

Statistical Techniques using R

Session 2: Simple Linear Regression Modelling

Dr. Kapil Kumar

Assistant Professor, Department of Statistics
Central University of Haryana, Mahendergarh, India
kapilstats@gmail.com

29 June 2018

One-month Faculty Induction Training Programme for faculty of universities, colleges and institutes (14 June – 14 July 2018)

Pandit Madan Mohan Malaviya National Mission on Teachers and Teaching (PMMMNTT),
Central University of Haryana, Mahendergarh

Keynote lecture

Where we left off

In Session 1 we learnt to

- build vectors and data frames, and read files with `read.csv("data/...")`
- draw graphs and write small functions
- compare groups with t -tests and a one-way ANOVA
- fit a first line with `lm()` and read it with `display()`

An open question For the trees data, `lm(Volume ~ Girth)` gave $R^2 = 0.94$, yet the points bent upwards at both ends. Is the straight line adequate, and what should we do if it is not?

Today the simple linear regression model properly: assumptions, estimation, inference, prediction and diagnostics, ending with an answer to that question.

Plan of the session

- 1 Regression ideas: models and assumptions
- 2 Least squares, by hand and with `lm()`
- 3 Inference: tests, intervals, ANOVA, R^2
- 4 Confidence and prediction intervals
- 5 Diagnostics and remedies
- 6 Workflow and take-aways

By the end you can state the simple linear regression model and its assumptions, compute the least-squares estimates by hand and with `lm()`, test and estimate the coefficients, read the ANOVA table and R^2 , tell a confidence interval from a prediction interval, and check the residuals and choose a remedy.

- 1 Regression ideas
- 2 Least squares
- 3 Inference
- 4 Prediction
- 5 Diagnostics and remedies
- 6 Workflow and take-aways

- Statistics is the science of learning from data
- Data carry information, but it has to be extracted with the right method
- Statistics forecasts with a stated uncertainty, not like an astrologer's parrot
- It cannot do everything, and good tools take time to develop

<i>Decision</i>	Correct data	Wrong data
Correct tool	correct	incorrect
Wrong tool	incorrect	incorrect

Moral A good decision needs *both* sound data and a method whose assumptions hold. Today's last section is about checking the second.

What is regression analysis?

- Models the relationship between a **dependent** (response) variable y and one or more **independent** (explanatory) variables x
- Aim: describe the process as well as possible, estimate effects and predict

Typical questions:

- What does one more year of education add to income?
- How will a tax on cigarettes change smoking in a state?
- How long will a patient survive with and without a treatment?

Where the name comes from

Francis Galton (1880s) found that children of very tall parents are tall, but less extreme than their parents: heights “regress” towards the mean. The name stuck to the method.

A model has two ingredients: *variables*, which we observe, and *parameters*, which we estimate from the data.

Mathematical and statistical models

Effect of fertilizer x on yield per acre y , if the relation were exactly $y = 2 + 3x$:

Fertilizer x	1	2	5	10	50	100	1000
Yield $y = 2 + 3x$	5	8	17	32	152	302	3002

Mathematical: $y = 2 + 3x$ holds exactly, for every x . More fertilizer always gives more yield.

Statistical: $y = \beta_0 + \beta_1 x + \varepsilon$, with random scatter, valid only over the range studied. Beyond some level the crop is damaged.

A statistical model is an approximation that must be checked against data, and it says nothing about values of x far outside the data.

Linear in the parameters

Model	$\partial y / \partial (\text{parameter})$	
$y = \beta_0 + \beta_1 x$	$1, x$	linear
$y = \beta_0 + \beta_1 x^2$	$1, x^2$	linear
$y = \beta_0 + \beta_1 \log x$	$1, \log x$	linear
$y = \exp(\beta_1 x)$	$x \exp(\beta_1 x)$	nonlinear
$y = \beta_0^2 + \beta_1 x$	$2\beta_0, x$	nonlinear

Rule A model is *linear* when no derivative with respect to a parameter depends on any parameter.

So “linear regression” means linear in the parameters: the variables may be squared or logged. Linear models are preferred because their estimates have closed forms and exact theory.

The simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$
$$E(y \mid x) = \beta_0 + \beta_1 x$$

- y : response (study, dependent) variable
- x : explanatory (independent) variable
- β_0 : intercept, the mean of y at $x = 0$
- β_1 : slope, the change in mean y per unit increase in x

The error term ε collects the effect of all factors left out of the model, randomness in behaviour and responses, measurement error, and qualitative influences.

The fitted model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ replaces the unknown parameters by estimates. How good it is depends on how well the parameters are estimated.

Assumptions

- 1 **Linearity:** $E(\varepsilon_i) = 0$, so the mean of y is a straight line in x
- 2 **Independence:** the errors are independent
- 3 **Constant variance:** $\text{Var}(\varepsilon_i) = \sigma^2$ for every x (homoscedasticity)
- 4 **Normality:** the errors are normal; needed for exact t and F inference in small samples
- 5 x is fixed, or measured without error

$$\varepsilon_1, \dots, \varepsilon_n \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Not assumed that x or y themselves are normally distributed, or that x and y have equal variances. The assumptions are about the *errors*. Collinearity among regressors is a concern of multiple regression only.

- 1 Regression ideas
- 2 Least squares**
- 3 Inference
- 4 Prediction
- 5 Diagnostics and remedies
- 6 Workflow and take-aways

Packages used in this session

- Base R: `lm()`, `summary()`, `confint()`, `anova()`, `predict()`, `plot()` for a fitted model
- **lmtest**: `bptest()` for non-constant variance, `resettest()` for a missing curve
- **MASS**: `boxcox()` to choose a transformation

```
# install once (needs internet)
# install.packages(c("lmtest", "MASS"))
library(lmtest)
library(MASS)
# colours used in all our plots
col_main <- "#3333B3"
col_alt  <- "#B45309"
```

The rocket propellant data

- A rocket motor is made by bonding an igniter propellant and a sustainer propellant inside a metal housing
- The **shear strength** of the bond (psi) is a key quality characteristic
- It is suspected to depend on the **age** (weeks) of the sustainer propellant batch
- 20 batches; a classic textbook example (Montgomery, Peck and Vining)
- No missing values or duplicates; values are plausible

```
rocket <- read.csv("data/rocket_propellant.csv")  
str(rocket)
```

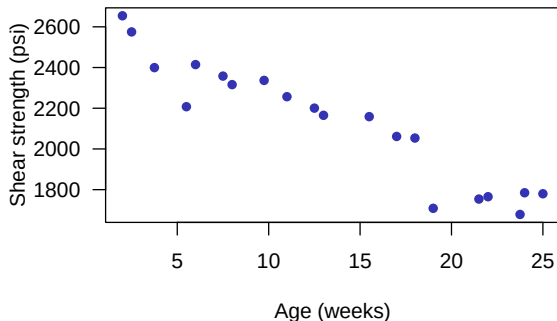
```
## 'data.frame': 20 obs. of 2 variables:  
## $ shear: num 2159 1678 2316 2061 2208 ...  
## $ age : num 15.5 23.8 8 17 5.5 ...
```

```
summary(rocket)
```

```
##      shear      age  
## Min.   :1678   Min.   : 2.00  
## 1st Qu.:1784   1st Qu.: 7.12  
## Median :2183   Median :12.75  
## Mean   :2131   Mean   :13.36  
## 3rd Qu.:2342   3rd Qu.:19.62  
## Max.   :2654   Max.   :25.00
```

Look first: the scatter plot

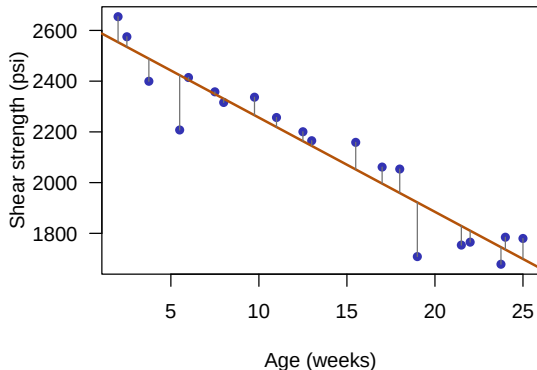
```
par(mar = c(4, 4, 1, 1), las = 1)
plot(shear ~ age, data = rocket, pch = 16,
     col = col_main, xlab = "Age (weeks)",
     ylab = "Shear strength (psi)")
```



Reading it A strong, decreasing and roughly straight-line pattern: older propellant bonds less strongly. The correlation, `cor(rocket$age, rocket$shear)`, is $r = -0.95$.

Keep an eye on two batches, near 5 and 19 weeks, that lie well below the others. We return to them in the diagnostics.

The least-squares principle



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

- Choose $\hat{\beta}_0, \hat{\beta}_1$ that make S as small as possible
- The grey segments are the vertical distances being squared: we predict y from x (“direct” regression)
- Under normal errors, least squares is also maximum likelihood

Normal equations and their solution

$$\frac{\partial S}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Rightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$S_{xx} = \sum (x_i - \bar{x})^2$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

A minimum? Yes: the matrix of second derivatives is positive definite whenever $S_{xx} > 0$, that is, whenever the x values are not all equal.

Least squares by hand in R

```
x <- rocket$age
y <- rocket$shear
n <- length(x)
Sxx <- sum((x - mean(x))^2)
Sxy <- sum((x - mean(x)) * (y - mean(y)))
b1 <- Sxy / Sxx
b0 <- mean(y) - b1 * mean(x)
c(xbar = mean(x), ybar = mean(y), Sxx = Sxx,
  Sxy = Sxy, b0 = b0, b1 = b1)

##      xbar      ybar      Sxx      Sxy      b0
##    13.36   2131.36   1106.56 -41112.65  2627.82
##      b1
##   -37.15
```

Result

$$\hat{\beta}_1 = \frac{-41112.7}{1106.56} = -37.154$$

$$\hat{\beta}_0 = 2131.36 + 37.154 \times 13.36 \\ = 2627.82$$

A few lines of R reproduce the textbook formulas exactly.

The same fit with `lm()`

```
fit <- lm(shear ~ age, data = rocket)
coef(fit)

## (Intercept)      age
##      2627.82     -37.15

all.equal(unname(coef(fit)), c(b0, b1))

## [1] TRUE

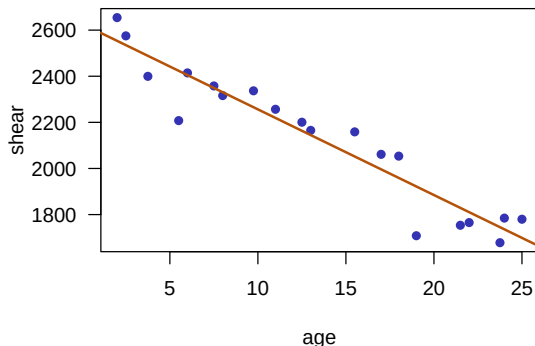
names(fit)

## [1] "coefficients" "residuals"      "effects"
## [4] "rank"         "fitted.values"  "assign"
## [7] "qr"          "df.residual"    "xlevels"
## [10] "call"         "terms"         "model"
```

- $\text{shear} \sim \text{age}$: response on the left, explanatory variable on the right
- The fitted model is a list; use extractor functions rather than `$`: `coef()`, `fitted()`, `residuals()`, `summary()`, `confint()`, `anova()`, `predict()`

Interpreting the fitted line

```
par(mar = c(4, 4, 1, 1), las = 1)
plot(shear ~ age, data = rocket, pch = 16,
     col = col_main)
abline(fit, col = col_alt, lwd = 2)
```



$$\widehat{\text{shear}} = 2627.82 - 37.15 \times \text{age}$$

Reading it Each additional week of age lowers the mean shear strength by about 37.2 psi. The intercept estimates the strength of freshly made propellant; age 0 lies just outside the data (2–25 weeks).

At 3 weeks: $2627.82 - 37.15 \times 3 = 2516.4$ psi.

Residuals and their properties

```
e <- residuals(fit)           # y - fitted
round(c(sum(e), sum(x * e)), 8)

## [1] 0 0

c(sum(fitted(fit)), sum(y))

## [1] 42627 42627

c(predict(fit, data.frame(age = mean(x))),
  mean(y))

##      1
## 2131 2131
```

$$e_i = y_i - \hat{y}_i$$

$$\sum e_i = 0, \quad \sum x_i e_i = 0$$

$$\sum \hat{y}_i = \sum y_i$$

the line passes through $(\bar{x}, \bar{y})$

All four facts are just the normal equations rewritten. They hold for every least-squares line, good or bad, so they say nothing about whether the model fits.

Estimating the error variance σ^2

```
SSE <- sum(e^2)           # residual SS
s2 <- SSE / (n - 2)        # mean square error
c(SSE = SSE, s2 = s2, s = sqrt(s2))

##      SSE      s2      s
## 166254.86  9236.38  96.11

summary(fit)$sigma        # the same s

## [1] 96.11
```

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \frac{\text{SSE}}{n-2} = \text{MSE}\end{aligned}$$

Two parameters were estimated, so $n - 2$ degrees of freedom remain; dividing by $n - 2$ makes $\hat{\sigma}^2$ unbiased.

Reading it $\hat{\sigma} = 96.1$ psi, the *residual standard error*: a typical batch lies about 96 psi above or below the line.

- 1 Regression ideas
- 2 Least squares
- 3 Inference**
- 4 Prediction
- 5 Diagnostics and remedies
- 6 Workflow and take-aways

Sampling distributions and standard errors

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$$

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2 [1/n + \bar{x}^2/S_{xx}])$$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \quad se(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-2}, \quad \frac{SSE}{\sigma^2} \sim \chi_{n-2}^2$$

```
se_b1 <- sqrt(s2 / Sxx)
se_b0 <- sqrt(s2 * (1/n + mean(x)^2 / Sxx))
c(se_b0 = se_b0, se_b1 = se_b1)
```

```
## se_b0 se_b1
## 44.184 2.889
```

The slope is estimated more precisely when the x values are spread out (large S_{xx}). In practice t and χ^2 quantiles come from `qt()` and `qchisq()` rather than printed tables.

$H_0 : \beta_1 = \beta_{10}$ versus $H_1 : \beta_1 \neq \beta_{10}$

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\text{se}(\hat{\beta}_1)} \sim t_{n-2} \text{ under } H_0$$

Reject H_0 if $|t_0| > t_{\alpha/2, n-2}$, i.e. if $p < \alpha$

Usually $\beta_{10} = 0$: “is there a linear relationship at all?” The test for β_0 is built the same way. (If σ^2 were known, t_0 would be a z statistic compared with $N(0, 1)$.)

```
t1 <- b1 / se_b1
p1 <- 2 * pt(-abs(t1), df = n - 2)
round(c(t = t1, t_crit = qt(0.975, df = n - 2)), 3)

##          t    t_crit
## -12.860    2.101

coef(summary(fit)) # p1 is the last entry

##          Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2627.82    44.184   59.47 4.064e-22
## age          -37.15     2.889  -12.86 1.643e-10
```

Reading the t -tests

- Slope: $t_0 = -12.86$ on 18 df, far beyond the critical value ± 2.1 ; $p \approx 1.6 \times 10^{-10}$. There is strong evidence that shear strength falls as the propellant ages.
- Intercept: $H_0 : \beta_0 = 0$ (“new propellant has zero strength”) is not a sensible question here, even though software always prints this test.
- The t value is the estimate measured in standard errors: the slope lies 12.9 standard errors away from zero.

Careful Failing to reject $H_0 : \beta_1 = 0$ does not prove that x is unimportant: the effect may be small, curved rather than straight, or the sample too small to detect it.

Confidence intervals for β_0 and β_1

$$100(1 - \alpha)\% \text{ CI: } \hat{\beta}_j \pm t_{\alpha/2, n-2} \text{ se}(\hat{\beta}_j)$$

```
b1 + c(-1, 1) * qt(0.975, n - 2) * se_b1
```

```
## [1] -43.22 -31.08
```

```
confint(fit, level = 0.95)
```

```
##           2.5 %   97.5 %
```

```
## (Intercept) 2535.00 2720.65
```

```
## age        -43.22  -31.08
```

Reading it With 95% confidence the mean shear strength falls by between 31.1 and 43.2 psi per week of age. The interval excludes 0, matching the t -test.

“95%” describes the method: in repeated samples, 95% of intervals built this way would contain the true β_1 .

A confidence interval for σ^2

$$\frac{SSE}{\chi^2_{\alpha/2, n-2}} \leq \sigma^2 \leq \frac{SSE}{\chi^2_{1-\alpha/2, n-2}}$$

$\chi^2_{\alpha/2, \nu}$: the point with area $\alpha/2$ to its right

```
ci_s2 <- SSE / qchisq(c(0.975, 0.025), df = n - 2)
ci_s2      # for sigma^2

## [1] 5274 20199

sqrt(ci_s2)      # for sigma

## [1] 72.62 142.12
```

Reading it σ lies between 72.6 and 142.1 psi with 95% confidence. The interval is not symmetric about $\hat{\sigma} = 96.1$ because the χ^2 distribution is skewed.

Caution Unlike the t intervals, this interval relies heavily on normal errors.

$$SST = SSR + SSE$$

$$SST = \sum (y_i - \bar{y})^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$\text{df: } n - 1 = 1 + (n - 2)$$

$$F_0 = \frac{SSR/1}{SSE/(n-2)}$$

$$F_0 \sim F_{1, n-2} \text{ if } \beta_1 = 0$$

```
print(anova(fit), signif.stars = FALSE)

## Analysis of Variance Table
##
## Response: shear
##           Df Sum Sq Mean Sq F value Pr(>F)
## age         1 1527483 1527483    165 1.6e-10
## Residuals  18  166255    9236

t1^2           # F = t^2 in simple regression

## [1] 165.4
```

Reading it SSR = 1,527,483 is explained by age, SSE = 166,255 is left over. $F_0 = 165.4 = t_0^2$: the same test as the slope's t -test.

Coefficient of determination R^2

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$
$$R^2_{\text{adj}} = 1 - \frac{SSE/(n-2)}{SST/(n-1)}$$

```
SST <- sum((y - mean(y))^2)
c(R2 = 1 - SSE / SST, r_squared = cor(x, y)^2,
  adjR2 = 1 - (SSE / (n - 2)) / (SST / (n - 1)))
```

```
##      R2 r_squared      adjR2
## 0.9018 0.9018 0.8964
```

Reading it Age explains 90.2% of the variation in shear strength. In simple regression $R^2 = r^2$.

What R^2 does not say A high R^2 does not prove that a straight line is the right shape, nor that x causes y . It also grows with the spread of the x values sampled.

Everything at once: summary()

```
print(summary(fit), signif.stars = FALSE)

##
## Call:
## lm(formula = shear ~ age, data = rocket)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -216.0   -50.7    28.7    66.6   106.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2627.82     44.18    59.5  < 2e-16
## age          -37.15       2.89   -12.9  1.6e-10
##
## Residual standard error: 96.1 on 18 degrees of freedom
## Multiple R-squared:  0.902, Adjusted R-squared:  0.896
## F-statistic: 165 on 1 and 18 DF,  p-value: 1.64e-10
```

- Estimate: $\hat{\beta}_0, \hat{\beta}_1$
- Std. Error: $se(\hat{\beta}_j)$
- t value, $Pr(>|t|)$: test of $\beta_j = 0$
- Residual standard error: $\hat{\sigma}$ on $n - 2$ df
- Multiple R-squared: R^2
- F-statistic: the ANOVA F test

- 1 Regression ideas
- 2 Least squares
- 3 Inference
- 4 Prediction**
- 5 Diagnostics and remedies
- 6 Workflow and take-aways

Two questions at a new value x_0

Point estimate: $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

Confidence interval for the **mean** response $E(y \mid x_0)$: $\hat{y}_0 \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

Prediction interval for a **new** observation y_0 : $\hat{y}_0 \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

Confidence interval Where the *average* strength of all batches aged x_0 weeks lies.

Prediction interval Where the strength of *one* new batch aged x_0 will fall. The extra 1 is that batch's own error, so the interval never shrinks to zero.

Both are narrowest at $x_0 = \bar{x}$ and widen as x_0 moves away from it.

predict() and a check by hand

```
new <- data.frame(age = c(3, 13, 20))
predict(fit, new, interval = "confidence")
```

```
##      fit   lwr   upr
## 1 2516 2439 2594
## 2 2145 2100 2190
## 3 1885 1824 1945
```

```
predict(fit, new, interval = "prediction")
```

```
##      fit   lwr   upr
## 1 2516 2300 2733
## 2 2145 1938 2352
## 3 1885 1674 2096
```

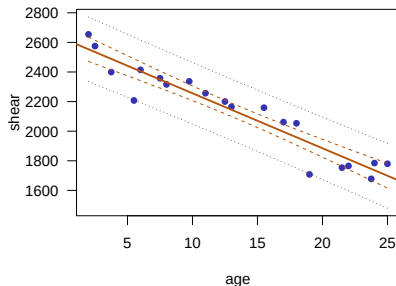
```
x0 <- 3
y0 <- b0 + b1 * x0
h0 <- 1/n + (x0 - mean(x))^2 / Sxx
y0 + c(-1, 1) * qt(0.975, n - 2) *
  sqrt(s2 * (1 + h0)) # PI at 3

## [1] 2300 2733
```

The hand computation reproduces the first prediction interval.

Confidence and prediction bands

```
grid <- data.frame(age = seq(2, 25, by = 0.5))
cb <- predict(fit, grid, interval = "confidence")
pb <- predict(fit, grid, interval = "prediction")
par(mar = c(4, 4, 1, 1), las = 1)
plot(shear ~ age, data = rocket, pch = 16,
     col = col_main, ylim = range(pb))
abline(fit, col = col_alt, lwd = 2)
matlines(grid$age, cb[, 2:3], lty = 2, col = col_alt)
matlines(grid$age, pb[, 2:3], lty = 3, col = "grey30")
```



Reading it At 3 weeks the *mean* strength is 2516 psi, 95% CI 2439–2594; a *single* new batch: 95% PI 2300–2733 psi.

Do not extrapolate At 60 weeks the line would predict 399 psi, but no batch older than 25 weeks was tested.

- 1 Regression ideas
- 2 Least squares
- 3 Inference
- 4 Prediction
- 5 Diagnostics and remedies**
- 6 Workflow and take-aways

Residuals, leverage and influence

$$\text{Leverage: } h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

$$\text{Standardized: } r_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}$$

$$\text{Studentized: } t_i = \frac{e_i}{\hat{\sigma}_{(i)} \sqrt{1 - h_{ii}}}$$

$$\text{Cook's distance: } D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}, \quad p = 2$$

$\hat{\sigma}_{(i)}$: residual standard error with case i left out.

Rules of thumb $|t_i| > 2$
deserves a look, $|t_i| > 3$ is
suspicious.

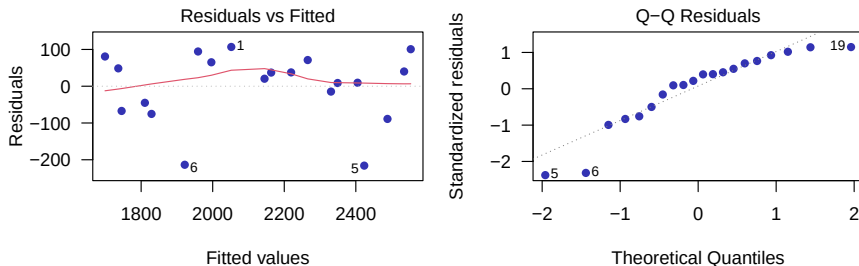
$h_{ii} > 2p/n$ marks an unusual x
value.

D_i near or above 1 marks a case
that moves the fit.

An *outlier* has an unusual y for its x ; a
high-leverage point has an unusual
 x ; an *influential* point changes the
fitted line when removed.

Residual plots

```
par(mfrow = c(1, 2), mar = c(4, 4, 2, 1), las = 1)
plot(fit, which = 1:2, pch = 16, col = col_main)
```



- **Residuals vs fitted:** no curve and an even spread, so linearity and constant variance look reasonable
- **Normal Q-Q:** batches 5 and 6 pull the lower tail away from the line

```
shapiro.test(residuals(fit))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit)
## W = 0.88, p-value = 0.01

bptest(fit)          # Breusch-Pagan

##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 0.013, df = 1, p-value = 0.9
```

Reading it Shapiro–Wilk:
 $p = 0.014$, so normality of the errors is rejected at 5%, driven by batches 5 and 6.

Breusch–Pagan: $p = 0.91$, no evidence that the variance changes with age.

The batches were not tested in time order, so a Durbin–Watson test for autocorrelation has no meaning here.

Which batches stand out?

```
infl <- data.frame(resid = e,
                   rstudent = rstudent(fit),
                   hat = hatvalues(fit),
                   cook = cooks.distance(fit))
round(infl[abs(infl$rstudent) > 2, ], 3)

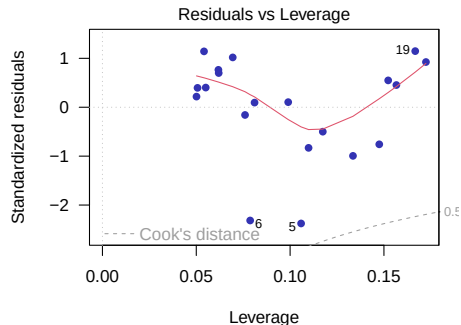
##      resid rstudent    hat   cook
## 5 -216.0    -2.788 0.106 0.334
## 6 -213.6    -2.686 0.079 0.229

2 * 2 / n           # leverage cut-off 2p/n

## [1] 0.2
```

Large residuals, but ordinary leverage and Cook's distances well below 1.

```
par(mar = c(4, 4, 2, 1), las = 1)
plot(fit, which = 5, pch = 16,
     col = col_main)
```



Does it matter? Refit without batches 5 and 6

```
fit56 <- lm(shear ~ age, data = rocket[-c(5, 6), ])
rbind(all = coef(fit), without_5_6 = coef(fit56))

##              (Intercept)      age
## all                2628 -37.15
## without_5_6        2659 -37.69

c(summary(fit)$sigma, summary(fit56)$sigma)

## [1] 96.11 62.97

shapiro.test(residuals(fit56))$p.value

## [1] 0.3798
```

Reading it The slope hardly moves, but $\hat{\sigma}$ drops from 96.1 to 63 psi and normality is no longer rejected. The two batches widen every interval without steering the line.

Delete only with a reason such as a testing error. Otherwise keep the case, report both fits, or use robust regression (`MASS::rlm()`).

Remedies when an assumption fails

Problem	What you see	Remedy
Curvature	a U or an arch in the residuals	transform x , or add x^2
Non-constant variance	a funnel in the residuals	transform y ; weighted LS
Non-normal errors	a curved Q-Q plot	transform y ; robust methods
Outliers	$ t_i > 3$, large D_i	check the data; robust fit
Dependent errors	a pattern over time	time-series models

After any remedy refit, and check the residuals again. Then interpret the model on its new scale.

Back to the trees data from Session 1

```
m1 <- lm(Volume ~ Girth, data = trees)
bptest(m1)$p.value      # constant variance?

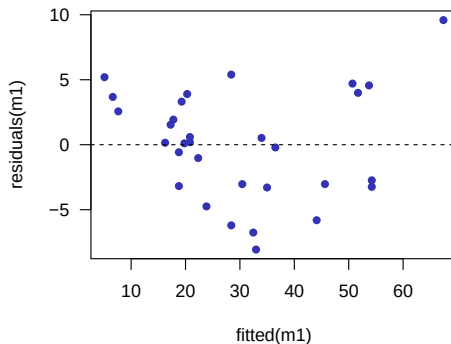
##          BP
## 0.01776

resettest(m1)$p.value   # missing curvature?

## [1] 0.0005878
```

Reading it The residuals form a clear U (RESET $p = 0.0006$) and fan out ($p = 0.018$): the straight line on the original scale fails two assumptions despite $R^2 = 0.94$.

```
par(mar = c(4, 4, 1, 1), las = 1)
plot(fitted(m1), residuals(m1), pch = 16,
     col = col_main)
abline(h = 0, lty = 2)
```



Box-Cox: letting the data choose a transformation

$$y^{(\lambda)} = \frac{y^\lambda - 1}{\lambda} \quad (\lambda \neq 0), \quad y^{(0)} = \log y$$

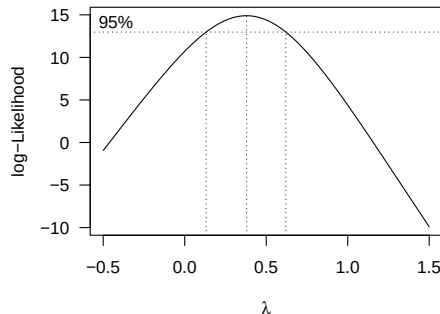
$\lambda = 1$: none; $\frac{1}{2}$: square root; $\frac{1}{3}$: cube root; 0: log

```
par(mar = c(4, 4, 1, 1), las = 1)
bc <- boxcox(m1, lambda = seq(-0.5, 1.5, 0.01))
bc$x[which.max(bc$y)]           # best lambda

## [1] 0.38

range(bc$x[bc$y > max(bc$y) - qchisq(0.95, 1)/2])

## [1] 0.14 0.62
```



The 95% interval for λ (last line) excludes 1 and 0 but contains $\frac{1}{3}$. Geometry agrees: the cube root of a volume is a length, which can grow linearly with diameter.

Refit on the cube-root scale

```
m2 <- lm(I(Volume^(1/3))) ~ Girth, data = trees)
round(coef(summary(m2))[, 1:3], 4)
```

```
##           Estimate Std. Error t value
## (Intercept)  0.8123      0.0904   8.985
## Girth        0.1668      0.0066  25.095
```

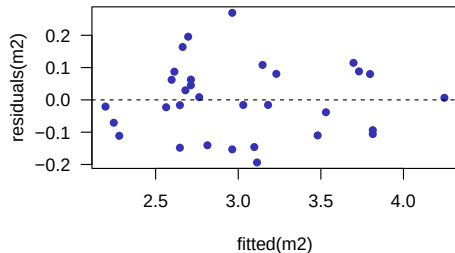
```
bptest(m2)$p.value; resettest(m2)$p.value
```

```
##      BP
## 0.7975
## [1] 0.6403
```

```
shapiro.test(residuals(m2))$p.value
```

```
## [1] 0.6286
```

```
par(mar = c(4, 4, 1, 1), las = 1)
plot(fitted(m2), residuals(m2), pch = 16,
     col = col_main)
abline(h = 0, lty = 2)
```



The U-shape and the funnel are gone;
no check rejects.

Predicting on the original scale

```
small <- data.frame(Girth = 8.5)  # a thin tree
p1 <- predict(m1, small, interval = "prediction")
p2 <- predict(m2, small, interval = "prediction")^3
round(rbind(straight = p1, cube_root = p2), 1)

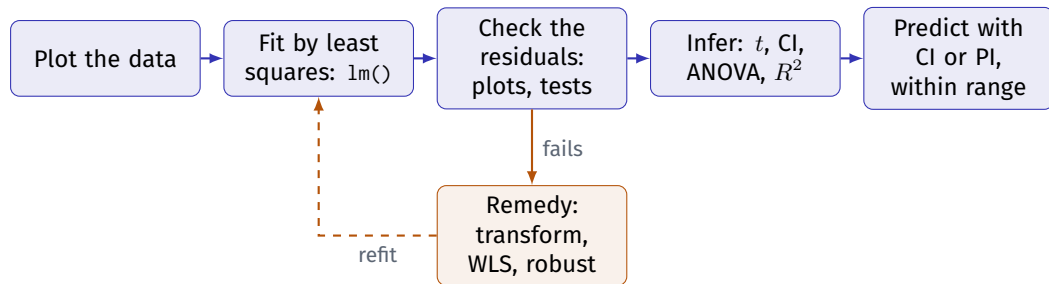
##      fit  lwr  upr
## 1   6.1 -3.0 15.3
## 1  11.1  7.8 15.2
```

Reading it Cubing gives an interval for the volume itself. The straight line predicts 6.1, with an impossible lower limit of -3 ; the cube-root model predicts 11.1, close to the three thinnest trees (about 10.3 cubic feet).

Answer to the open question No, the straight line was not adequate. A transformation suggested by Box-Cox, and by geometry, gives a model that passes its checks and makes sensible predictions.

- 1 Regression ideas
- 2 Least squares
- 3 Inference
- 4 Prediction
- 5 Diagnostics and remedies
- 6 Workflow and take-aways**

The regression workflow



Inference and prediction come *after* the residual checks: a p -value from a model that fails its assumptions is not worth reporting.

Key take-aways

- 1 **Linear** means linear in the parameters; x and y may be transformed.
- 2 **The assumptions concern the errors:** linear mean, independence, constant variance, normality.
- 3 **Least squares** has closed-form estimates, $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, which `lm()` reproduces.
- 4 **Report intervals**, not only p -values; in simple regression $F = t^2$ and $R^2 = r^2$.
- 5 **A prediction interval** for one new case is always wider than the confidence interval for the mean; never extrapolate far.
- 6 **Check residuals before you trust anything**, and fix problems with a transformation or a robust fit rather than by deleting data.

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