Hands-on Data Analytics with R

Dr. Kapil Kumar
Associate Professor, Department of Statistics
Central University of Haryana, Mahendergarh, India
kapilstats@gmail.com

Hands-on Training on Data Analytics through Statistical Tools

Bhaskara Association, Department of Mathematics, Faculty of Science, SGT University, Gurugram

6 April 2023 · **Resource person**

Plan of the session

1. Workflow and set-up
2. Describing data quickly
3. Categorical data: tables and association
4. Comparing means: t-tests and ANOVA
5. Regression with diagnostics

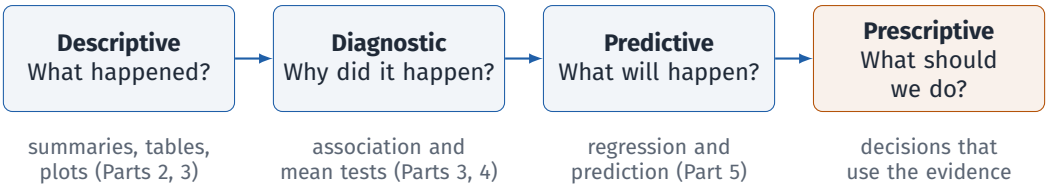
By the end you can read and check a data file, summarise it, test an association in a two-way table, compare two or more means, fit and diagnose a regression line, and state every result in plain words.

Hands-on Type the code as we go. Every box runs top to bottom in a fresh R session.

PART 1

Workflow and set-up

From data to decisions



Today's rule Every analysis ends with a plain-language sentence: what was tested, the size of the effect, how uncertain it is, and what it does *not* show.

Organise the work as a project

hands_on/ (an RStudio project)

- `analysis.R` script, run top to bottom
- **data/**
  - `marks.csv`
  - `sugar_patients_data.csv`
  - `smoker_data.csv`
  - `data15.txt`

- Install R (`cran.r-project.org`), then RStudio Desktop (`posit.co`).
- *File* → *New Project*: RStudio then works inside that folder, so `"data/marks.csv"` is found on any computer.
- Write code in the script and send lines to the console with `Ctrl+Enter`.
- Help: `?t.test`; search: `??regression`.

Packages and colours used today

- Base R does most of the work: `table()`, `t.test()`, `aov()`, `lm()`.
- **MASS**: `studres()`, studentized residuals
- **lmtest**: `dwtest()`, `bptest()`, `coeftest()`
- **car**: `hccm()`, robust standard errors
- **nortest**: `lillie.test()` for normality

```r
# install once:
# install.packages(c("lmtest", "car",
#                    "nortest"))
library(MASS)
library(lmtest)
library(car)
library(nortest)

col_main <- "#1F5FA8"   # plot colours
col_alt  <- "#B45309"
```

Reading today's four data files

```
marks <- read.csv("data/marks.csv")                          # two tests, 20 students
sugar <- read.csv("data/sugar_patients_data.csv", stringsAsFactors = TRUE)
smoke <- read.csv("data/smoker_data.csv")                    # before/after, 8 smokers
batch <- read.table("data/data15.txt", header = TRUE)        # space-separated, row names
sapply(list(marks = marks, sugar = sugar, smoke = smoke, batch = batch), dim)
```

```
##      marks sugar smoke batch
## [1,]    20    20     8     5
## [2,]     3     3     2     3
```

Data note Rows are observations, columns are variables. All four files are small illustrative teaching data; the sugar-patient data are hypothetical.

PART 2

Describing data quickly

A first look at the marks data

```
head(marks, 3)
anyNA(marks)       # any missing values?
sum(duplicated(marks))

##   Roll_No. m1 m2
## 1        1 95  2
## 2        2 99  1
## 3        3 67 56
## [1] FALSE
## [1] 0
```

```
summary(marks[, c("m1", "m2")])

##        m1              m2
##  Min.   : 10.0   Min.   : 1.0
##  1st Qu.: 32.8   1st Qu.:20.0
##  Median : 53.5   Median :46.0
##  Mean   : 56.5   Mean   :44.5
##  3rd Qu.: 77.5   3rd Qu.:65.2
##  Max.   :100.0   Max.   :95.0
```

Checks before analysis No missing values, no duplicated rows; marks lie on a 0–100 scale. But the minimum of `m2` is 1: worth a second look.

Summaries with `sapply()` and our own function

```r
m <- marks[, c("m1", "m2")]
cv <- function(x) 100 * sd(x) / mean(x)
round(sapply(m, mean), 2)
round(sapply(m, sd), 2)
round(sapply(m, cv), 1)        # CV in %
```

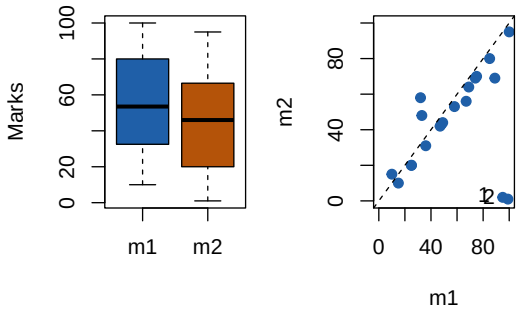
```
##    m1    m2
## 56.55 44.50
##    m1    m2
## 28.68 26.75
##   m1   m2
## 50.7 60.1
```

$$\mathrm{CV} = 100 \times s/\bar{x} \quad \text{(in \%)}$$

Reading it Test 2 has the lower mean (44.5 against 56.5) and the larger relative spread (CV 60% against 51%).

Look at the two tests together

```
par(mfrow = c(1, 2), mar = c(4, 4, 1, 1))
boxplot(m, col = c(col_main, col_alt),
        ylab = "Marks")
plot(m2 ~ m1, data = marks, pch = 19,
     col = col_main, xlim = c(0, 100),
     ylim = c(0, 100))
abline(0, 1, lty = 2)          # m2 = m1
text(marks$m1[1:2], marks$m2[1:2],
     labels = 1:2, pos = 2)  # rows 1, 2
```



Most students lie just below the line $m_2 = m_1$; rows 1 and 2 scored above 90 in test 1 and almost nothing in test 2.

Diagnostic analytics: why is the correlation weak?

```
cor(marks$m1, marks$m2)
marks$diff <- marks$m1 - marks$m2
table(marks$diff)
marks[marks$diff > 50, ]
cor(marks$m1[-(1:2)], marks$m2[-(1:2)])
```

```
## [1] 0.415
##
## -26 -15  -5   5  11  20  93  98
##   1   1   1  13   1   1   1   1
##   Roll_No. m1 m2 diff
## 1        1 95  2   93
## 2        2 99  1   98
## [1] 0.9303
```

Reading it $r = 0.42$ for all 20 students but $r = 0.93$ without rows 1–2. 13 students have $m_2 = m_1 - 5$ exactly.

Do not just delete Marks of 2 and 1 after 95 and 99 suggest absence or an entry error. Check the records; report results with and without them.

PART 3

Categorical data

The sugar-patient data

```
head(sugar, 4)
summary(sugar)

##   diabetes   age gender
## 1       no   old female
## 2       no young female
## 3      yes   old   male
## 4      yes   old   male
##  diabetes    age        gender
##  no :13   old  : 8   female:11
##  yes: 7   young:12   male  : 9
```

- 20 hypothetical patients and three categorical variables: diabetes (no/yes), age group (old/young) and gender.
- `stringsAsFactors = TRUE` stored them as factors, so `summary()` gives counts.

> **Question** Is diabetes associated with age group? With gender?

Two-way tables and row proportions

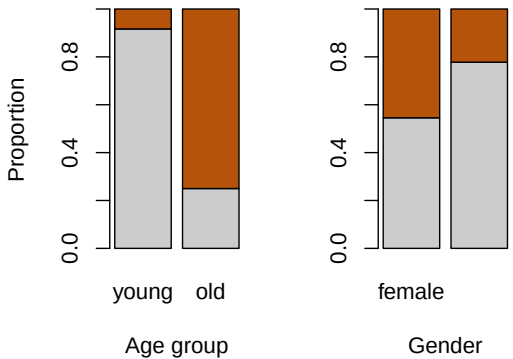
```
sugar$age <- factor(sugar$age, levels = c("young", "old"))
tab <- table(age = sugar$age, diabetes = sugar$diabetes)
addmargins(tab)
round(prop.table(tab, margin = 1), 2)

##        diabetes
## age     no yes Sum
##   young 11   1  12
##   old    2   6   8
##   Sum   13   7  20
##        diabetes
## age       no  yes
##   young 0.92 0.08
##   old   0.25 0.75
```

- Explanatory variable (age) in the rows, *row* proportions: the share with diabetes in each age group.

Reading it 6 of 8 older patients (75%) have diabetes, against 1 of 12 younger ones.

A picture of the two associations

```
tab_g <- table(gender = sugar$gender,
               diabetes = sugar$diabetes)
par(mfrow = c(1, 2), mar = c(4, 4, 1, 1))
barplot(t(prop.table(tab, 1)),
        col = c("grey80", col_alt),
      ylab = "Proportion", xlab = "Age group")
barplot(t(prop.table(tab_g, 1)),
        col = c("grey80", col_alt),
        xlab = "Gender")
```



Stacked row proportions; amber = diabetes. The share differs sharply by age group and only mildly by gender.

Chi-square test: check the expected counts first

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad E_{ij} = \frac{R_i\, C_j}{n}$$

H_0: no association; then $X^2 \approx \chi^2_{(r-1)(c-1)}$

Rule of thumb The approximation needs expected counts of about 5 or more.

```
chisq.test(tab)$expected
```

```
##         diabetes
## age        no yes
##   young 7.8 4.2
##   old   5.2 2.8
```

Here Two expected counts (2.8 and 4.2) are below 5, so we use Fisher's exact test.

Fisher's exact test

```
ft <- fisher.test(tab)        # age group x diabetes: exact, no approximation
ft

##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 0.004
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##     1.821 1625.380
## sample estimates:
## odds ratio
##      25.14
```

Reading the categorical analysis

Age group $p = 0.0044$: strong evidence that diabetes and age group are associated. The odds of diabetes for older patients are estimated at about 25 times those for younger ones.

Gender `fisher.test(tab_g)` gives $p = 0.37$: no evidence of an association in these data.

How sure are we? The 95% interval for the odds ratio, 1.8 to 1625, is very wide: with 20 patients the direction is clear, the size is not.

What it does not show Association is not causation, and “no evidence” for gender is not evidence of no effect: a small sample has little power.

PART 4

Comparing means

Two batches: data and hypotheses

```
batch
sapply(batch[, c("batchA", "batchB")], mean)
sapply(batch[, c("batchA", "batchB")], sd)

##   batchA batchB batchC
## 1   20.0     40     10
## 2   20.5     42     15
## 3   13.0     56     12
## 4   16.0     38     10
## 5   11.0     57      3
## batchA batchB
##   16.1   46.6
## batchA batchB
##  4.189  9.154
```

$H_0 : \mu_A = \mu_B$ vs $H_1 : \mu_A \neq \mu_B$

Welch: $t = \dfrac{\bar{x}_A - \bar{x}_B}{\sqrt{s_A^2/n_A + s_B^2/n_B}}$

Scores of five students in each of three batches (illustrative). Different students in each batch, so the samples are *independent*.

For data in long format use the formula version: `t.test(values ~ ind)`.

The two-sample t-test

```
fit_t <- t.test(batch$batchA, batch$batchB)   # Welch: variances not assumed equal
fit_t

##
##  Welch Two Sample t-test
##
## data:  batch$batchA and batch$batchB
## t = -6.8, df = 5.6, p-value = 0.0007
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -41.71 -19.29
## sample estimates:
## mean of x mean of y
##      16.1      46.6
```

Reading the two-sample test

```
names(fit_t)
round(fit_t$conf.int, 2)

##  [1] "statistic"   "parameter"   "p.value"
##  [4] "conf.int"    "estimate"    "null.value"
##  [7] "stderr"      "alternative" "method"
## [10] "data.name"
## [1] -41.71 -19.29
## attr(,"conf.level")
## [1] 0.95
```

A test result is a list: pull out what you need with `$`.

In plain words Batch B scored on average 30.5 marks more than batch A (95% CI 19.3 to 41.7). With $t = -6.77$ on 5.6 df and $p = 0.0007$, a difference this large is very unlikely by chance.

Welch or pooled? Welch is a safe default. The pooled test (`var.equal = TRUE`) gives the same t here (equal n) on 8 df, $p = 0.0001$.

A paired design: does the film reduce smoking?

```
smoke$d <- smoke$Before - smoke$After
smoke

##   Before After  d
## 1     28    26  2
## 2     29    27  2
## 3     31    32 -1
## 4     44    44  0
## 5     35    35  0
## 6     20    16  4
## 7     50    47  3
## 8     25    23  2
```

- Mean daily cigarettes for 8 heavy smokers, in the month before and the month after an anti-smoking film.
- The same people are measured twice, so we analyse the differences d_i: a one-sample t-test on d.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \text{ on } n-1 \text{ df}$$

$H_0 : \mu_d = 0$ vs $H_1 : \mu_d > 0$ (a decline)

The paired t-test

```
pt <- t.test(smoke$Before, smoke$After, paired = TRUE, alternative = "greater")
pt

##
##  Paired t-test
##
## data:  smoke$Before and smoke$After
## t = 2.5, df = 7, p-value = 0.02
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  0.3678    Inf
## sample estimates:
## mean difference
##             1.5
```

Reading the paired test

In plain words Consumption fell by 1.5 cigarettes a day on average. One-sided $p = 0.02$: at the 5% level the data support a decline. The 95% lower bound is 0.37 cigarettes.

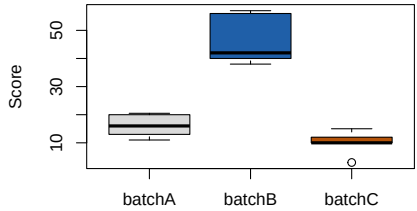
Why one-sided? The question asked about a *decline*, fixed before seeing the data. The two-sided p would be 0.04.

Limits of this design No control group, so the drop may have other causes. With $n = 8$, check the differences: Shapiro–Wilk gives $p = 0.56$.

Wrong test, wrong answer An unpaired test ignores that each smoker is their own control; here it gives $p = 0.39$.

Three batches: one-way ANOVA

```r
scores <- stack(batch)   # long format: values, ind
fit_aov <- aov(values ~ ind, data = scores)
summary(fit_aov)

##             Df Sum Sq Mean Sq F value    Pr(>F)
## ind          2   3845    1923    47.7 0.0000019 ***
## Residuals   12    483      40
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$
$$H_0 : \tau_A = \tau_B = \tau_C = 0$$
$$F = \mathrm{MS}_{\text{between}} / \mathrm{MS}_{\text{within}}$$

Which batches differ? Tukey's HSD

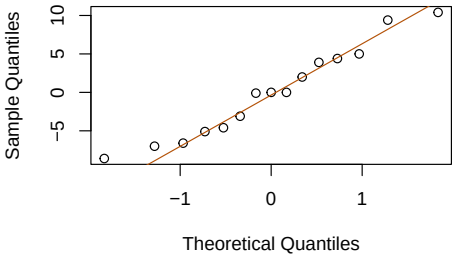
```
tk <- TukeyHSD(fit_aov)
round(tk$ind, 4)
tapply(scores$values, scores$ind, mean)

##                 diff    lwr      upr  p adj
## batchB-batchA  30.5  19.79  41.209 0.0000
## batchC-batchA  -6.1 -16.81   4.609 0.3167
## batchC-batchB -36.6 -47.31 -25.891 0.0000
## batchA batchB batchC
##   16.1   46.6   10.0
```

Reading it The ANOVA $F = 47.7$ ($p < 0.001$) rejects equal means. Tukey's family-wise 95% intervals put batch B above both A and C; A and C do not differ significantly ($p = 0.32$).

Tukey's method controls the chance of *any* false difference across all three comparisons.

Checking the ANOVA assumptions

```r
e <- residuals(fit_aov)
shapiro.test(e)$p.value
bartlett.test(values ~ ind, data = scores)$p.value
par(mar = c(4, 4, 1, 1))
qqnorm(e, main = ""); qqline(e, col = col_alt)

## [1] 0.6125
## [1] 0.2269
```



Independence holds by design. Normality (Shapiro–Wilk) and equal variances (Bartlett): no evidence against.

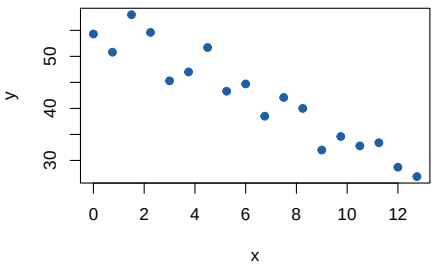
Small samples Five per group: low power, so a large p reassures but does not prove.

PART 5

Regression with diagnostics

The data and a first look

```r
x <- seq(0, 12.75, by = 0.75)          # 18 values
y <- c(54.3, 50.8, 58, 54.6, 45.3, 47, 51.7, 43.3,
       44.7, 38.5, 42.1, 40, 32, 34.6, 32.8, 33.4,
       28.7, 26.9)
par(mar = c(4, 4, 1, 1))
plot(x, y, pch = 19, col = col_main)
cor(x, y)
cor(x, y, method = "spearman")

## [1] -0.9487
## [1] -0.9422
```



18 paired observations (illustrative data from the lecture). Pearson and Spearman correlations agree: a strong, negative trend.

Is the correlation significant?

```
(ct <- cor.test(x, y))

##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = -12, df = 16, p-value = 2e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9811 -0.8650
## sample estimates:
##     cor
## -0.9487
```

Reading it $r = -0.949$ (95% CI -0.98 to -0.87), $p < 0.001$: a strong negative *linear* association.

Remember
Correlation says nothing about cause.

Fitting the regression line

```
fit <- lm(y ~ x)
coef(summary(fit))
round(confint(fit), 3)
c(R2 = summary(fit)$r.squared,
  sigma = summary(fit)$sigma)
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   56.418     1.3921   40.53 1.481e-17
## x             -2.238     0.1864  -12.01 2.040e-09
##              2.5 % 97.5 %
## (Intercept) 53.466 59.369
## x           -2.633 -1.843
##     R2  sigma
## 0.9001 3.0771
```

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$
ε_i independent $N(0, \sigma^2)$

- Least squares minimises $\sum (y_i - \hat{y}_i)^2$.
- `summary(fit)` prints everything; here we pick the parts we need.

Reading the fitted model

```
anova(fit)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## x          1   1365    1365     144  2e-09 ***
## Residuals 16    151       9
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In plain words $\hat{y} = 56.42 - 2.24\,x$. Each unit increase in x lowers the mean of y by 2.24 (95% CI 1.84 to 2.63); x explains 90% of the variation in y.

The ANOVA F equals t^2 for the slope. All of this rests on the assumptions, checked next.

What to check, and with which tool

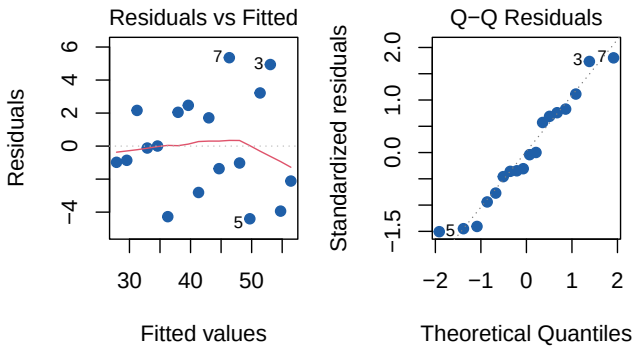
Assumption	Plot	Test
Linearity	residuals vs fitted	—
Independence	residuals vs order	Durbin–Watson, `dwtest()`
Constant variance	residuals vs fitted	Breusch–Pagan, `bptest()`
Normality of errors	normal Q–Q plot	Shapiro–Wilk, Lilliefors
No unusual points	studentized residuals	$\lvert r_i^* \rvert > 2$ deserves a look

Plots first Tests answer “is there evidence of a problem?”; plots show what the problem is and how serious it is. With 18 points, use both.

Residual plots

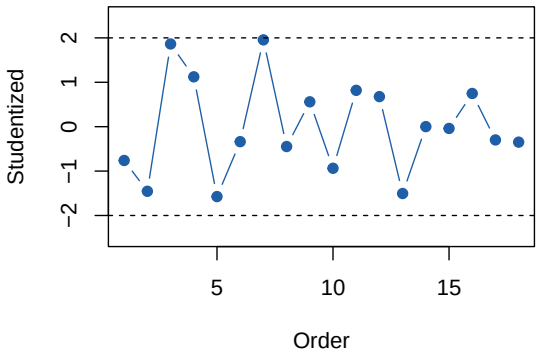
```
par(mfrow = c(1, 2),
    mar = c(4, 4, 2, 1))
plot(fit, which = 1:2,
     pch = 19, col = col_main)
```

Reading it No curvature. The spread is wider for small x (large fitted values): residual SD 3.7 for $x < 6.5$ against 2.3 above. The Q–Q plot is close to a line.



Studentized residuals: any unusual points?

```
r_stud <- studres(fit)          # MASS
round(range(r_stud), 2)
par(mar = c(4, 4, 1, 1))
plot(r_stud, type = "b", pch = 19,
     col = col_main, ylim = c(-2.5, 2.5),
     xlab = "Order", ylab = "Studentized")
abline(h = c(-2, 2), lty = 2)

## [1] -1.57  1.96
```



Reading it All $|r_i^*| < 2$: no outlier. The signs alternate, with no long runs.

Independence: the Durbin–Watson test

```
dw <- dwtest(fit)              # H1: positive autocorrelation of the errors
dw$method
c(DW = unname(dw$statistic), p = dw$p.value)

## [1] "Durbin-Watson test"
##     DW      p
## 2.4492 0.7626
```

$DW = \sum_{t=2}^{n}(e_t - e_{t-1})^2 / \sum_{t=1}^{n} e_t^2 \approx 2(1 - r_1)$, where r_1 is the lag-1 autocorrelation of the residuals ($DW \approx 2$: none).

Reading it $DW = 2.45$, $p = 0.76$: no evidence of positive autocorrelation.

Constant variance: the Breusch–Pagan test

```
bptest(fit)

##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 4.4, df = 1, p-value = 0.04
```

Regress e_i^2 on x_i; under H_0 (constant variance) $BP = nR^2_{\text{aux}} \approx \chi^2_1$.

Reading it $p = 0.035 < 0.05$: some evidence that the error variance changes with x, as the residual plot suggested.

Remedy Keep the line but use heteroscedasticity-consistent standard errors (next slide), or model the variance.

Robust standard errors, and normality of residuals

```
rob <- coeftest(fit, vcov = hccm(fit))
rob[, 1:4]
shapiro.test(residuals(fit))$p.value
lillie.test(residuals(fit))$p.value

##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   56.418     1.7948   31.43 8.223e-16
## x             -2.238     0.1956  -11.44 4.087e-09
## [1] 0.5399
## [1] 0.7814
```

Reading it The robust SE of the slope is 0.196 (ordinary: 0.186); the slope stays highly significant. Neither normality test finds a problem.

A common mistake `ks.test()` with mean and SD estimated from the residuals: p too large.

Prediction with intervals

```
new <- data.frame(x = c(6, 14))
pc <- predict(fit, new, interval = "confidence")
pp <- predict(fit, new, interval = "prediction")
round(pc, 2)
round(pp, 2)
range(x)

##     fit   lwr   upr
## 1 42.99 41.44 44.53
## 2 25.08 21.70 28.47
##     fit   lwr   upr
## 1 42.99 36.29 49.69
## 2 25.08 17.74 32.43
## [1]  0.00 12.75
```

At $x = 6$ Mean of y: 43.0 (95% CI 41.4 to 44.5). A single new y: 36.3 to 49.7, wider because it adds the scatter σ.

At $x = 14$ Outside the observed range of x: an extrapolation the data cannot support.

Key take-aways

1. **Check the data first:** missing values, duplicates, odd values. Two records moved r from 0.42 to 0.93.
2. **Small tables:** look at expected counts; use Fisher's exact test when they are small.
3. **Match the test to the design:** independent groups (Welch), the same subjects twice (paired), three or more groups (ANOVA, then Tukey).
4. **Report effects with intervals**, not just p-values.
5. **Regression does not end at `summary()`:** plot residuals, test assumptions, use robust SEs when the variance is not constant, do not extrapolate.

References

- Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5), 1287–1294.
- Dalgaard, P. (2008). *Introductory Statistics with R*, 2nd ed. Springer.
- Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*, 3rd ed. Sage.
- Montgomery, D. C., Peck, E. A. and Vining, G. G. (2021). *Introduction to Linear Regression Analysis*, 6th ed. Wiley.
- Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591–611.
- Verzani, J. (2014). *Using R for Introductory Statistics*, 2nd ed. CRC Press.
- 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