Practical Utility of Correlation and Regression through R

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

One Week Online National Workshop on Statistical Data Analysis using SPSS & R (SDASR), 5th Edition (11–15 March 2024)
Department of Statistics, Central University of Haryana, Mahendergarh
12 March 2024 · Lecture 2 of 2 (online)

Plan of the session

1. Relationships between two variables
2. Measuring correlation
3. Simple linear regression
4. Regression diagnostics
5. Prediction, and practice

By the end you can draw and read a scatterplot, compute and test Pearson and Spearman correlations, spot unpaired data and outliers, fit and interpret a straight-line regression, check its assumptions with plots and tests, and predict without extrapolating.

Where we left off

Session 1: testing of hypothesis

- H_0 and H_1 come from the question; α is fixed in advance
- p-value: how surprising the data are if H_0 is true; reject when $p \leq \alpha$
- One-sample, two-sample (pooled, Welch) and paired t tests on the blood pressure, batch, smoker and IQ data
- Report an estimate and a confidence interval, not only p

Today: same logic, new parameters

$H_0 : \mu = \mu_0 \quad \rightarrow \quad H_0 : \rho = 0$

$H_0 : \mu_1 = \mu_2 \quad \rightarrow \quad H_0 : \beta_1 = 0$

No correlation, no slope: tested with a t statistic, a p-value and a confidence interval, exactly as before.

We reuse session 1's data where it fits: the smokers, the two batches and the blood pressure groups.

Packages used in this session

- **ggpubr**: `ggscatter()`, scatterplot with fitted line and r
- **Hmisc**: `rcorr()`, correlation matrix with p-values
- **arm**: `display()`, compact regression summary
- **MASS**: `studres()`, studentized residuals
- **lmtest**: `dwtest()`, `bptest()`
- **nortest**: `lillie.test()`, normality test with estimated parameters

```r
# install once
# install.packages(c("ggpubr", "Hmisc", "arm",
#                    "MASS", "lmtest", "nortest"))
library(ggpubr);  library(Hmisc)
library(arm);     library(MASS)
library(lmtest);  library(nortest)

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

PART 1

Relationships between two variables

Two questions about two variables

Correlation: do X and Y move together, and how strongly?

- symmetric: $r_{XY} = r_{YX}$
- a single, unit-free number between -1 and 1

Regression: how does Y change with X, and what Y do we expect for a given X?

- asymmetric: response Y, predictor X
- slope in the units of Y per unit of X

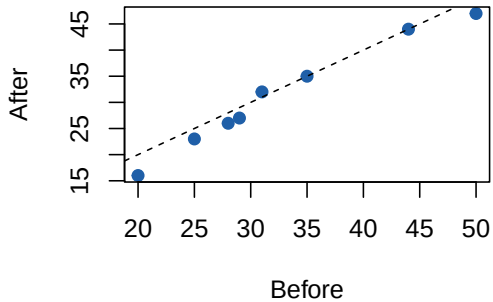
Why it matters Linear regression is the basis of much of applied statistics: analysis of variance and covariance are special cases of it, and generalised linear models extend it.

First step, always Plot Y against X. A number cannot tell you whether the pattern is straight, curved or driven by one odd point.

Back to the smokers: why pairing worked

```r
smoke <- read.csv("data/smoker_data.csv")
cor(smoke$Before, smoke$After)

## [1] 0.9878

par(mar = c(4, 4, 1, 1))
plot(After ~ Before, data = smoke,
     pch = 19, col = col_main)
abline(0, 1, lty = 2)   # line of no change
```

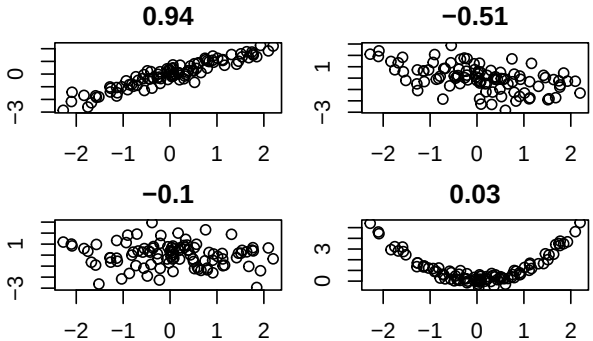
Link to session 1 $r = 0.988$: each smoker's two values are close, so the paired test is sensitive.



Cigarettes per day, one point per smoker: most lie just below the line of no change.

What different values of r look like

```r
set.seed(21)
x  <- rnorm(100)
y1 <- x + rnorm(100, sd = 0.4)   # strong
y2 <- -0.5 * x + rnorm(100)      # moderate
y3 <- rnorm(100)                 # unrelated
y4 <- x^2 + rnorm(100, sd = 0.3) # curved
par(mfrow = c(2, 2), cex = 0.85,
    mar = c(2, 2, 2, 1))
plot(x, y1, main = round(cor(x, y1), 2))
plot(x, y2, main = round(cor(x, y2), 2))
plot(x, y3, main = round(cor(x, y3), 2))
plot(x, y4, main = round(cor(x, y4), 2))
```



Bottom right: a strong *curved* relationship, yet $r \approx 0$; r measures *linear* association only.

PART 2

Measuring correlation

Pearson's correlation coefficient

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\,\sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{\operatorname{cov}(x, y)}{s_x\, s_y}$$

Test of $H_0 : \rho = 0$ (bivariate normal data):

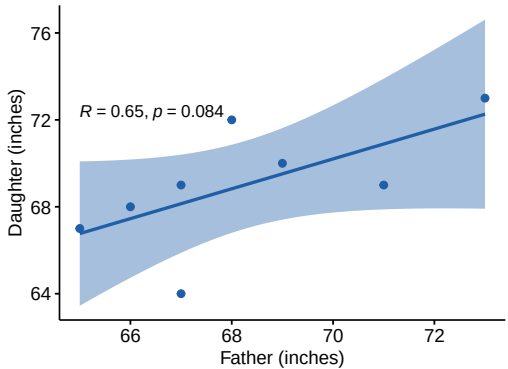
$$t = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}} \sim t_{n-2}$$

Properties $-1 \le r \le 1$; the sign gives the direction. r has no units and does not change if either variable is shifted or rescaled. $r = \pm 1$ only when all points lie on a line.

Not a slope r says how closely points follow a line, not how steep it is.

Heights of fathers and daughters

```
height_F <- c(65, 66, 67, 67, 68, 69, 71, 73)
height_D <- c(67, 68, 64, 69, 72, 70, 69, 73)
hts <- data.frame(height_F, height_D)
ggscatter(hts, x = "height_F", y = "height_D",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, color = col_main,
          xlab = "Father (inches)",
          ylab = "Daughter (inches)")
```



Eight father–daughter pairs (illustrative). The source labelled the heights in cm; values of 64–73 can only be inches.

Testing the correlation: `cor.test()`

```
cor.test(hts$height_F, hts$height_D)

##
##  Pearson's product-moment correlation
##
## data:  hts$height_F and hts$height_D
## t = 2.0717, df = 6, p-value = 0.08369
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1080788  0.9281049
## sample estimates:
##       cor
## 0.6457766
```

Reading it $r = 0.65$, a fairly strong positive association, but $p = 0.084$ and the 95% CI (-0.11, 0.93) includes 0.

Sample size With only 8 pairs even $r = 0.65$ is not significant at 5%; the interval is very wide.

Spearman's rank correlation

Replace each variable by its ranks and compute Pearson's r on the ranks. Without ties:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

d_i: difference between the two ranks of unit i

Use it for ordinal data, for monotone but non-linear relationships, or when a few extreme values worry you.

```
cor(height_F, height_D, method = "spearman")

## [1] 0.7651

cor.test(height_F, height_D,
         method = "spearman",
         exact = FALSE)$p.value

## [1] 0.02698
```

Ties (two 67s, two 69s) rule out an exact p-value: `exact = FALSE`. Pearson and Spearman disagree at 5% here; choose the measure in advance.

Pitfall 1: correlation needs paired observations

```r
d3 <- read.table("data/data3.txt", header = TRUE)
d3$batchA
d3$batchB
cor(d3$batchA, d3$batchB)
# session 1 listed batch B in another order
cor(d3$batchA, c(40, 42, 38, 56, 57))
```

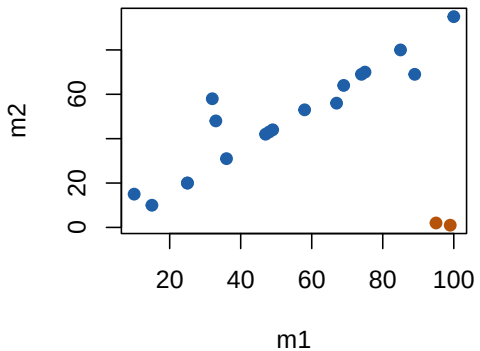
```
## [1] 20.0 20.5 13.0 16.0 11.0
## [1] 40 42 56 38 57
## [1] -0.8299
## [1] -0.4778
```

- Batch A and batch B are scores of *different* students (we compared them with a two-sample t test in session 1)
- Row i of A has nothing to do with row i of B, so any pairing is arbitrary
- Reordering B changes r from -0.83 to -0.48

> **Rule** Correlate two measurements made on the *same* units.

Pitfall 2: one or two odd points

```r
marks <- read.csv("data/marks.csv")
odd <- marks$Roll_No. %in% c(1, 2)
cor(marks$m1, marks$m2)                # all 20
cor(marks$m1[!odd], marks$m2[!odd])  # 18
par(mar = c(4, 4, 1, 1))
plot(m2 ~ m1, data = marks, pch = 19,
     col = ifelse(odd, col_alt, col_main))

## [1] 0.415
## [1] 0.9303
```



Reading it Students 1 and 2 (orange) pull r from 0.93 down to 0.42.

What to do with suspicious points

1. **Check the source**: marks of 2 and 1 after 95 and 99 look like an absence or a data-entry error
2. **Correct** a proven error; do not delete a point only because it is inconvenient
3. **Report both analyses** if the point is genuine but unusual
4. **Rank-based measures help only partly**: here Spearman's $r_s = 0.45$ is pulled down too, so no single number rescues bad data

Always plot A correlation of 0.42 would suggest a weak relationship between the two papers. The plot shows a near-perfect relationship for 18 students and two clear anomalies.

Data note `marks.csv` holds illustrative marks of 20 students identified only by roll number.

Many variables at once: the correlation matrix

```
cars6 <- mtcars[, c("mpg", "disp", "hp",
                    "drat", "wt", "qsec")]
round(cor(cars6), 2)

##        mpg  disp    hp  drat    wt  qsec
## mpg   1.00 -0.85 -0.78  0.68 -0.87  0.42
## disp -0.85  1.00  0.79 -0.71  0.89 -0.43
## hp   -0.78  0.79  1.00 -0.45  0.66 -0.71
## drat  0.68 -0.71 -0.45  1.00 -0.71  0.09
## wt   -0.87  0.89  0.66 -0.71  1.00 -0.17
## qsec  0.42 -0.43 -0.71  0.09 -0.17  1.00
```

- `mtcars`: 32 cars from a 1974 US magazine road test (built into R)
- mpg: miles per gallon; disp: engine size; hp: horsepower; wt: weight (1000 lb); qsec: quarter-mile time
- Fuel economy falls with weight ($r = -0.87$) and engine size ($r = -0.85$)

p-values for every pair: `Hmisc::rcorr()`

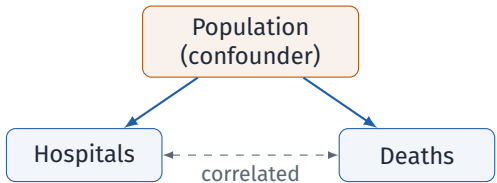
```
res <- rcorr(as.matrix(cars6))
round(res$P, 3)       # p-values; diagonal is NA

##        mpg  disp   hp drat    wt  qsec
## mpg     NA 0.000 0.00 0.00 0.000 0.017
## disp 0.000    NA 0.00 0.00 0.000 0.013
## hp   0.000 0.000   NA 0.01 0.000 0.000
## drat 0.000 0.000 0.01   NA 0.000 0.620
## wt   0.000 0.000 0.00 0.00    NA 0.339
## qsec 0.017 0.013 0.00 0.62 0.339    NA
```

- `rcorr()` returns `$r` (correlations), `$n` and `$P`
- Only drat–qsec and wt–qsec are not significant at 5%

15 tests at once With many pairs some will be “significant” by chance. Adjust (e.g. `p.adjust(p, "holm")`) or treat the matrix as exploratory.

Correlation is not causation

- Across districts, numbers of hospitals and of deaths are positively correlated. Hospitals do not cause deaths: both grow with *population*.
- A third variable that drives both is a confounder.
- Only a designed experiment (randomisation) or careful adjustment for confounders supports a causal claim.



Say it carefully "X and Y are associated", not "X affects Y", unless the design allows it.

PART 3

Simple linear regression

The simple linear regression model

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

Least squares minimises $\sum_i (y_i - b_0 - b_1 x_i)^2$, giving

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r\,\frac{s_y}{s_x}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

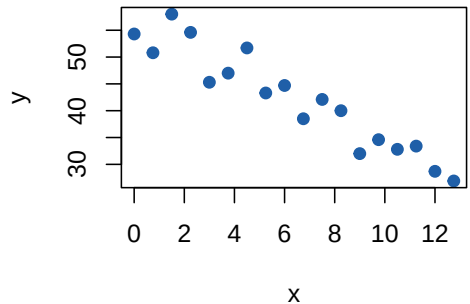
β_1: change in the mean of Y for a one-unit increase in X.

Assumptions (LINE) **L**inear mean, **I**ndependent errors, **N**ormal errors, **E**qual variance. Part 4 checks each one.

Direction matters Regressing Y on X is not the same as X on Y. Put the variable you want to explain on the left of ~.

An example with 18 observations

```r
x <- c(0, 0.75, 1.5, 2.25, 3, 3.75, 4.5, 5.25, 6,
       6.75, 7.5, 8.25, 9, 9.75, 10.5, 11.25, 12,
       12.75)
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)
cor(x, y)

## [1] -0.9487

par(mar = c(4, 4, 1, 1))
plot(x, y, pch = 19, col = col_main)
```



Illustrative workshop data: y at 18 equally spaced values of x. The points fall close to a straight line, $r = -0.949$.

Fitting the line with `lm()`

```
M1 <- lm(y ~ x)
display(M1, detail = TRUE)     # arm

## lm(formula = y ~ x)
##             coef.est coef.se t value Pr(>|t|)
## (Intercept)  56.42     1.39   40.53    0.00
## x            -2.24     0.19  -12.01    0.00
## ---
## n = 18, k = 2
## residual sd = 3.08, R-Squared = 0.90
```

`summary(M1)` gives the same numbers in the usual, longer layout.

Reading it Fitted line: $\hat{y} = 56.42 - 2.24\,x$. Each unit increase in x lowers the mean of y by 2.24 ($t = -12.01$, $p < 0.001$). $R^2 = 0.9$: the line explains 90% of the variation in y; the residual SD is 3.08.

Confidence intervals and the ANOVA table

```
round(confint(M1), 3)
print(anova(M1), signif.stars = FALSE)

##               2.5 % 97.5 %
## (Intercept) 53.466 59.369
## x           -2.633 -1.843
## 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
```

$H_0 : \beta_1 = 0$:
$t = b_1/\mathrm{SE}(b_1) \sim t_{n-2}$
SST = SSR + SSE, R^2 = SSR/SST
$F = \mathrm{MSR}/\mathrm{MSE} = t^2$

Slope: 95% CI (-2.63, -1.84). In simple regression the slope t test, the ANOVA F test and `cor.test()` for $\rho = 0$ coincide: $t = -12.01$ in `cor.test(x, y)`, and $t^2 = F = 144.2$.

Session 1 revisited: a t test is a regression

```r
bp <- read.csv("data/bp_data.csv")
fit_bp <- lm(mmhg ~ status, data = bp)
coef(summary(fit_bp))
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   112.92      2.209   51.11 8.965e-47
## status         31.31      2.991   10.47 1.660e-14
```

```r
t.test(mmhg ~ status, data = bp,
       var.equal = TRUE)$statistic
```

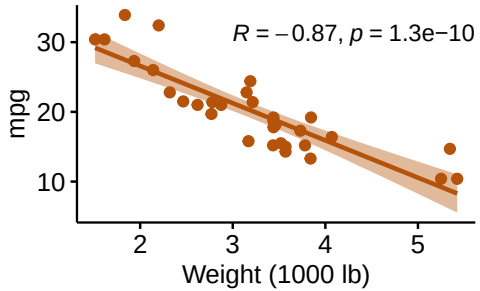
```
##      t
## -10.47
```

- `status`: 0 healthy, 1 hypertensive
- Intercept = healthy mean, 112.9 mmHg
- Slope = difference in means, 31.3 mmHg
- Same t as session 1's pooled test (the sign flips because `t.test()` takes group 0 minus group 1)

One framework Regression on a 0/1 predictor reproduces the pooled two-sample t test.

Weight and fuel economy of cars

```
M2 <- lm(mpg ~ wt, data = mtcars)
coef(summary(M2))

##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  37.285    1.8776  19.858 8.242e-19
## wt           -5.344    0.5591  -9.559 1.294e-10

ggscatter(mtcars, x = "wt", y = "mpg",
  add = "reg.line", conf.int = TRUE,
  cor.coef = TRUE, color = col_alt,
 cor.coeff.args = list(label.x = 3, label.y = 31),
  xlab = "Weight (1000 lb)", ylab = "mpg")
```



Reading it Each extra 1000 lb: about 5.34 mpg fewer; $R^2 = 0.75$.

PART 4

Regression diagnostics

What to check, and how

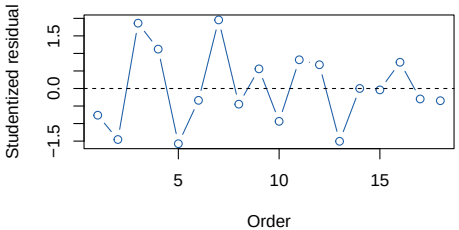
Assumption	Plot	Formal test	R
Linear mean	residuals vs fitted	–	`plot(M1, 1)`
Independent errors	residuals vs order	Durbin–Watson	`lmtest::dwtest()`
Equal variance	residuals vs fitted	Breusch–Pagan	`lmtest::bptest()`
Normal errors	normal Q–Q plot, histogram	Shapiro–Wilk	`shapiro.test()`
No dominant points	residuals vs leverage	Cook's distance	`plot(M1, 4)`

Plots first Tests answer "is there evidence of a problem?"; plots show *what* the problem is. With small n the tests have little power, with large n they flag trivial departures. `PASWR::checking.plots()` draws the main residual plots in one call.

Independence: Durbin–Watson test

```r
str_res <- studres(M1)      # MASS
par(mar = c(4, 4, 1, 1))
plot(1:18, str_res, type = "b", col = col_main,
     xlab = "Order", ylab = "Studentized residual")
abline(h = 0, lty = 2)
dw <- dwtest(M1)    # H1: positive autocorrelation
c(dw$statistic, p = dw$p.value)

##     DW      p
## 2.4492 0.7626
```

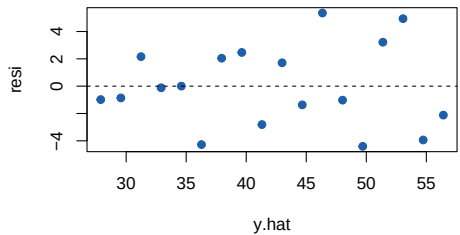
$$DW = \sum_{t=2}^{n}(e_t - e_{t-1})^2 \Big/ \sum_{t=1}^{n} e_t^2$$



$DW \approx 2(1 - \hat{\rho}_1)$, so values near 2 mean no lag-1 autocorrelation. $DW = 2.45$, $p = 0.76$: no positive autocorrelation. The test makes sense only when the order (time, or x) means something.

Constant variance: Breusch–Pagan test

```
resi <- residuals(M1)
y.hat <- fitted(M1)
par(mar = c(4, 4, 1, 1))
plot(y.hat, resi, pch = 19, col = col_main)
abline(h = 0, lty = 2)
bpt <- bptest(M1)       # H0: constant variance
c(bpt$statistic, p = bpt$p.value)

##      BP    p.BP
## 4.44138 0.03508
```

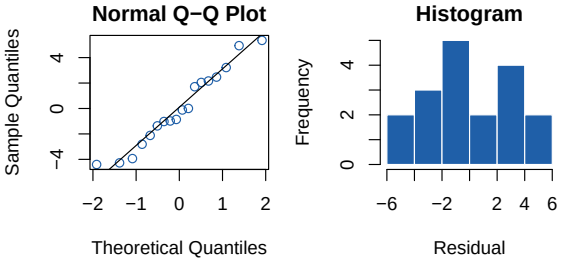


$BP = 4.44$ on 1 df, $p = 0.035$: some evidence that the spread changes with the fitted value. With 18 points, judge it together with the plot.

Remedies Transform y, weighted least squares, or robust standard errors.

Normality of the residuals

```
par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))
qqnorm(resi, col = col_main); qqline(resi)
hist(resi, col = col_main, border = "white",
     main = "Histogram", xlab = "Residual")
```




```
shapiro.test(resi)

##
##  Shapiro-Wilk normality test
##
## data:  resi
## W = 0.95673, p-value = 0.5399
```

$p = 0.54$: no evidence against normal errors.

A correction: the KS test with estimated parameters

```r
# KS with mean and sd taken from the data:
# p-value is too large (test is conservative)
ks.test(resi, "pnorm", mean = mean(resi),
        sd = sd(resi))$p.value

## [1] 0.9542

# Lilliefors correction for estimated
# parameters (nortest)
lillie.test(resi)$p.value

## [1] 0.7814
```

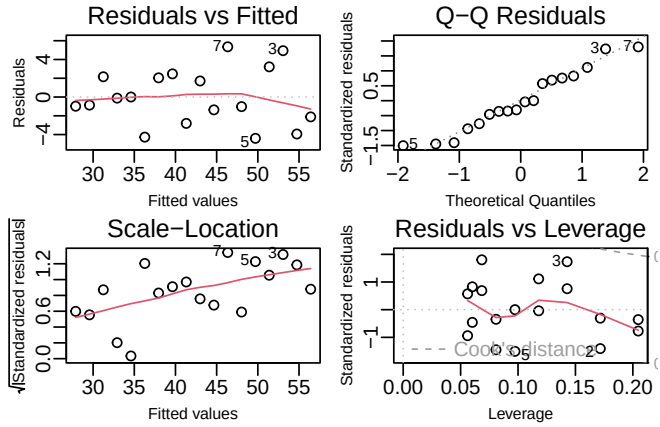
Two myths "Use Shapiro–Wilk below $n = 50$ and KS above" is not a rule: `shapiro.test()` works for $3 \leq n \leq 5000$. And the plain KS test assumes the mean and SD are *known*; estimating them from the same data needs the Lilliefors version.

All three tests agree here: the residuals look normal.

All the checks in one graph

```
par(mfrow = c(2, 2), cex = 0.75,
    mar = c(2.6, 2.6, 1.4, 0.4),
    mgp = c(1.5, 0.45, 0),
    cex.lab = 0.8)
plot(M1, cex.caption = 0.9)
```

- Residuals vs fitted: no curve
- Q–Q: close to the line
- Scale–location: mild rise (BP test)
- No point near Cook's distance 0.5



PART 5

Prediction, and practice

Advertising and sales

```
marketing <- data.frame(
  sales   = c(2, 4, 6, 9, 12, 34, 45),
  youtube = c(1, 2, 4, 7, 9, 11, 15))
model <- lm(sales ~ youtube, data = marketing)
coef(summary(model))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.364     4.6608  -1.151 0.301856
## youtube        3.052     0.5531   5.518 0.002677
```

```
summary(model)$r.squared
```

```
## [1] 0.8589
```

- Seven periods of advertising spend on YouTube and sales (illustrative)
- Slope 3.05: one more unit of spend goes with about 3.1 more units of sales, $p = 0.0027$
- The intercept is not significant and has no sensible meaning here (spend 0 lies outside the data)

Prediction, and the danger of extrapolation

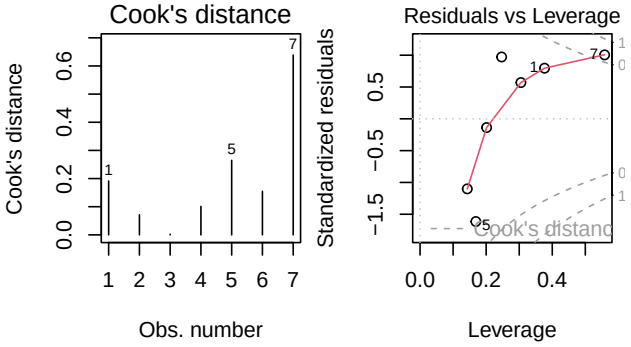
```
range(marketing$youtube)

## [1]  1 15

new <- data.frame(youtube = c(5, 0, 1000))
round(predict(model, new,
              interval = "prediction"), 1)

##      fit    lwr  upr
## 1    9.9   -9.2   29
## 2   -5.4  -26.7   16
## 3 3046.6 1634.5 4459
```

- Spend 5 lies inside the data (1 to 15): a sensible prediction, with a wide interval because $n = 7$
- Spend 0 gives *negative* sales
- Spend 1000 is 67 times the largest value observed; the straight line has no support there

Rule Predict only within the range of the data used to fit the model.

Influential points: Cook's distance and leverage

```
par(mfrow = c(1, 2), cex = 0.8,
    mar = c(4, 4, 2, 0.5))
plot(model, 4)   # Cook's distance
plot(model, 5, cex.caption = 0.85)
round(cooks.distance(model), 2)

##    1    2    3    4    5    6    7
## 0.19 0.07 0.00 0.10 0.26 0.15 0.64
```



Reading it Period 7: leverage 0.56, largest Cook's distance. With $n = 7$ the slope leans on it.

Exercise

Baseline blood pressure (`base_bp`) and plasma epinephrine concentration (`epi`) of 30 subjects, from the workshop assignment:

```
base_bp <- c(163, 195, 170, 171, 148, 195, 151, 166, 177, 165, 174, 164, 152, 202, 171,
             143, 166, 160, 170, 150, 159, 141, 160, 160, 162, 214, 179, 178, 198, 176)
epi <- c(248, 339, 193, 307, 80, 550, 70, 66, 129, 281, 296, 217, 88, 288, 265,
         279, 160, 263, 184, 227, 92, 139, 173, 224, 183, 441, 612, 401, 132, 130)
```

1. Draw the scatter diagram. Compute r from the formula with `sum()` and check it with `cor()`.
2. Test $H_0 : \rho = 0$ with `cor.test()`, then with `method = "spearman"`. Which points look unusual?
3. Fit `lm(epi ~ base_bp)`, interpret the slope, and run the residual checks of Part 4. Is the spread of `epi` constant?

Key take-aways

1. **Plot first**: a curve, an unpaired pairing or two odd marks can all hide behind a single r.
2. **Correlation needs pairs** on the same units, and it is not causation.
3. **Test what you estimate**: $\rho = 0$ and $\beta_1 = 0$ use the same t logic as session 1; report r or b_1 with a confidence interval.
4. **Check LINE**: residual plots first, then Durbin–Watson, Breusch–Pagan and Shapiro–Wilk (not a plain KS test).
5. **Predict inside the data**, and look for influential points before trusting the line.

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.
- Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to Linear Regression Analysis*, 6th Ed., Wiley, Hoboken.
- Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, Cambridge.
- 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.

Thank you

Questions and discussion