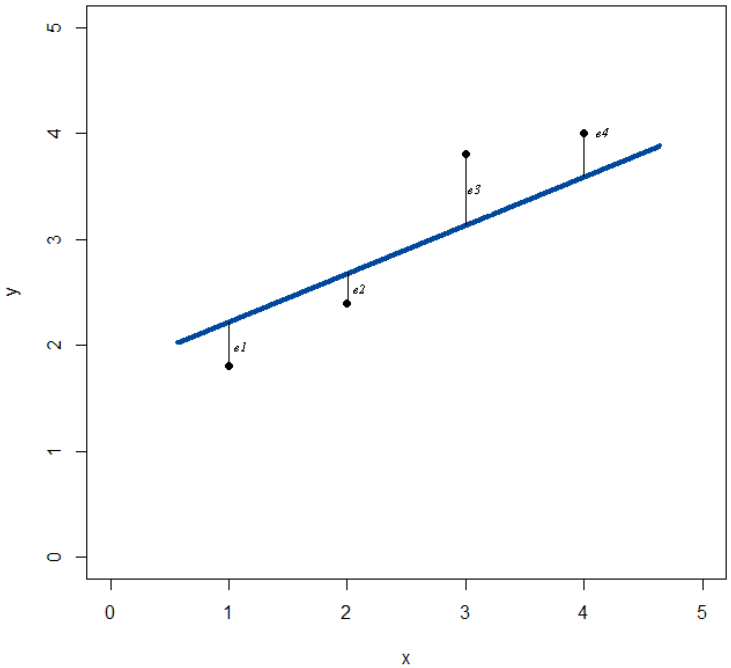
- Non-linear Model – Non-linear on regression coefficients

$$\begin{aligned} Y &= \frac{\beta_0}{1 + \beta_1 X_1} \\ Y &= \beta_0 + X_1^{\beta_1} \end{aligned}$$

Non-linear equations can be linearized by transforming the variables,

$$Y = \beta_0 X_1^{-\beta_1}, \; Y = \frac{\beta_0 X_1}{\beta_1 + X_2}$$

Simple linear Regression Model



- Linear model with one predictor

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

β_0 and β_1 are regression coefficient; ϵ is assumed to be a random error or measurement error, having zero mean and an unknown variance σ^2

- Given data with n observations from the above model can be written

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, \ldots, n$$

- Various of estimating method coefficients β_0, β_1, we use the least squares method , available in the R function `lm`

Example: Berkeley Guidance Study (BGS)

The study enrolled children born in Berkeley, between January 1928 and June 1929 and then measured them periodically until age 18. The dataset is available in the `alr4` library and is named `BGSall`, the variables in the dataset are given below

Variable	Description
Sex	0 = males, 1 = females
WT2	Age 2 weight (kg)
HT2	Age 2 height (cm)
WT9	Age 9 weight (kg)
HT9	Age 9 height (cm)
LG9	Age 9 leg circumference (cm)
ST9	Age 9 strength (kg)
WT18	Age 18 weight (kg)
HT18	Age 18 height (cm)
LG18	Age 18 leg circumference (cm)
ST18	Age 18 strength (kg)
BMI18	Body Mass Index, $WT18/(HT18/100)^2$, rounded to one decimal.
Soma	Somatotype, a 1 to 7 scale of body type.

BGS Example continued

- Draw a scatterplot matrix of the data, use different colors for males and females, what conclusion do you draw about the variables,
  `pairs(BGSall[,-1], col=BGSall$Sex+1)`
- Estimate the correlation matrix for boys.
  `round( cor(BGSall[BGSall$Sex==0, -1]), 2 )`
- For only boys, fit a simple linear regression model regressing `Soma` on `WT9` using the `lm` function
  - `boys.fit <- lm( Soma ~ WT9, data=BGSall[ BGSall$Sex==0, ])`
    `boys.fit <- lm( Soma ~ WT9, data=BGSall, subset = (Sex==0))`
  - Call `summary(boys.fit)`
- Use `coef` to obtain the coefficients
- Obtain the confidence intervals, predictions using `confint, predict`

Review - Composition of a significance test

- H_0 – Null hypothesis; the claim we seek evidence against,
- H_A – Alternative hypothesis; the claim we are trying to find evidence for. Can be one sided or two sided
- G(Sample data) – Test statistic; the evidence from the sample
- p-value – The p-value is the probability that the test statistics will take on a value as extreme or more extreme than that actually observed.
  - Small p-values are evidence against the null hypothesis
  - Large p-values support the null hypothesis
- α – Significance level; this is a fixed value which helps compare the p-value. **Hypothesis tests and confidence intervals for coefficients, fitted values and prediction use Student-t distribution which is dependent on the assumption that $\epsilon \sim NID(0, \sigma)^2$ holds.**

  A quick tool for doing this is a plot of residuals versus fit `plot(bgs.fit)`

Model Estimates

- Coefficients – $\hat{\beta}_0, \hat{\beta}_1$, call with `coef(bgs.fit)` and their confidence interval estimates with `confint(bgs.fit)`
- Fitted values – $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ call with `predict(bgs.fit)`
- Residuals/error –$e_i = y_i - \hat{y}_i$ call with `resid(bgs.fit)`
- standard error of the regression (SER), –

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

call with `(summary(BGS.fit))$sigma`

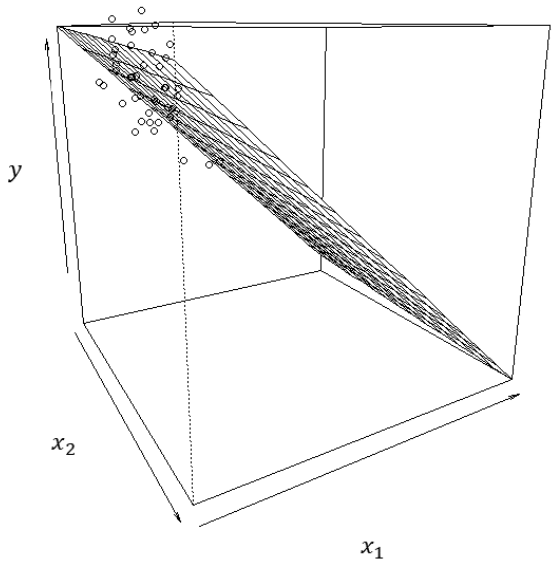
Some facts about the simple linear regression estimates

- $\hat{\beta}_0$ represents the response when $x = 0$; this interpretation can be made more meaningful by:
- transformation let ${y_i}^* = y_i - \bar{y}$, ${x_i}^* = x_i - \bar{x}$, then rewrite the LS regression line as

$$ {y_i}^* = \beta_1 {x_i}^* $$

- The distinction between response and explanatory variable is important (results are different if we interchange response and explanatory variable)
- The sign on β_1 is the same sign on the pearson correlation coefficient r
- The least squares line passes through the point $(\bar{x}, \bar{y})$
- The slope of the regression line is independent of the origin of co-ordinates
- R-square which is a measure of the proportion of the total variation in Y explained by X, is the square of the Pearson correlation coefficient

Multiple Linear Regression



- For a dataset with n observation, a multiple linear regression equation regressing Y on p predictors $X_1, \ldots, X_p$ can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \ i = 1, \ldots, n$$

where ϵ is the random variable with mean of zero and constant variance σ^2 which represents the measurement error.

- The R commands for a least square fit: `lm( y~ x1 + x2 + x3 ..., data=Data)`
- Other useful functionality:
  - Analysis of variance for model comparison

Example: BGS Data

Regressing Soma on WT9 and LG9

```
> BGS.fit2 <- lm(Soma~WT9 + LG9, data = BGSgirls)
> summary(BGS.fit2)

Call:
lm(formula = Soma ~ WT9 + LG9, data = BGSgirls)

Residuals:
     Min       1Q   Median       3Q      Max
-1.57125 -0.38297 -0.03819  0.36787  1.36617

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.49059    1.24348   1.199   0.2349
WT9          0.07148    0.03110   2.299   0.0246 *
LG9          0.03692    0.07391   0.499   0.6191
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6418 on 67 degrees of freedom
Multiple R-squared:  0.3843,    Adjusted R-squared:  0.3659
F-statistic: 20.91 on 2 and 67 DF,  p-value: 8.782e-08
```

p-value for overall F-test,
$H_0: \beta_1 = \beta_2 = 0\ vs$
$H_A: either \beta_1 \neq 0\ or\ \beta_2 \neq 0$

Regressing Soma on WT9

```
> summary(BGS.fit)

Call:
lm(formula = Soma ~ WT9, data = BGSgirls)

Residuals:
     Min       1Q   Median       3Q      Max
-1.62809 -0.37394 -0.03102  0.37507  1.38508

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.07405    0.42406   4.891 6.45e-06 ***
WT9          0.08553    0.01319   6.483 1.19e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6382 on 68 degrees of freedom
Multiple R-squared:  0.382,     Adjusted R-squared:  0.3729
F-statistic: 42.04 on 1 and 68 DF,  p-value: 1.193e-08
```

Class Exercise

Using the BGSgirls dataset

- Regress `Soma` on `WT9` and save the residuals to `soma.resid`
- Regress `LG9` on `WT9` and save the residuals to `LG9.resid`
- Draw a scatterplot of `soma.resid` versus `LG9.resid`. This plot is known as an added-variable plot.
- Regress `soma.resid` on `LG9.resid`. Compare the coefficient for `LG9.resid` with the coefficient for `LG` in the fit in the previous page and comment.

Including qualitative predictors

- In regression analysis, quantitative predictors (E.g. Sex, treatment, Race, Insurance, Blood Type) are known as factors
- The categories within a factor are known as levels
- Factors are represented numerically by dummy variables (E.g (0, 1), (-1, 1), (1, 2), differences in statistical software in how dummy variables are coded
- Factor rule - A factor with d-levels can be represented by at most $d - 1$ dummy variables (once you know whether which group an item is by examining $d - 1$ levels then you know which group the item belongs to)

Example of including qualitative predictors

The `pots` dataset is from a nutrition experiment to help measure the iron content from stews cooked with three different types of pots: aluminum, clay and iron.

```
> pot.data
   pot.type   fe
1        Al 1.77
2        Al 2.36
3        Al 1.96
4        Al 2.14
5        Cl 1.28
6        Cl 2.27
7        Cl 2.48
8        Cl 2.68
9        Ir 5.27
10       Ir 5.17
11       Ir 4.06
12       Ir 4.22
```

- Create dummy variable for each pot type: U1, U2, U3

$$U1 = \begin{cases} 1, & \text{if pot.type} = \texttt{Al} \\ 0, & \text{otherwise} \end{cases}$$

$$U2 = \begin{cases} 1, & \text{if pot.type} = \texttt{Cl} \\ 0, & \text{otherwise} \end{cases}$$

$$U3 = \begin{cases} 1, & \text{if pot.type} = \texttt{Ir} \\ 0, & \text{otherwise} \end{cases}$$

- Find the pot type group averages: hint use the function `ave`
- Regress `fe` on `U2, U3`, compare the coeffients with the group averages
- Regress `fe` on `pot.type`, compare the results of the fit with the previous bullet point's results
- Regress `fe` on U3, compare the coefficients

Data type Factors

- Can code vectors having distinct values as factors (i.e. group into distint categories)
- Advantages of this format is to allow you to label categories which helps in data summaries
- Enables you to change reference group in analyses involving categorical variables
- Helps keep track of categories of interest, e.g. missing information
- Helps order categories in a variable
- You can change a character vector or numeric vector by using `factor()` or changeback using `as.numeric`

It is always a good idea to code none numeric variables as factors: Exercise for the NSCH dataset, convert the ACE variables to factors and label them

Vizualizing quantitative Data

- Barchart (`barplot, lattice: barchart,`
- cleveland dot plot `dotchart, lattice: dotplot.` With both can group by independent variable
- pie-charts (something fun: try `pie(1:24, col=rainbow(24))`)
- When there is an associated indpendent variable, grouped histograms and boxplots can be used (grouped by the nominal variable)

Bivariate association with qualitative outcome

– Quantitative variable

- Difference in the mean in two populations with a two-sample t-test – `t.test(x, y, ...)`
- Analysis of variance `anova()`
- Kruskal-Wallis test `kruskall.test`

–Qualitative variable

- Chisquare test for independence `summary(xtabs (Freq ., ))`
- Fisher exact test
- Mantel haezel test

Backup slides

Definition

Study design refers to the methods and methodologies used in research to gather the data needed to achieve one or more research objectives
Flowchart showing types of quantitative study designs at: ▸ Link

Terminology Review

- Population – The collection of individuals or objects that one is interested in studying/generalizing/drawing valid conclusions
- Sample — A collection of members of the population whose characteristics are to be measured so as to generalize over the population
- Simple random sample — Individuals in the population are equally likely to be chosen for the sample; every possible sample has an equal chance to be chosen
- Random variable – Measurement from individuals in a population, could be qualitative or quantitative
- Population parameter – Quantities that determine a distribution of the random variable pertaining to the population
- Sample statistics – Estimates of population parameters determined from a sample
- Response and explanatory variables — Terminology used when studying the association between one or more variables, where one variable is the outcome of interest and the other variables influence the outcome. The outcome is the response variable. The other variables are the explanatory variables (predictors).

Key design consideration in a randomized control study

- **Treatment allocation or randomization**
- Blinding (reduce bias)
- **Power/Sample size consideration**, further considerations:
  1. endpoint
  2. target effect
  3. claim (superiority, non-inferiority)
  4. interim analysis/looks
  5. Duration of study
  6. Drop-out rate
  7. accrual rates

Sample size computations

- One sample to detect a population proportion: A simple random sample of Chicago adults will be used to estimate the proportion of adults with asthma. How large a sample such that the 95% confidence interval is within 15% of the true proportion?
- Two sample to detect a difference in population means: Suppose we wanted to find the sample size necessary to detect a difference in mean response of 20 units between two treatments with 90% power using a t-test (two-sided) at the .05 level of significance. We expect the population standard deviation of response to be about 60 units
- A new drug is expected to increase the response rate by 10% from the standard of care's response rate of 35%. A clinical trial will randomize patients with equal allocation to either the new treatment (treatment 1) or the standard treatment. How large a sample size is necessary to detect a clinically important difference with 90% power using a one-sided test at the .025 level of significance?

Estimates of Regression Coefficients

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1}$$

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

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

$$= Cor(X, Y) \cdot s_y / s_x$$

Parameter estimation by least squares

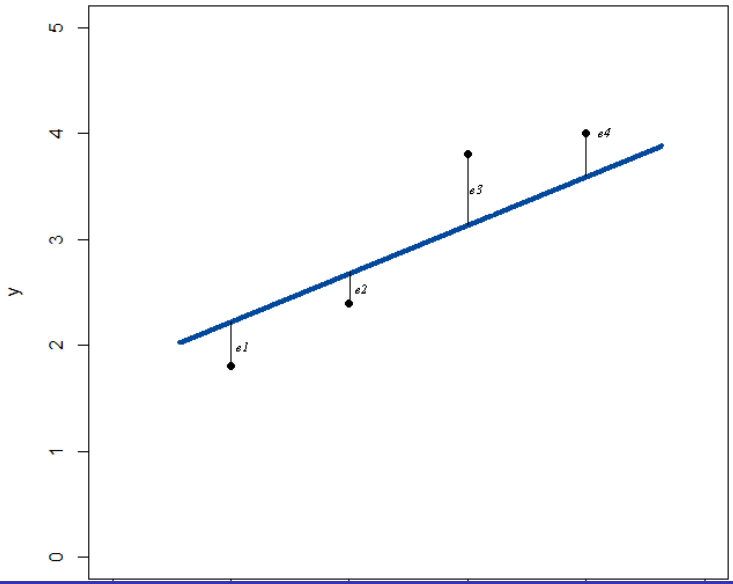
Least squares line, find the "the best-fitting line" by minimizing the sum of square errors

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

This reduces to solving a pair of simultaneous equations for β_0, β_1

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0$$



For Further Reading I

Weisberg

Applied Linear Regression.

Link

Harrell

Regression modeling strategies: with applications.

Link

NSCH Code Book

Census Bureau.

Link