Data Analysis using R

Dr. Kapil Kumar
Associate Professor, Department of Statistics
Chaudhary Charan Singh University, Meerut, India
kapilstats@gmail.com

Two-Day National Workshop on Data Analysis with R (20–21 February 2026)
Department of Economics, Chaudhary Charan Singh University, Meerut

21 February 2026 · Day 2 lecture

Where we left off

Day 1: *Introduction to R*

- Vectors and vectorised arithmetic; logical indexing `data[condition, cols]`
- Factors, matrices, lists and data frames; `str()` and `summary()`
- `read.csv()`, `na.strings`, CSV, RDS and Excel with relative paths
- `ifelse()`, loops, functions and `tapply()`

Loose ends we noticed `Region` has stray spaces, schooling has missing values, Titanic cabins are blank strings, and the median literacy score sat far above its mean. Today we deal with these first, then move from describing the data to testing and modelling.

Plan of the session

1. Cleaning the data
2. Descriptive statistics and grouped summaries
3. Visualisation with base R graphics
4. Statistical inference: t, ANOVA, χ^2, correlation
5. Regression for economists

By the end you can find and fix common data problems, summarise variables overall and by group, draw and read the standard graphs, choose and interpret a basic test, and fit, interpret and check a regression with log variables.

PART 1

Cleaning the data

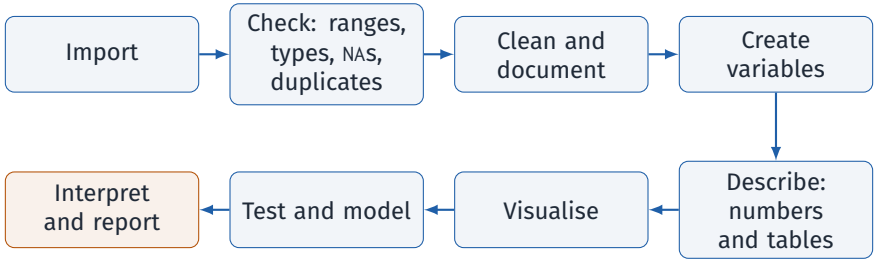
Packages, colours and the data

```r
# install.packages("lmtest")    # once
library(lmtest)           # Breusch-Pagan test
options(digits = 5)       # digits in printed output
col_main <- "#1F5FA8"     # plot colours
col_alt  <- "#B45309"
socio   <- read.csv("data/socio_economic.csv")
sales   <- read.csv("data/sales.csv")
titanic <- read.csv("data/titanic.csv",
                    na.strings = c("", "NA"))
```

- The same three files as on Day 1, read afresh so that this script runs on its own
- **`lmtest`** adds the Breusch–Pagan test; everything else is base R

Data note `socio` and `sales` are illustrative teaching data with unstated units; Titanic is a classic real data set.

The analysis workflow



Workshop insight Cleaning is not optional. Most wrong conclusions come from data problems that a summary, a plot or a count of missing values would have revealed. Keep a log of every cleaning decision.

Plausibility checks

```
vars <- c("age", "education_years", "Annual_Income")
summary(socio[, vars])
sum(duplicated(socio$id))       # repeated ids?
```

```
##       age        education_years Annual_Income
##  Min.   :19.0   Min.   : 0.00   Min.   :10695
##  1st Qu.:28.5   1st Qu.: 4.25   1st Qu.:21114
##  Median :46.0   Median : 9.50   Median :25754
##  Mean   :41.1   Mean   :10.00   Mean   :25891
##  3rd Qu.:52.2   3rd Qu.:16.00   3rd Qu.:30307
##  Max.   :60.0   Max.   :20.00   Max.   :46375
##                 NA's   :2
## [1] 0
```

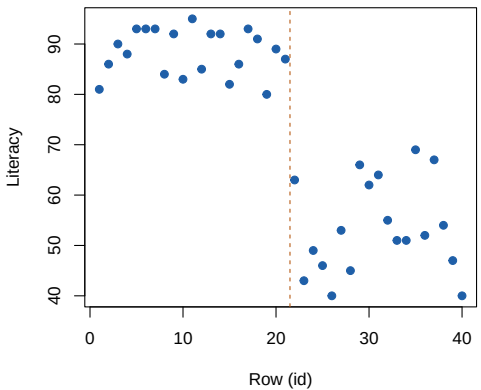
Reading it Ages 19–60 and schooling 0–20 years are plausible; 2 schooling values are missing; no id is repeated.

For every variable ask: is the range possible, are the units known, how much is missing?

A hidden problem: literacy depends on row order

```r
par(mar = c(4, 4, 1, 1))
plot(Literacy ~ id, data = socio, pch = 19,
     col = col_main, xlab = "Row (id)")
abline(v = 21.5, lty = 2, col = col_alt)
```

What to do Rows 1–21 score 80–95, rows 22–40 only 40–69: two batches were probably joined. Ask the provider; until then keep `Literacy` out of models.



Missing values: three choices

```r
colSums(is.na(titanic))[c("Age", "Cabin", "Embarked")]
nrow(na.omit(titanic))       # complete rows only
age_imp <- titanic$Age       # median imputation
age_imp[is.na(age_imp)] <- median(titanic$Age,
                                  na.rm = TRUE)
c(sd_observed = sd(titanic$Age, na.rm = TRUE),
  sd_imputed  = sd(age_imp))
```

```
##      Age    Cabin Embarked
##      177      687        2
## [1] 183
## sd_observed  sd_imputed
##      14.526      13.020
```

1. **Drop rows**: only 183 of 891 survive, as `Cabin` is mostly empty
2. **Impute**: easy, but distorts
3. **Keep** `NA`; each analysis drops what it must

> **Caution** Median imputation shrinks the sd of age from 14.5 to 13: later tests would look more precise than they are.

Duplicates that `duplicated()` misses

```r
sum(duplicated(sales))                  # exact copies
key <- sales[, c("month", "region")]
sum(duplicated(key))                    # same month twice
sales[sales$month == "01-05-2025", ]
sales <- sales[!duplicated(key), ]      # keep the first
nrow(sales)
```

```
## [1] 0
## [1] 1
##       month region sales marketing_spend Monthly_Revenue
## 5 01-05-2025  North   724             104             323
## 6 01-05-2025  North   724             104             114
## [1] 24
```

- No row is an exact copy, so `duplicated()` on the whole data says 0
- But May 2025 (North) appears twice, with revenue 323 and 114
- We keep the first, log the decision and ask the data owner

Text, factors and dates

```
unique(socio$Region)                # note the spaces
socio$Region <- trimws(socio$Region)
table(socio$Region)
socio$sector <- factor(socio$sector)
sales$date <- as.Date(sales$month, format = "%d-%m-%Y")
range(sales$date)

## [1] " West "  " East "  " North " " South "
##
##  East North South  West
##     7    12    10    11
## [1] "2025-01-01" "2026-12-01"
```

- `trimws()` strips outer spaces; `" West "` and `"West"` would otherwise be two groups
- `as.Date()` needs the format: day-month-year here

Check The months run to December 2026, after this workshop: the sales file is clearly illustrative.

Creating new variables with `cut()`

```
inc  <- socio$Annual_Income
cuts <- quantile(inc, c(0, 1/3, 2/3, 1))
socio$income_grp <- cut(inc, cuts, include.lowest = TRUE,
                        labels = c("Low", "Middle", "High"))
summary(socio$income_grp)
titanic$age_group <- cut(titanic$Age, c(0, 18, 60, Inf),
                         right = FALSE,
                       labels = c("Child", "Adult", "Senior"))
summary(titanic$age_group)      # NA = age unknown

##    Low Middle   High
##     14     13     13
##  Child  Adult Senior   NA's
##    113    575     26    177
```

- Terciles as breaks give three income groups of (nearly) equal size
- `right = FALSE` gives $[0, 18)$, $[18, 60)$, $[60, \infty)$: the same as nested `ifelse()` with `< 18` and `< 60`
- Other new variables: `log()`, `scale()` for z-scores, ratios

The cleaning log

Problem	Found by	Decision
Spaces in `Region`	`unique()`	`trimws()`
2 missing schooling values	`summary()`	keep `NA`; models drop 2 rows
Literacy in two batches	plot against row order	flag; ask provider; not modelled
May 2025 (North) twice	`duplicated()` on key	keep first; confirm with owner
Blank Titanic cabins and ports	counting `""`	read as `NA` (`na.strings`)
177 missing Titanic ages	`is.na()`	no imputation

Why keep a log Anyone, including you in six months, can see what was changed and why, and can rerun the script to get the same cleaned data.

PART 2

Descriptive statistics

Location and spread of income

```
round(c(mean = mean(inc), median = median(inc),
        sd = sd(inc), IQR = IQR(inc)))
round(100 * sd(inc) / mean(inc), 1)      # CV (%)
quantile(inc, c(0.10, 0.90))
skew <- mean((inc - mean(inc))^3) / sd(inc)^3
round(skew, 2)                           # skewness

##   mean median     sd    IQR
##  25891  25754   7923   9192
## [1] 30.6
##   10%   90%
## 16162 35042
## [1] 0.42
```

$$s^2 = \tfrac{1}{n-1}\sum (x_i - \bar{x})^2$$
$$\text{CV} = 100\, s/\bar{x}$$
$$g_1 = \tfrac{1}{n}\sum (x_i - \bar{x})^3 \big/ s^3$$

Reading it Mean and median almost coincide (25,891 vs 25,754) and skewness is only 0.42: unusually symmetric for incomes. CV = 31%.

Frequency tables and cross-tabulations

```
tab <- table(socio$sector, socio$gender)
addmargins(tab, 2)                  # add row totals
round(100 * prop.table(tab, margin = 1), 1)

##
##               Female Male Sum
##   Agriculture      4    8  12
##   Industry         5    4   9
##   Services        13    6  19
##
##               Female Male
##   Agriculture   33.3 66.7
##   Industry      55.6 44.4
##   Services      68.4 31.6
```

`addmargins()` adds totals; `prop.table()` with `margin = 1` gives row percentages (each sector adds to 100).

Reading it Women are 68% of services workers but 33% of those in agriculture: keep this in mind when we compare incomes by gender.

Grouped summaries

```
round(tapply(inc, socio$sector, mean))
aggregate(Annual_Income ~ sector + gender,
          data = socio, FUN = mean)

## Agriculture    Industry    Services
##       27673       23659       25823
##        sector gender Annual_Income
## 1 Agriculture Female         31673
## 2    Industry Female         27352
## 3    Services Female         28125
## 4 Agriculture   Male         25673
## 5    Industry   Male         19043
## 6    Services   Male         20836
```

`tapply()`: one summary per group. `aggregate()`: every combination, returned as a data frame.

Reading it In every sector women's mean income is higher than men's in this illustrative sample, so the gender gap is not just a sector effect. Cells hold only 4–13 people, so the means are noisy.

Survival on the Titanic: a socio-economic gradient

```
cls <- table(titanic$Pclass, titanic$Survived)
round(100 * prop.table(cls, margin = 1), 1)
surv <- tapply(titanic$Survived,
           list(titanic$Sex, titanic$Pclass), mean)
round(100 * surv, 1)        # % survived

##
##        0    1
##   1 37.0 63.0
##   2 52.7 47.3
##   3 75.8 24.2
##           1    2    3
## female 96.8 92.1 50.0
## male   36.9 15.7 13.5
```

Reading it 63% of first-class passengers survived, 24% in third class. Within every class women did far better: 97% of first-class women survived, 14% of third-class men.

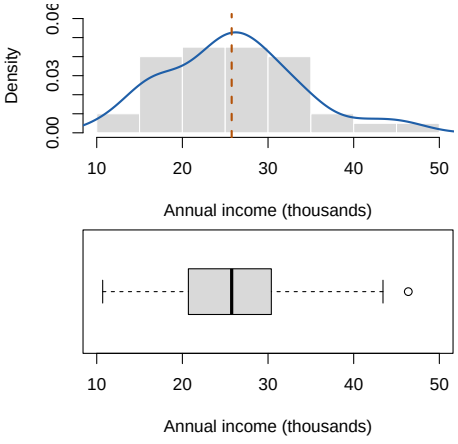
`Survived` is coded 0/1, so its mean is a proportion.

PART 3

Visualisation

One variable: histogram and box plot

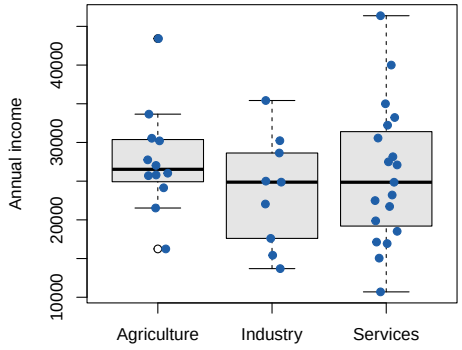
```r
inc_k <- inc / 1000                          # thousands
par(mfrow = c(2, 1), mar = c(4, 4, 0.5, 1))
hist(inc_k, breaks = 10, freq = FALSE,
     ylim = c(0, 0.06), col = "grey85",
     border = "white", main = "",
     xlab = "Annual income (thousands)")
lines(density(inc_k), col = col_main, lwd = 2)
abline(v = median(inc_k), col = col_alt, lwd = 2,
       lty = 2)
boxplot(inc_k, horizontal = TRUE, ylim = c(10, 50),
        col = "grey85",
        xlab = "Annual income (thousands)")
```



`lines()` and `abline()` add to a finished plot.

Comparing groups: box plots with the data

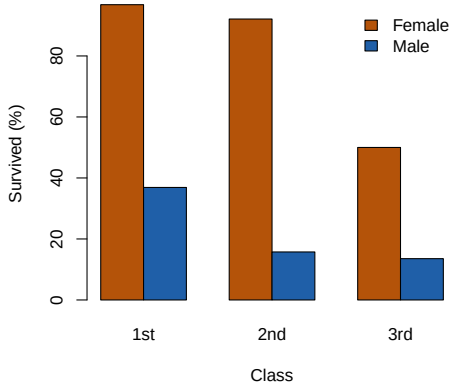
```r
par(mar = c(4, 4, 1, 1))
boxplot(Annual_Income ~ sector, data = socio,
        col = "grey90", xlab = "",
        ylab = "Annual income")
stripchart(Annual_Income ~ sector, data = socio,
           vertical = TRUE, method = "jitter",
           pch = 19, col = col_main, add = TRUE)
```



Reading it Box = middle 50%, line = median. The sectors overlap heavily; with only 9–19 people each, showing the points matters.

Categories: a grouped bar chart

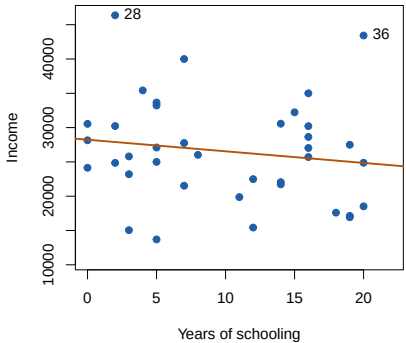
```r
par(mar = c(4, 4, 1, 1))
barplot(100 * surv, beside = TRUE,
        col = c(col_alt, col_main),
        names.arg = c("1st", "2nd", "3rd"),
        xlab = "Class", ylab = "Survived (%)",
        legend.text = c("Female", "Male"),
        args.legend = list(x = "topright",
                           bty = "n"))
```



Reading it Sex makes the bigger difference; class adds a gradient within each sex. Bars start at zero and show percentages, not counts.

Two numeric variables: scatter plot with additions

```r
par(mar = c(4, 4, 1, 1))
plot(Annual_Income ~ education_years, data = socio,
     pch = 19, col = col_main, xlim = c(0, 22),
     xlab = "Years of schooling", ylab = "Income")
line_fit <- lm(Annual_Income ~ education_years, socio)
abline(line_fit, col = col_alt, lwd = 2)
top <- which(inc > 40000)                # top earners
text(socio$education_years[top], inc[top],
     labels = socio$id[top], pos = 4)
```

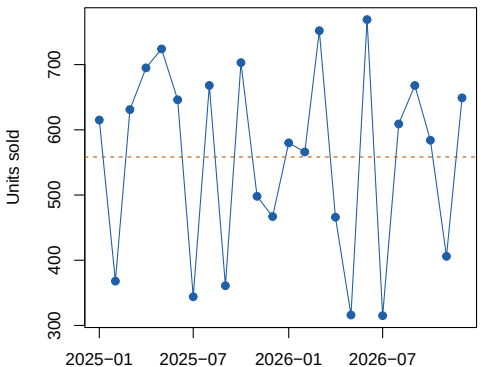


Reading it Flat line (slope -169 per year): no sign here that schooling raises income.

Ordered in time: a line plot

```r
par(mar = c(3, 4, 1, 1))
plot(sales ~ date, data = sales, type = "o",
     pch = 19, col = col_main,
     xlab = "", ylab = "Units sold")
abline(h = mean(sales$sales), col = col_alt,
       lty = 2)
```

Design flaw Each region is observed in a different half-year (North first, then South, East, West), so region and time period are completely confounded: no analysis can separate them.



PART 4

Statistical inference

The logic of a test, in one slide

1. State H_0 (no difference, no association) and H_1
2. Choose a test whose assumptions fit the data
3. Compute the test statistic from the sample
4. p-value: the probability, *if* H_0 *were true*, of a statistic at least as extreme as the one observed
5. Reject H_0 if $p < \alpha$ (usually 0.05); report the estimate and its confidence interval

Welch two-sample t statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Common misreadings p is *not* the probability that H_0 is true. "Not significant" does *not* prove "no effect". Significant does not mean large or important.

Do mean incomes differ by gender? Welch t-test

```
women <- inc[socio$gender == "Female"]
men   <- inc[socio$gender == "Male"]
(tt <- t.test(women, men))           # Welch is the default: variances may differ

##
##  Welch Two Sample t-test
##
## data:  women and men
## t = 2.6, df = 38, p-value = 0.013
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1324.4 10690.0
## sample estimates:
## mean of x mean of y
##     28594     22587
```

Reading the t-test like an economist

```
c(women = shapiro.test(women)$p.value,
  men   = shapiro.test(men)$p.value)

##   women     men
## 0.67202 0.85805
```

The numbers Means: women 28,594, men 22,587. Difference 6,007, 95% CI 1,324 to 10,690, $p = 0.013$. Shapiro–Wilk (above): normality within each group is plausible.

What it does not say

- It is an *association*, not a causal effect of gender
- Sector, age or schooling may explain part of the gap: regression adjusts
- 40 illustrative records say nothing about a real population

More than two groups: one-way ANOVA

```
fit_aov <- aov(Annual_Income ~ sector, data = socio)
summary(fit_aov)
bt <- bartlett.test(Annual_Income ~ sector, data = socio)
bt$p.value                  # equal variances?

##              Df     Sum Sq  Mean Sq F value Pr(>F)
## sector        2   83011213 41505607    0.65   0.53
## Residuals    37 2365264902 63926078
## [1] 0.52239
```

H_0: all three sector means are equal. F = between-group mean square / within-group mean square.

Reading it $F = 0.65$, $p = 0.53$: no evidence that mean income differs between sectors. Bartlett's $p = 0.52$: no reason to doubt equal variances.

Had ANOVA rejected, `TukeyHSD(fit_aov)` would show which pairs differ.

Two categorical variables: the χ^2 test

```r
ct <- chisq.test(cls)       # class x survived
ct
min(ct$expected)            # should exceed 5
cramer_v <- sqrt(ct$statistic / sum(cls))
round(unname(cramer_v), 2)

##
##  Pearson's Chi-squared test
##
## data:  cls
## X-squared = 103, df = 2, p-value <2e-16
##
## [1] 70.626
## [1] 0.34
```

$$\chi^2 = \sum (O - E)^2 / E$$
$$V = \sqrt{\chi^2 / (n\,(k-1))}$$
$$k = \min(\text{rows}, \text{columns})$$

Reading it $\chi^2 = 102.9$ on 2 df, $p < 0.001$: survival was not independent of class. Cramér's $V = 0.34$ ($k = 2$): a moderate association.

Two numeric variables: correlation test

```
ctest <- cor.test(sales$marketing_spend, sales$sales)
ctest

##
##  Pearson's product-moment correlation
##
## data:  sales$marketing_spend and sales$sales
## t = 1.29, df = 22, p-value = 0.21
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.15471  0.60401
## sample estimates:
##     cor
## 0.26524
```

Reading it $r = 0.27$ over 24 months, 95% CI -0.15 to 0.6, $p = 0.21$. The interval holds zero and large values alike: these data cannot tell whether marketing is linked to sales.

r measures *linear* association only, and is not evidence of cause.

PART 5

Regression for economists

The linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \qquad \text{OLS chooses } \hat{\beta} \text{ to minimise } \textstyle\sum_i (y_i - \hat{y}_i)^2$$

Assumptions on the errors: mean zero given the x's (correct form, no omitted confounder), independence, constant variance, and normality for small-sample tests.

Model	β_1 means
y on x	change in y per unit of x
$\ln y$ on x	about $100\,\beta_1$% per unit of x
$\ln y$ on $\ln x$	elasticity: % change per 1%

For a dummy in a log model the exact effect is $100\,(e^{\beta} - 1)\%$.

Sales and marketing: a simple regression

```
fit1 <- lm(sales ~ marketing_spend, data = sales)
round(coef(summary(fit1)), 3)
round(confint(fit1), 2)
round(summary(fit1)$r.squared, 3)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      461.047     80.852   5.702     0.00
## marketing_spend    0.696      0.540   1.290     0.21
##                  2.5 % 97.5 %
## (Intercept)     293.37 628.72
## marketing_spend  -0.42   1.82
## [1] 0.07
```

Reading it Each extra unit of marketing spend goes with 0.7 more units sold, but the 95% CI (-0.42 to 1.82) includes zero; marketing explains only 7% of the variation in sales.

The slope's t-test is the correlation test again: the same p-value.

The log–log model: an elasticity

```
fit2 <- lm(log(sales) ~ log(marketing_spend),
           data = sales)
round(coef(summary(fit2)), 3)
round(confint(fit2)[2, ], 2)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.588      0.672   8.311    0.000
## log(marketing_spend)   0.144      0.138   1.046    0.307
##  2.5 % 97.5 %
##  -0.14   0.43
```

$$\ln(\text{sales}) = \beta_0 + \beta_1 \ln(\text{marketing}) + \varepsilon$$

$\beta_1 = \partial \ln y / \partial \ln x$ (elasticity)

Reading it Elasticity 0.14: a 10% rise in marketing spend goes with about 1.4% more sales. The CI (-0.14 to 0.43) rules out neither no effect nor a sizeable one.

Elasticities are unit-free, which is why economists like them.

An earnings equation (Mincer type)

```
fit3 <- lm(log(Annual_Income) ~ education_years + age + gender + sector, data = socio)
round(coef(summary(fit3)), 3)

##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        10.505      0.196  53.695    0.000
## education_years    -0.008      0.007  -1.260    0.217
## age                -0.001      0.004  -0.168    0.867
## genderMale         -0.245      0.093  -2.623    0.013
## sectorIndustry     -0.261      0.130  -2.006    0.053
## sectorServices     -0.178      0.108  -1.649    0.109
```

$n = 38$ (the 2 rows with missing schooling are dropped), $R^2 = 0.257$, adjusted $R^2 = 0.141$. Reference groups: female, agriculture; `lm()` builds the dummies from the factors. Age stands in for experience.

Reading the earnings equation

```
b <- coef(fit3)[-1]
round(100 * (exp(b) - 1), 1)              # % effect of each variable

## education_years             age      genderMale  sectorIndustry  sectorServices
##            -0.8            -0.1           -21.7           -23.0           -16.3
```

Reading it

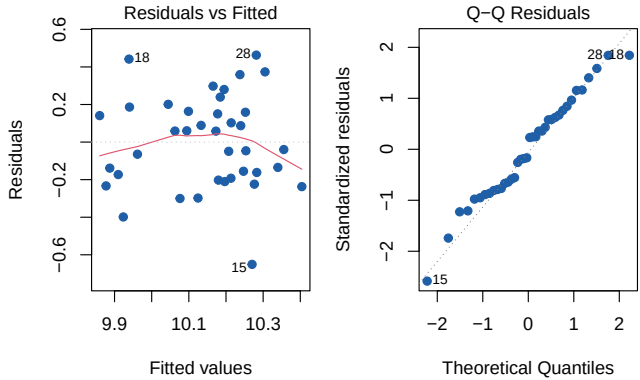
- Men: 22% less than women with the same schooling, age, sector ($p = 0.013$)
- Schooling: -0.8% per year, not significant
- Industry: -23% vs agriculture ($p = 0.053$)

Keep in mind Overall F-test: $p = 0.077$. Associations in 38 illustrative records; real data usually show clear positive returns to schooling.

Residual diagnostics

```r
par(mfrow = c(1, 2),
    mar = c(4, 4, 2, 1))
plot(fit3, which = c(1, 2),
     pch = 19, col = col_main)
# which = 3, 5: two more plots
```

Reading it Left: no curve, so a linear form fits. Right: points near the line, so normal errors are plausible.



Formal checks and what they tell us

```
bptest(fit3)                    # lmtest
shapiro.test(residuals(fit3))$p.value
max(cooks.distance(fit3))       # influence

##
##  studentized Breusch-Pagan test
##
## data:  fit3
## BP = 4.97, df = 5, p-value = 0.42
##
## [1] 0.64709
## [1] 0.18243
```

- **Breusch–Pagan**, H_0: constant variance; $p = 0.42$, no evidence of heteroscedasticity
- **Shapiro–Wilk** on residuals: $p = 0.65$
- **Cook's distance** at most 0.18: no observation comes near the common cut-off of 1

Verdict The assumptions look acceptable; the weak point is the small, illustrative sample.

Key take-aways

1. **Check before you analyse**: ranges, types, missing values, keys for duplicates, and plots against row order.
2. **Keep a cleaning log**; prefer leaving NA to silent imputation.
3. **Describe first**: centre, spread, shape, and summaries by group.
4. **Show the data**: histograms, box plots with points, bars that start at zero.
5. **Report estimates with confidence intervals**, not just p-values; association is not causation.
6. **Logs give percentages and elasticities**; read dummy effects as $100(e^{\beta} - 1)\%$.
7. **Diagnose every regression** with residual plots and tests.

References

- Dalgaard, P. (2008). *Introductory Statistics with R*, 2nd Ed., Springer, New York.
- Fox, J., and Weisberg, S. (2019). *An R Companion to Applied Regression*, 3rd Ed., Sage, Thousand Oaks.
- Jones, O., Harden, S., and Crawley, M. J. (2022). *The R Book*, 3rd Ed., Wiley, Hoboken.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.
- Breusch, T. S., and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5), 1287–1294.
- Shapiro, S. S., and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591–611.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. `https://www.R-project.org`

Thank you

Questions and discussion