GENERAL LINEAR MODELS

2 NOVEMBER 2018

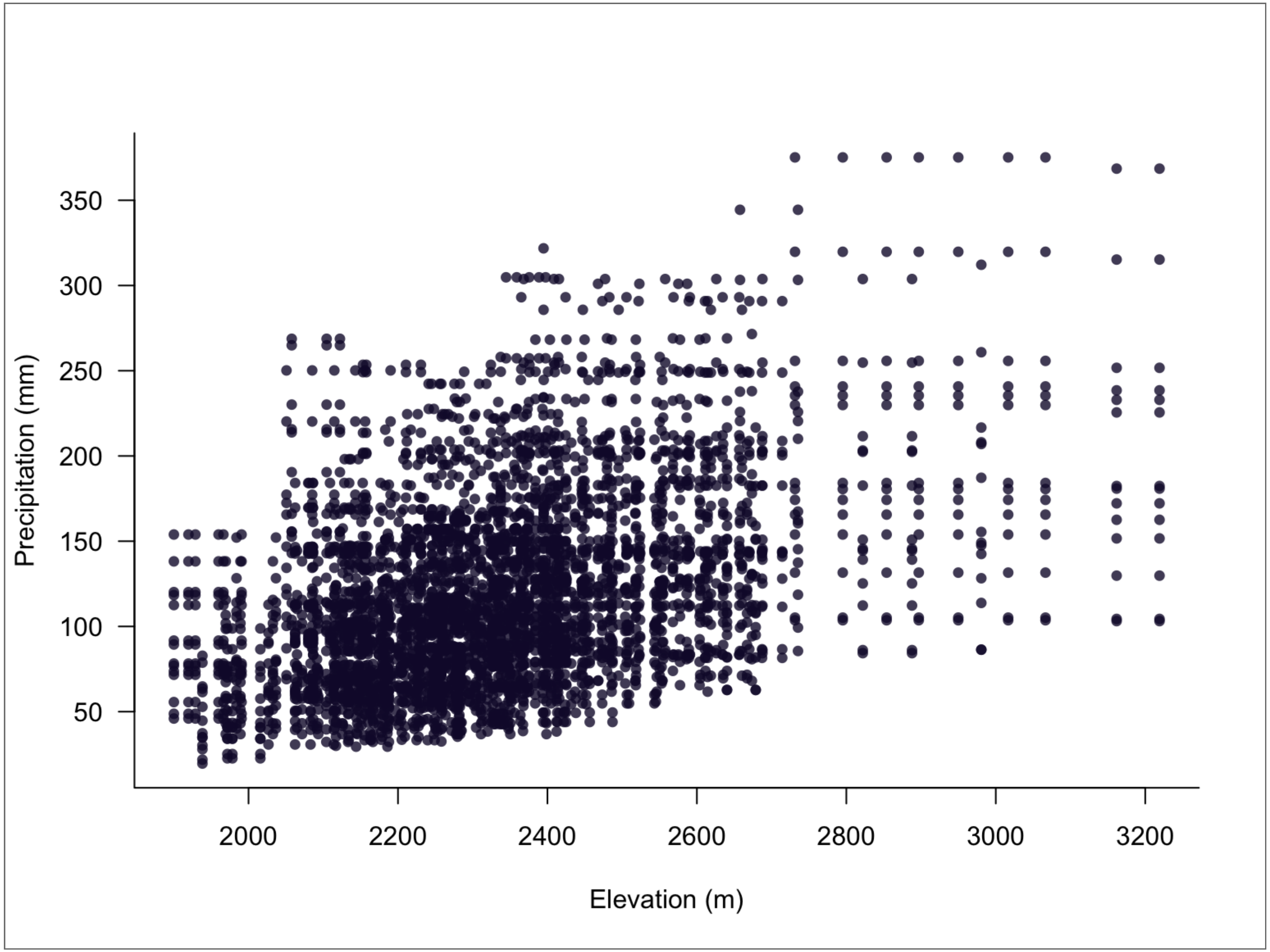
WORKSHOP OVERVIEW

- **https://github.com/goldingn/linear_models_workshop**
- general linear models
- mixed effects models
- generalised linear models
- Bayesian inference

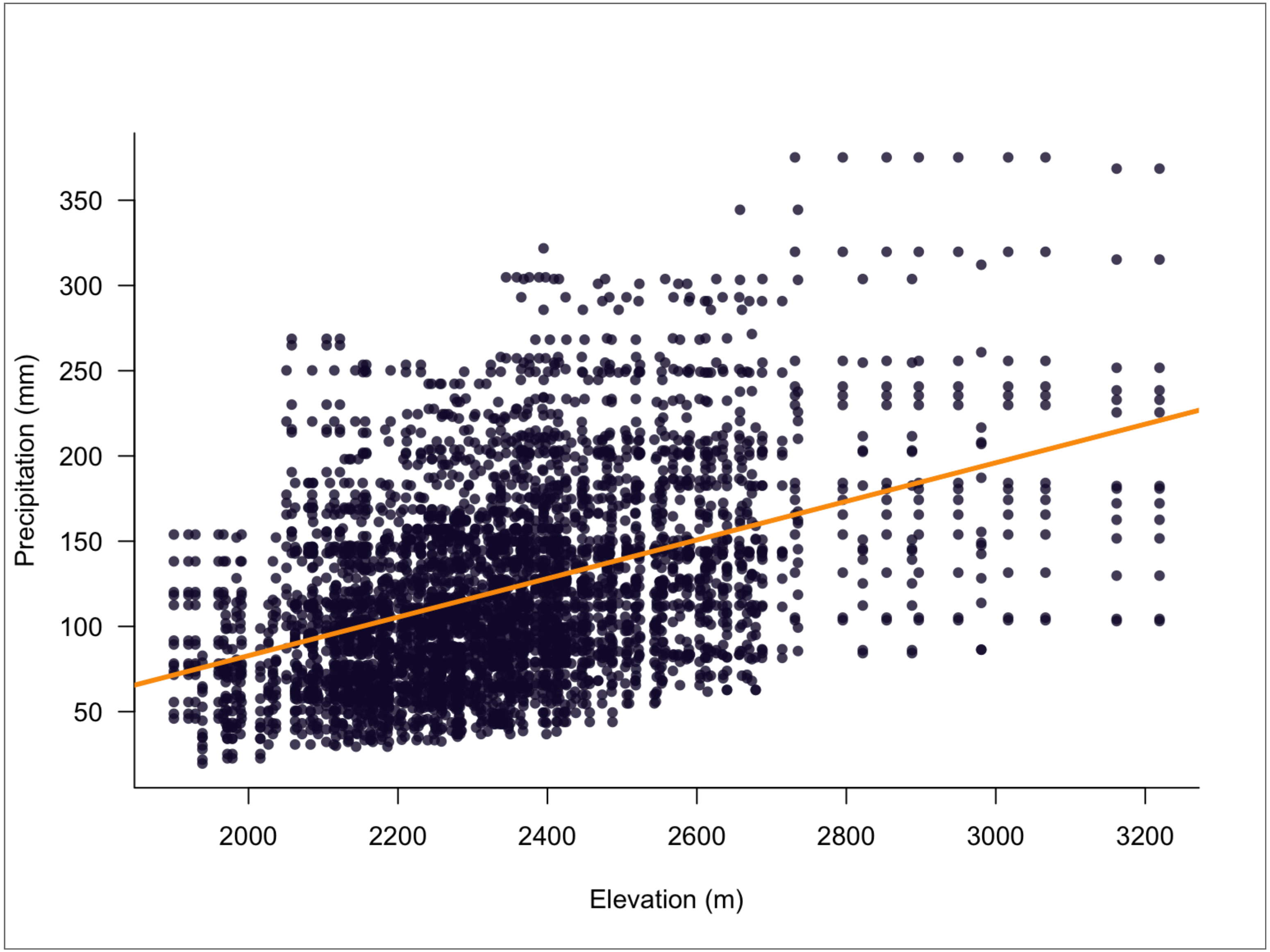
EXPECTED OUTCOMES

- learn some new terms
- identify appropriate models for your data
- understand assumptions of common models
- know where to look for help

AN EXAMPLE


Precipitation (mm)
350
300
250
200
150
100
50
2000
2200
2400
2600
2800
3000
3200
Elevation (m)

AN EXAMPLE



350
300
250
200
150
100
50
Precipitation (mm)
2000
2200
2400
2600
2800
3000
3200
Elevation (m)

LINEAR REGRESSION

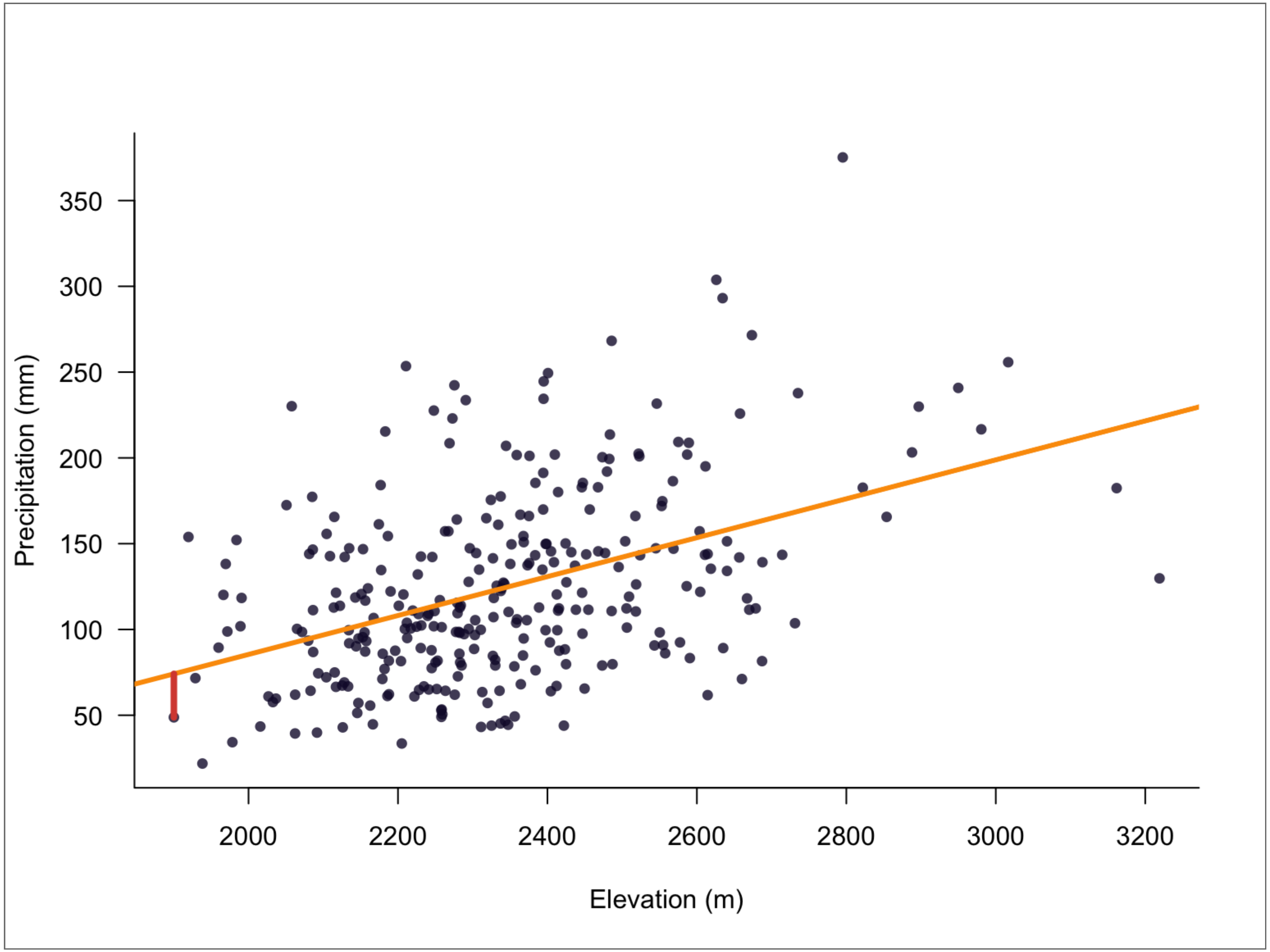
- what characterises this example?
  - continuous response
  - continuous predictor

LINEAR REGRESSION

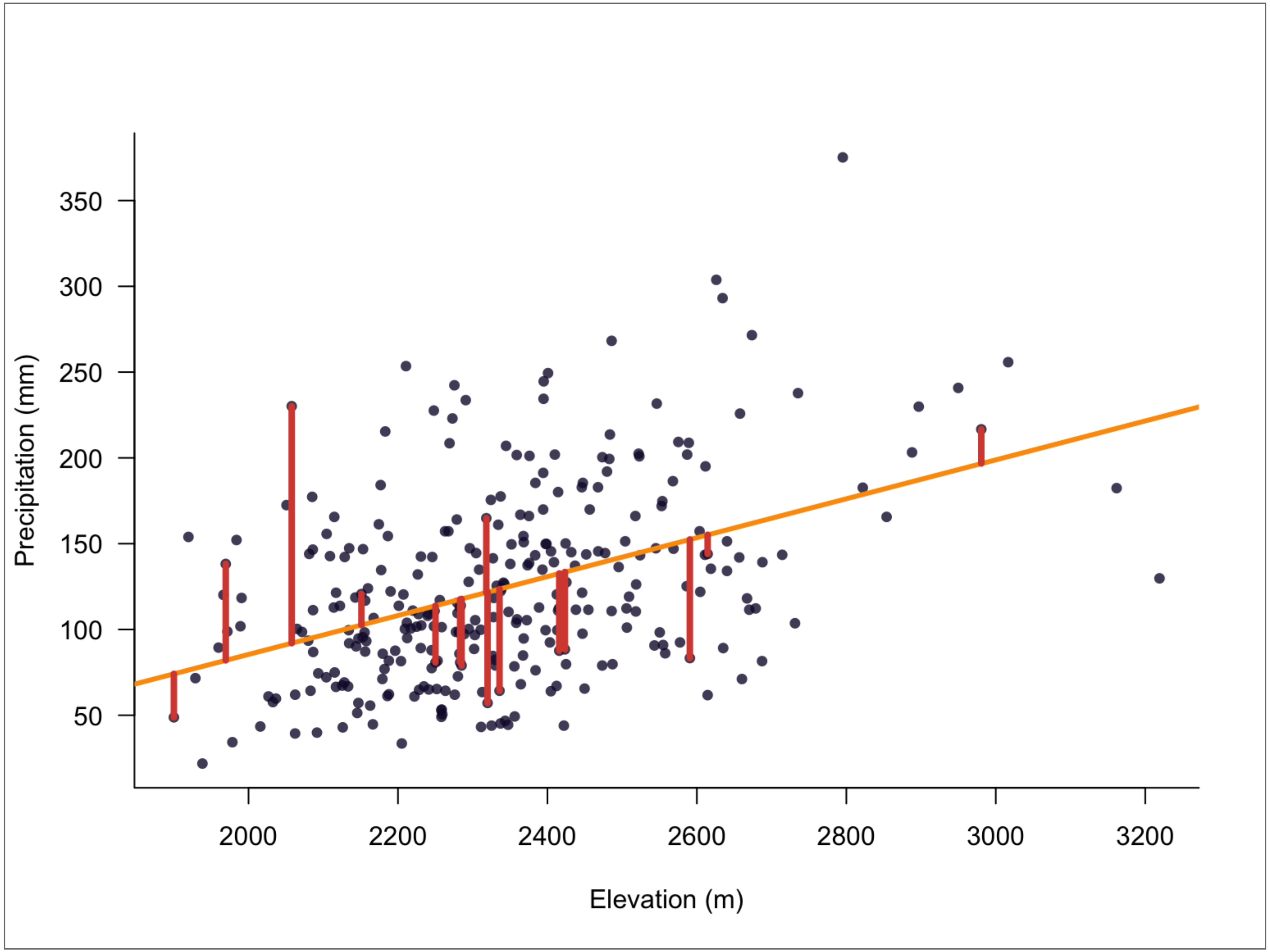
- need an equation for a line:
  - response = intercept + slope × predictor + residual
  - $y_i = \alpha + \beta x_i + \epsilon_i$

- our goal is to estimate the best values of α and β

LINEAR REGRESSION



LINEAR REGRESSION



LINEAR REGRESSION: ASSUMPTIONS

- observations are independent
- residuals are normally distributed
- residual variance is constant

LINEAR REGRESSION: INDEPENDENT OBSERVATIONS

- each independent observations gives a certain amount of information
- non-independent observations give less information
- how to avoid issues: good study design
- how to address issues: mixed models (second session)

LINEAR REGRESSION: RESIDUAL DISTRIBUTION

- $y_i = \alpha + \beta x_i + \epsilon_i$
- we assume ϵ_i is normally distributed
  - needed to define *likelihood*
- how to identify issues: diagnostic checks
- how to address issues: transformations, generalised linear models

LINEAR REGRESSION: CONSTANT RESIDUAL VARIANCE

- $y_i = \alpha + \beta x_i + \epsilon_i$
- we assume ϵ_i all come from one distribution
- how to identify issues: diagnostic checks
- how to address issues: transformations, GLMs, hierarchical models

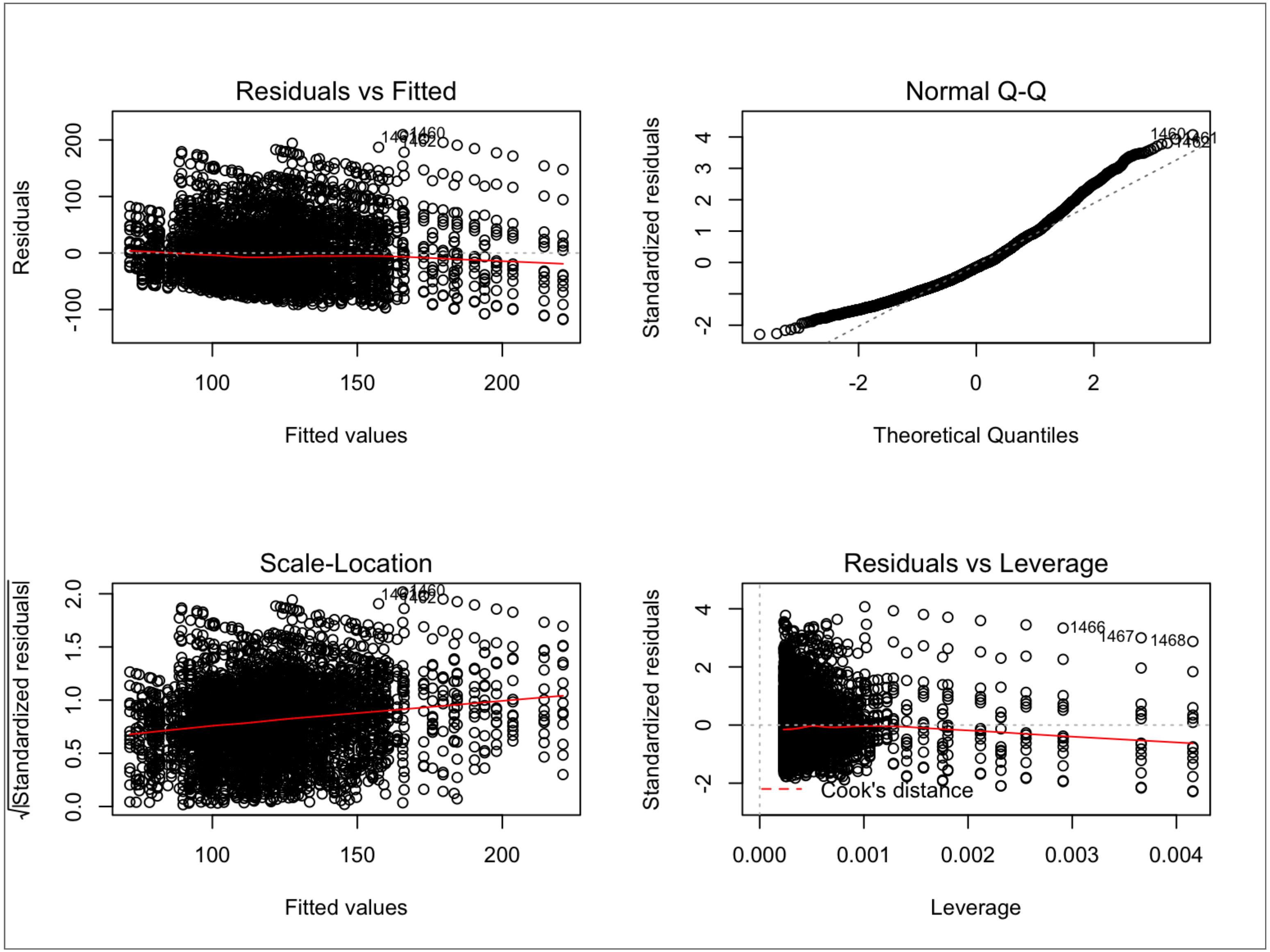
LINEAR REGRESSION: FITTING A MODEL IN R

```r
# fit model
mod <- lm(precipitation ~ elevation)

# check model assumptions
plot(mod)

# summarise model
summary(mod)
```

LINEAR REGRESSION: ASSESSING A FITTED MODEL

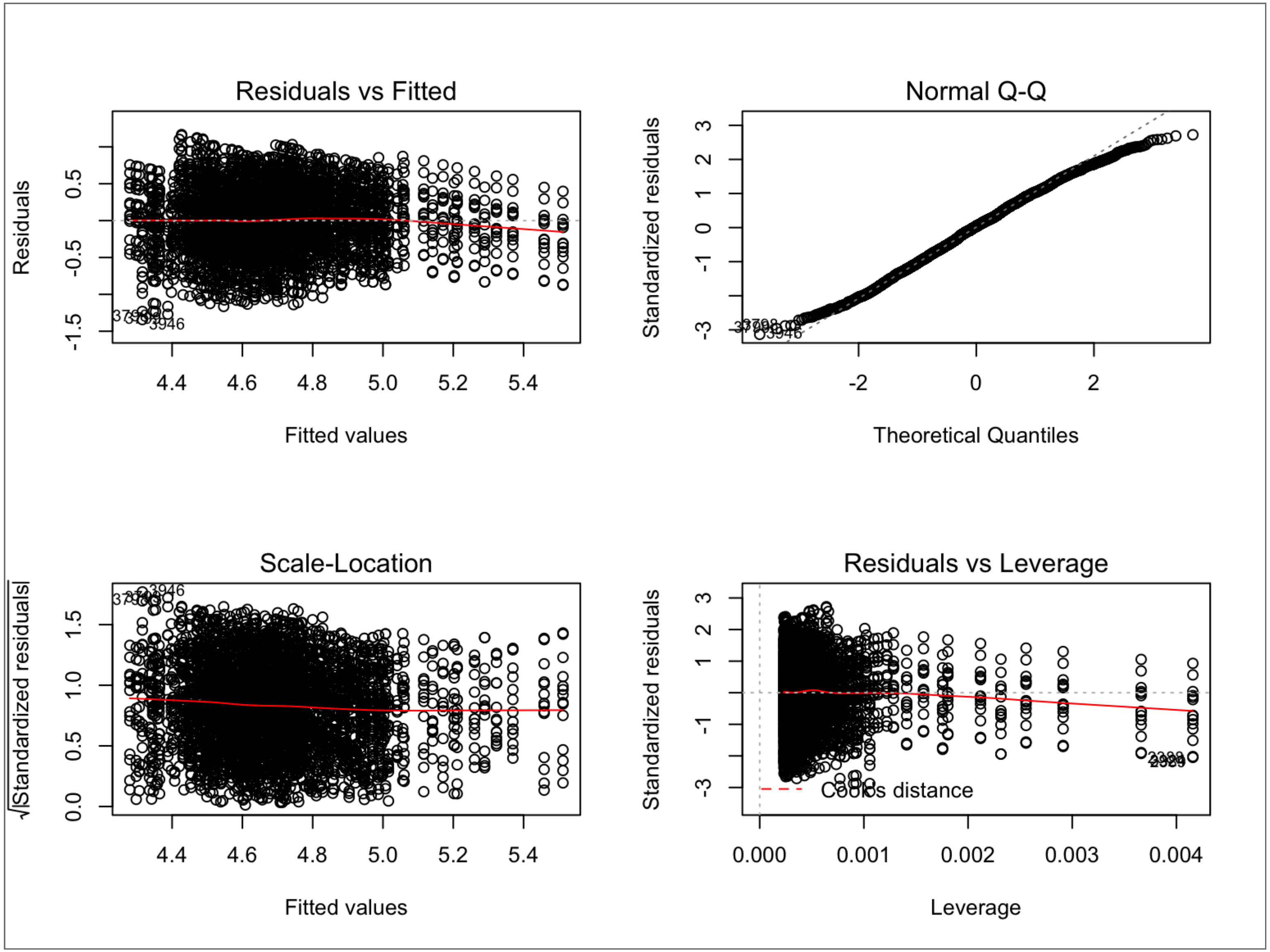


LINEAR REGRESSION: ASSESSING A FITTED MODEL

```r
# fit model with log-transformed response
mod <- lm(log(precipitation) ~ elevation)

# is it any better?
plot(mod)
```

LINEAR REGRESSION: ASSESSING A FITTED MODEL



LINEAR REGRESSION: INTERPRETING A FITTED MODEL

```r
summary(mod)
```

```
##
## Call:
## lm(formula = precipitation ~ elevation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -117.788  -37.852   -8.118   30.345  209.488
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.437e+02  8.632e+00  -16.64   <2e-16 ***
## elevation    1.133e-01  3.672e-03   30.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.48 on 4296 degrees of freedom
## Multiple R-squared:  0.1813, Adjusted R-squared:  0.1811
## F-statistic: 951.1 on 1 and 4296 DF,  p-value: < 2.2e-16
```

LINEAR REGRESSION: INTERPRETING A FITTED MODEL

- how well does the model fit?
  - typically use r^2

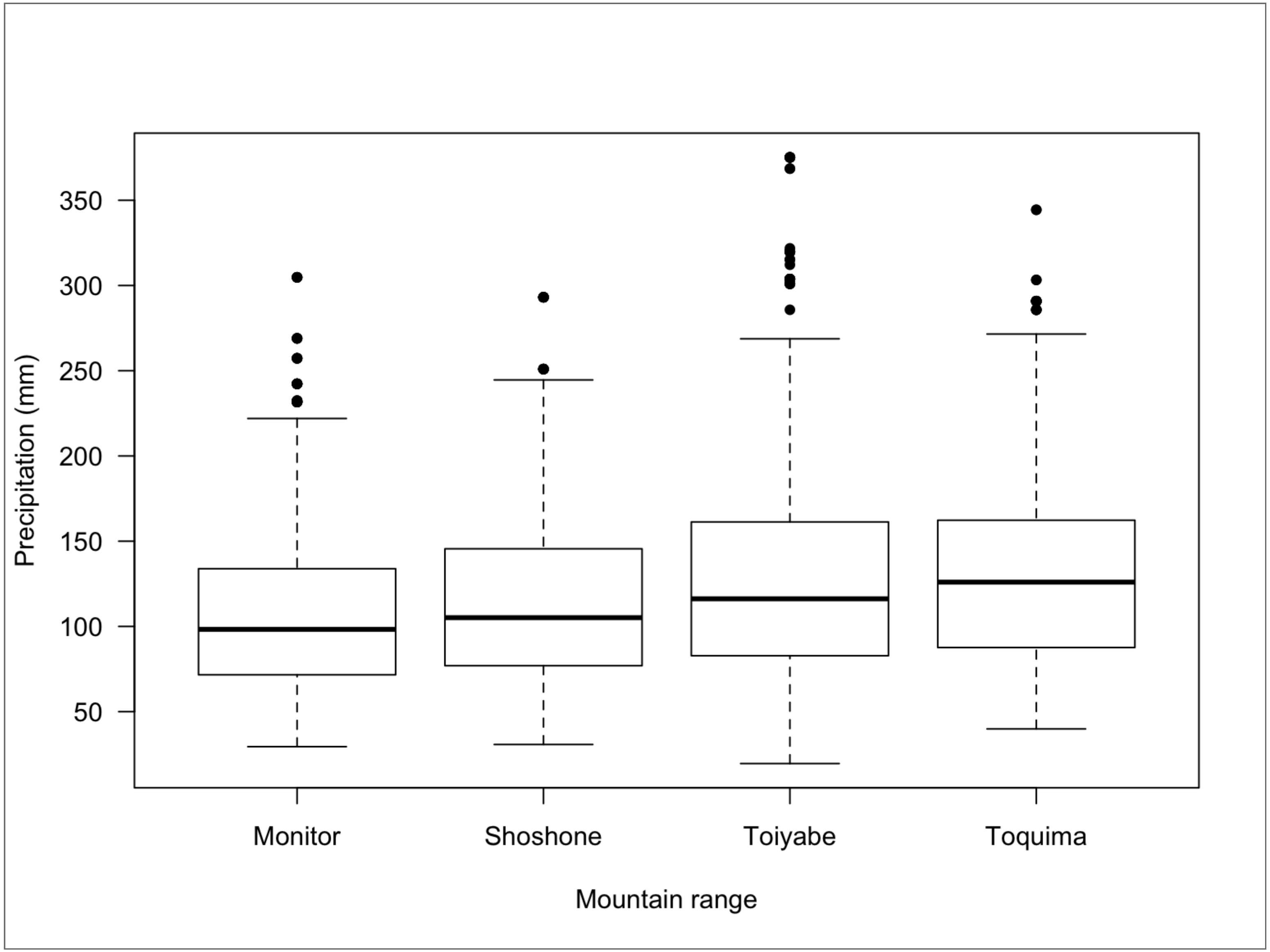
- is there statistical support for an association?
  - often use *p-values*

- is a statistically supported association meaningful?
  - look at the coefficients

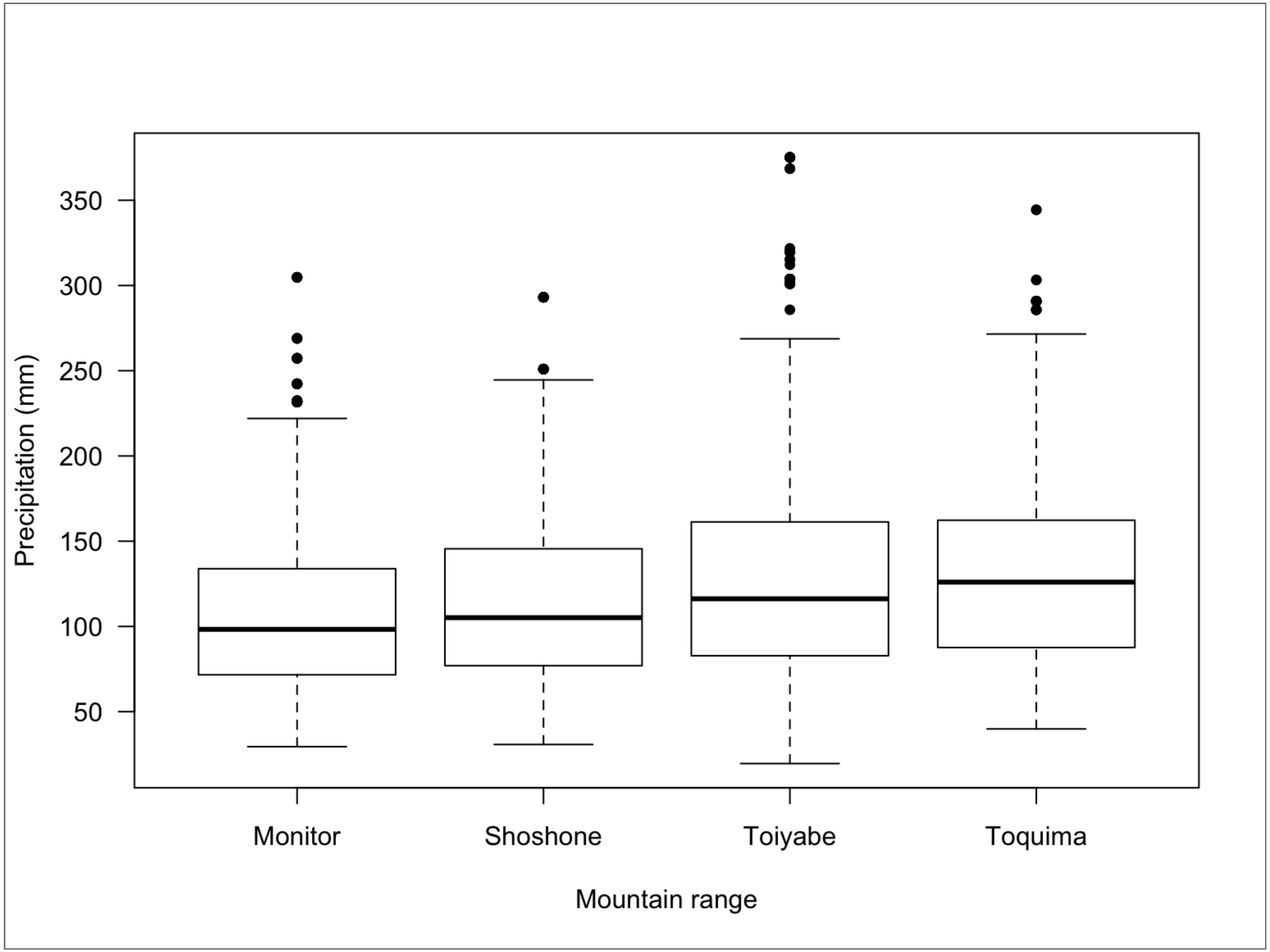
LINEAR REGRESSION: PRESENTING A FITTED MODEL

- is the model adequate? (assumptions, diagnostics)
- does the model fit the data? (diagnostics, r^2)
- is the model statistically meaningful? (p-values, test statistics)
- is the model actually meaningful? (parameter estimates)
- can I see it? (scatterplots)

ANOTHER EXAMPLE

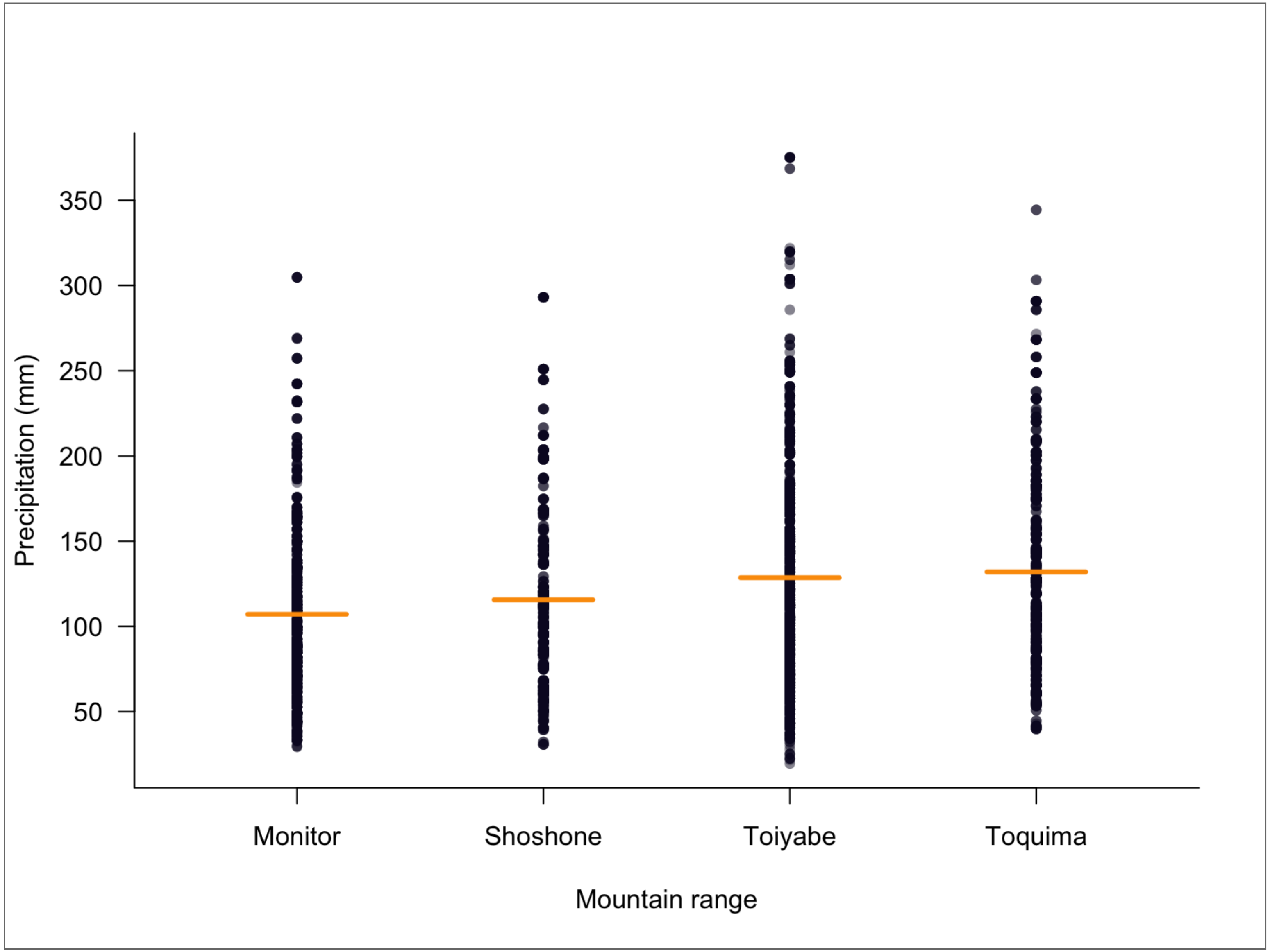


ANOVA



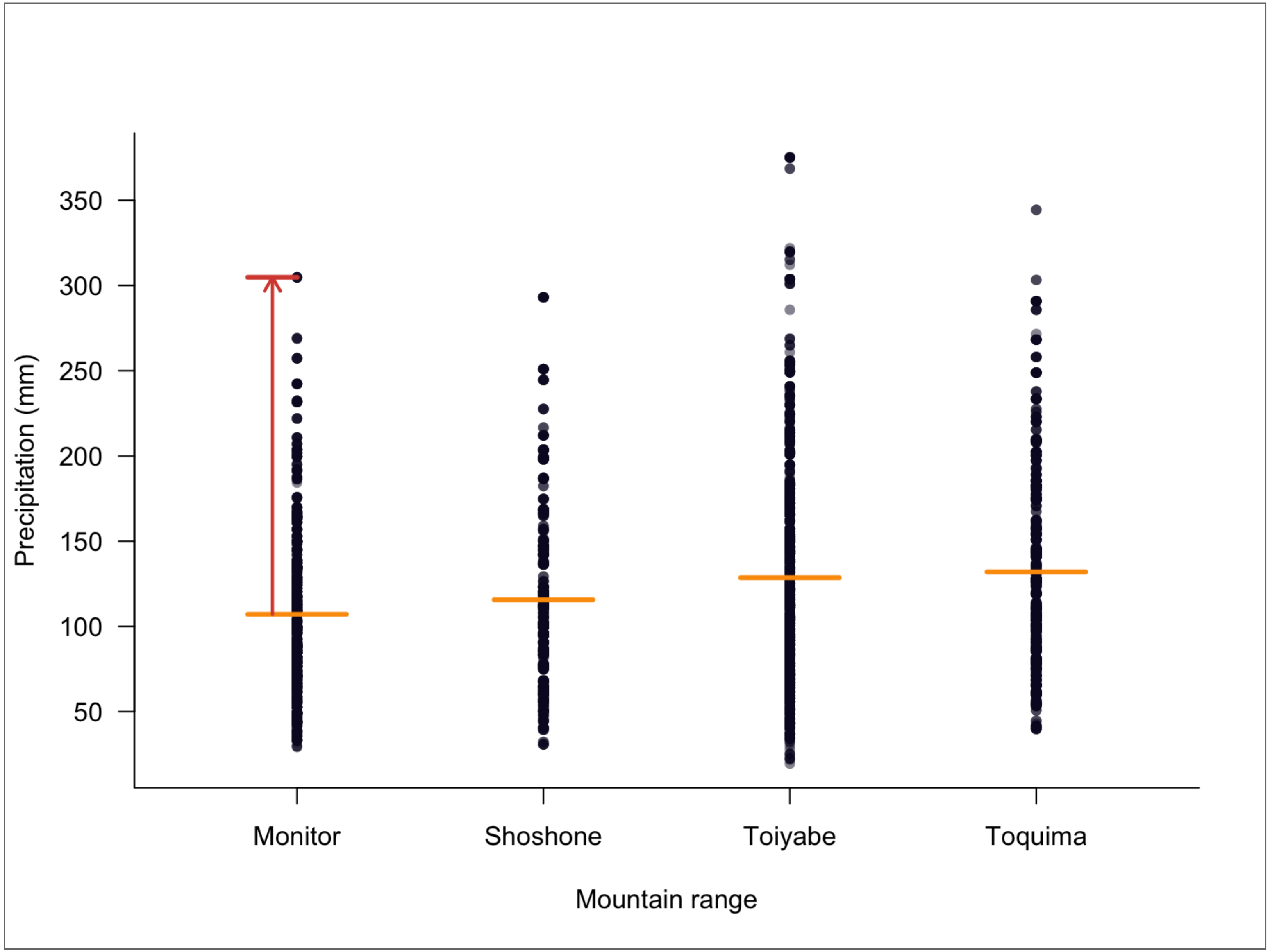
Precipitation (mm)
350
300
250
200
150
100
50
Monitor
Shoshone
Toiyabe
Toquima
Mountain range

ANOVA



Precipitation (mm)
350
300
250
200
150
100
50
Monitor
Shoshone
Toiyabe
Toquima
Mountain range

ANOVA



Precipitation (mm)
350
300
250
200
150
100
50
Monitor
Shoshone
Toiyabe
Toquima
Mountain range

ANOVA

- what characterises this example?
  - continuous response
  - *discrete* predictor

- assumptions: identical to linear regression
- response = overall intercept + group intercept + residual
- $y_i = \alpha + \beta_{g(i)} + \epsilon_i$

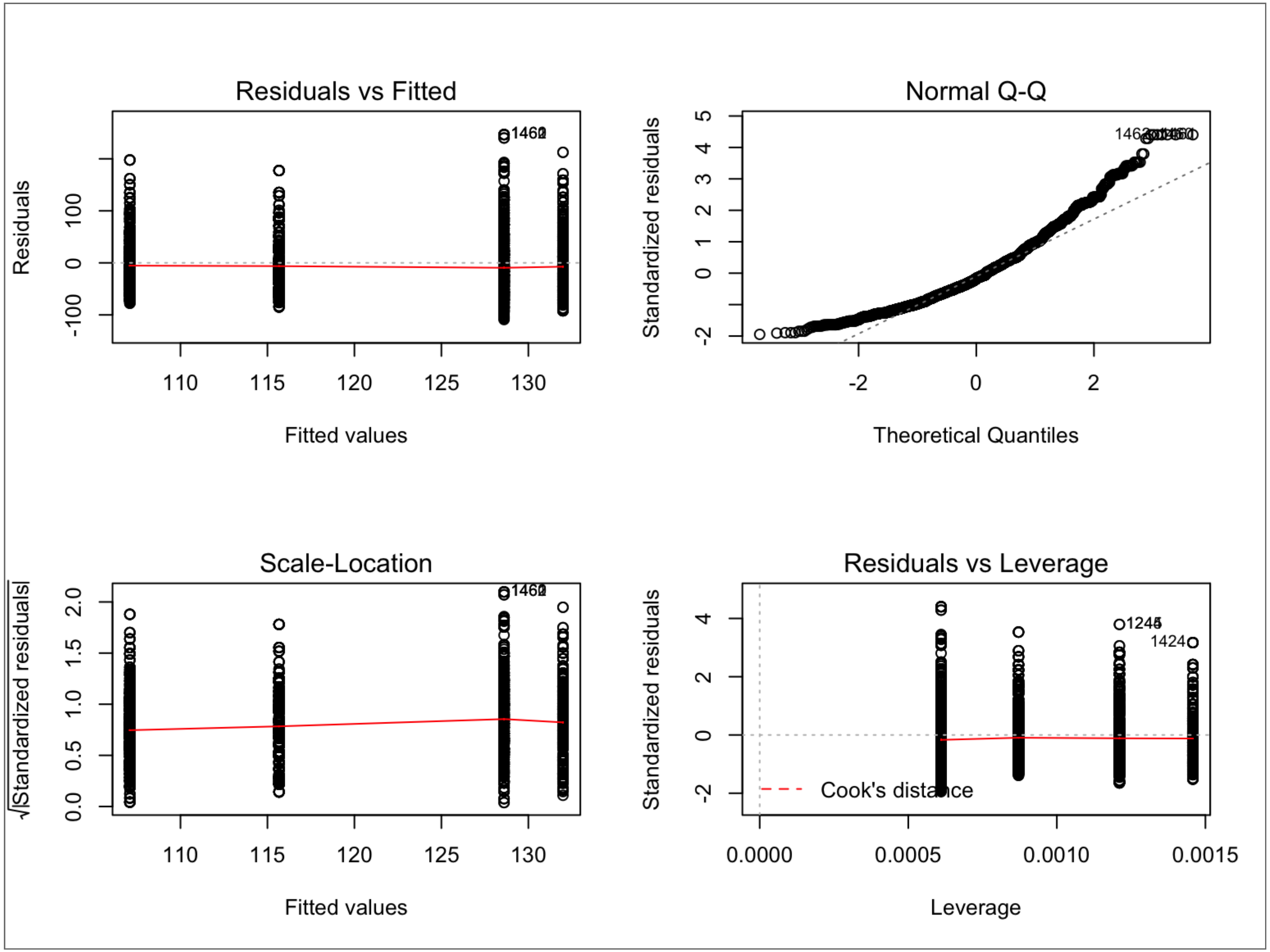
ANOVA: FITTING A MODEL IN R

```
# fit a model
mod <- lm(response ~ predictor, data = data_set)

# does the model meet assumptions?
plot(mod)

# summarise the model
summary(mod)
```

ANOVA: ASSESSING A FITTED MODEL



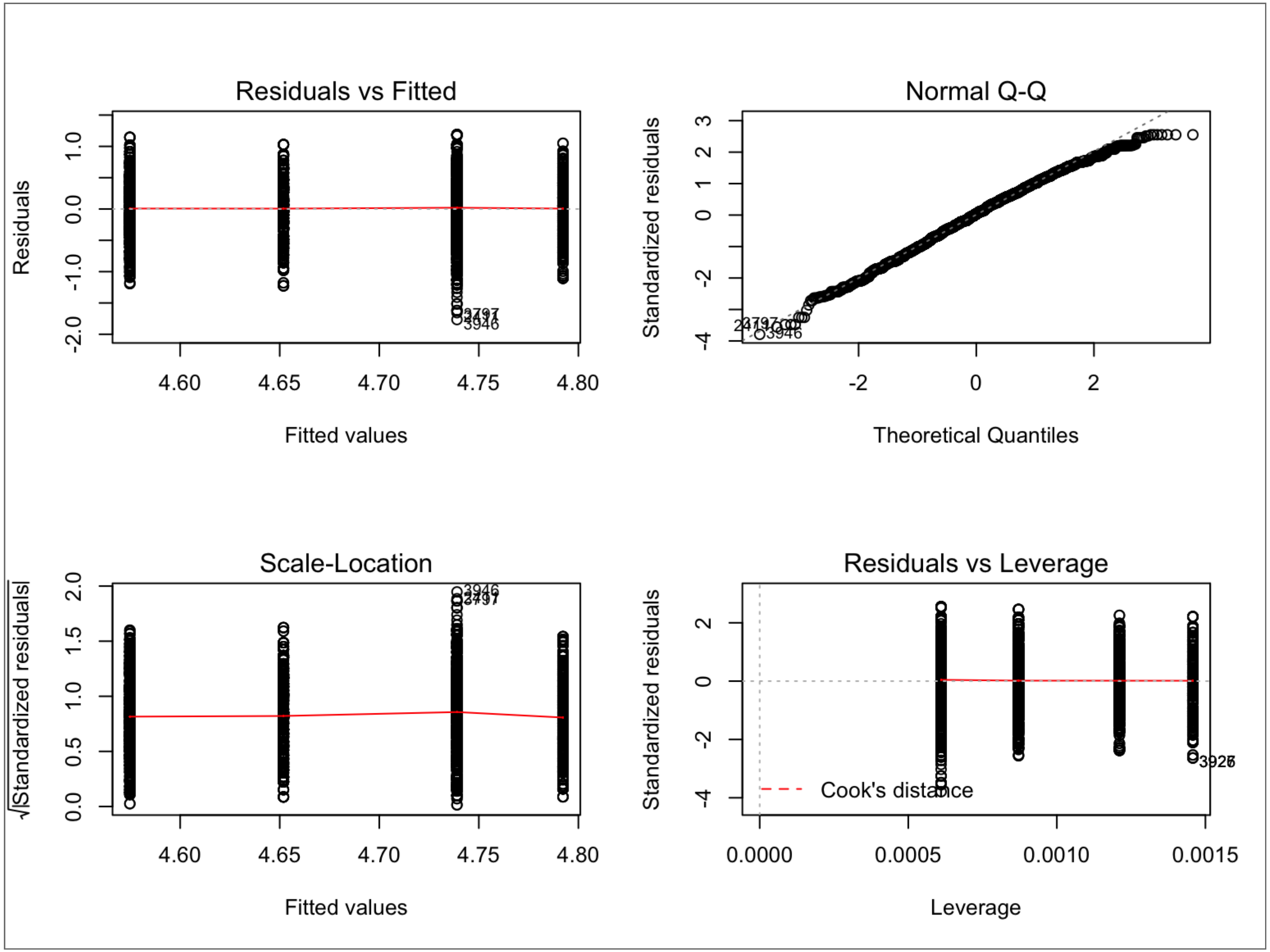
ANOVA: FITTING A MODEL IN R

```r
# fit a model to log-transformed data
mod <- lm(log(response) ~ predictor, data = data_set)

# does the model meet assumptions?
plot(mod)

# summarise the model
summary(mod)
```

ANOVA: ASSESSING A FITTED MODEL



ANOVA: INTERPRETING A FITTED MODEL

```
##
## Call:
## lm(formula = log(precipitation) ~ mountain_range)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76554 -0.31034  0.01769  0.32208  1.18833
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.57471    0.01375 332.777  < 2e-16 ***
## mountain_rangeShoshone  0.07717    0.02248   3.433 0.000602 ***
## mountain_rangeToiyabe   0.16431    0.01793   9.165  < 2e-16 ***
## mountain_rangeToquima   0.21755    0.02125  10.237  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4658 on 4294 degrees of freedom
## Multiple R-squared:  0.03001,    Adjusted R-squared:  0.02933
## F-statistic: 44.28 on 3 and 4294 DF,  p-value: < 2.2e-16
```

ANOVA: INTERPRETING A FITTED MODEL

- now we have lots of p-values. . .
- can use *post hoc* tests but not universally accepted
- can pre-specify *contrasts* for specific hypotheses

ANOVA: PRESENTING A FITTED MODEL

- is the model adequate? (assumptions, diagnostics)
- does the model fit the data? (diagnostics, r^2)
- is the model statistically meaningful? (p-values, test statistics)
- is the model actually meaningful? (parameter estimates)
- can I see it? (boxplots)

ASIDE: DISCRETE PREDICTOR WITH TWO LEVELS

- special case: t-test (it's still an ANOVA)

```
mod <- t.test(response ~ predictor, data = data_set)
summary(mod)
```

```
##
##  Welch Two Sample t-test
##
## data:  precipitation by region
## t = -4.2244, df = 4284.3, p-value = 2.446e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.652670  -3.899172
## sample estimates:
## mean in group East mean in group West
##           117.5083           124.7842
```

GENERAL LINEAR MODELS

- linear regression, ANOVA, t-test: they're all the same
- just needs a special setup for discrete predictors

MATRIX NOTATION

- $y_i = \alpha + \beta^T x_i + \epsilon_i$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}; \; x_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,k} \end{pmatrix}$$

$$\beta^T = \begin{pmatrix} \beta_1 & \dots & \beta_k \end{pmatrix}$$

ANOVA: HOW DOES THIS WORK?

- code the $x_{i,k}$ values as 1 or 0

$$\beta^T = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \end{pmatrix}$$

$$x_i = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\beta^T x_i = 0 + \beta_2 + 0$$

ALL TOO MUCH?

- R will do this for you! (this is one reason R has `factors`)

```r
model.matrix( ~ discrete_predictor)
```

```
##      (Intercept) mountain_rangeShoshone mountain_rangeToiyabe
## [1,]           1                      0                     0
## [2,]           1                      1                     0
## [3,]           1                      0                     0
## [4,]           1                      0                     0
## [5,]           1                      0                     1
##      mountain_rangeToquima
## [1,]                     0
## [2,]                     0
## [3,]                     0
## [4,]                     1
## [5,]                     0
```

MORE THAN ONE PREDICTOR

- same setup, but now the x_i values don't have to be 0 or 1

$$\beta^T = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \end{pmatrix}$$

$$x_i = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ x_{i,3} \end{pmatrix}$$

$$\beta^T x_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$$

MORE THAN ONE PREDICTOR

```
model.matrix( ~ predictor1 + predictor2 + predictor3)
```

```
##      (Intercept) predictor1 predictor2 predictor3
## [1,]           1     2285.1      78.99       20.8
## [2,]           1     2304.6      67.31       17.1
## [3,]           1     2330.1      64.27       16.7
## [4,]           1     2589.2     144.02       15.8
## [5,]           1     3016.6     235.46        6.2
```

MORE THAN ONE PREDICTOR

- the scale of the variables matters

- good to standardise continuous predictors

```r
# standardise continuous predictors
predictors_std <- scale(predictors)
```

```
##      predictor1 predictor2  predictor3
## [1,] -0.7065051 -0.5328227  1.00678496
## [2,] -0.6438887 -0.6923145  0.32702139
## [3,] -0.5620058 -0.7338261  0.25353344
## [4,]  0.2699889  0.3551696  0.08818554
## [5,]  1.6424108  1.6037937 -1.67552534
```

CONTINUOUS AND DISCRETE PREDICTORS

- can include continuous and discrete predictors in one model

```
mod <- lm(response ~ continuous1 + continuous2 + discrete)
```

```
##      (Intercept) continuous1 continuous2 discrete1 discrete2 discrete3
## [1,]           1      2285.1       78.99         0         0         0
## [2,]           1      2304.6       67.31         1         0         0
## [3,]           1      2330.1       64.27         0         0         0
## [4,]           1      2589.2      144.02         0         0         1
## [5,]           1      3016.6      235.46         0         1         0
```

MULTIPLE PREDICTORS: NEW ASSUMPTIONS

- all the same assumptions as before
- plus: predictors are assumed to be independent(ish)
  - technical term: *multicollinearity*
- if two predictors are highly correlated the model can't tell them apart
- how to address issues: careful predictor choice, remove correlated predictors

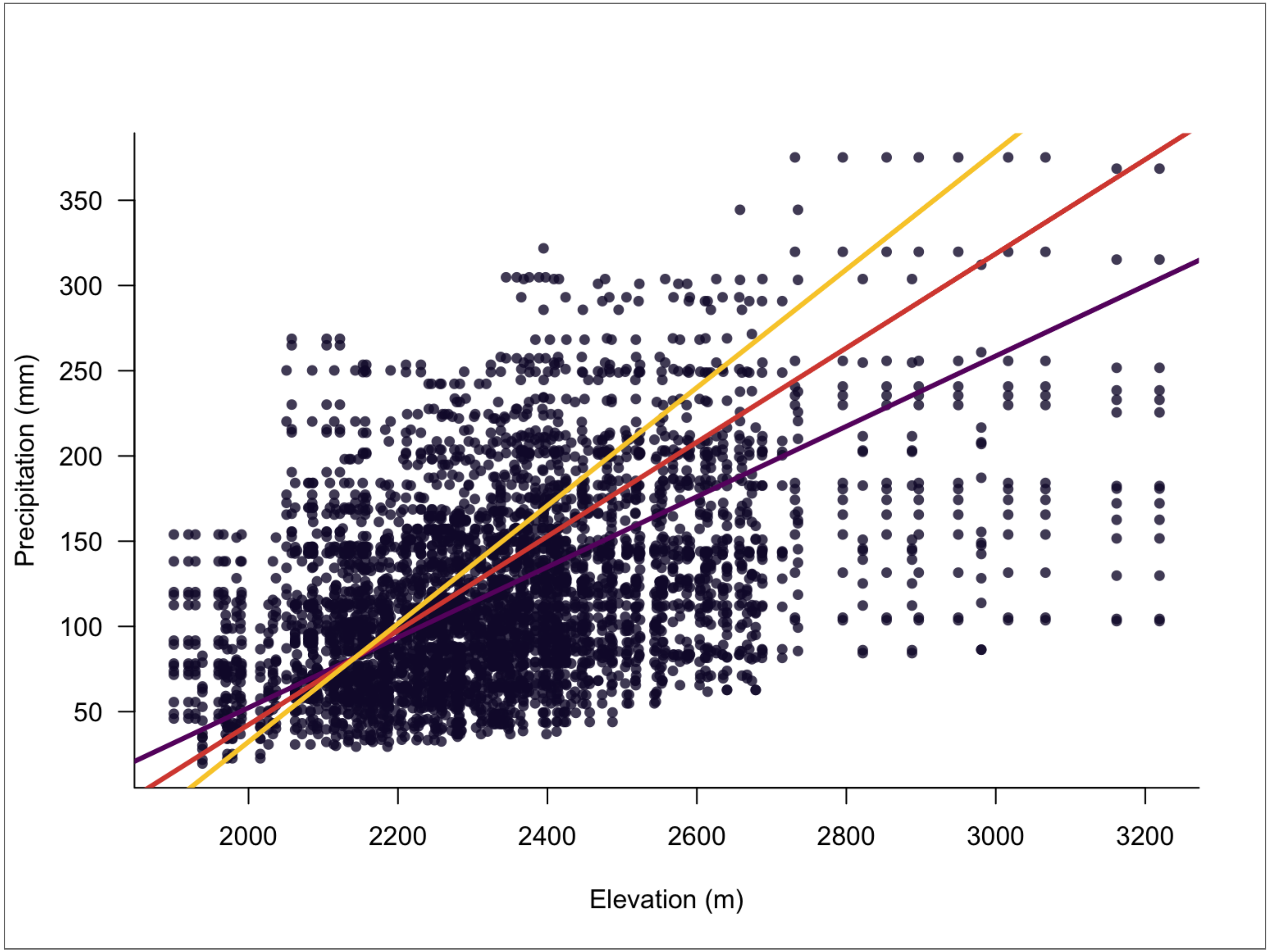
MULTICOLLINEARITY IN R

```
round(cor(predictor_variables), 2)
```

```
##            predictor1 predictor2 predictor3 predictor4
## predictor1       1.00       0.34       0.43      -0.53
## predictor2       0.34       1.00       0.20      -0.22
## predictor3       0.43       0.20       1.00      -0.52
## predictor4      -0.53      -0.22      -0.52       1.00
```

- solution: remove variables until none are highly correlated
  - removing `predictor4` is a good option here

MULTIPLE PREDICTORS: INTERACTIONS



MULTIPLE PREDICTORS: INTERACTIONS

```r
mod <- lm(precipitation ~ elevation * mountain_range)
summary(mod)
```

```
##
## Call:
## lm(formula = precipitation ~ elevation * mountain_range)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.75  -35.97   -9.48   30.03  202.73
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -2.104e+02  3.185e+01  -6.607  4.4e-11
## elevation                        1.364e-01  1.366e-02   9.981  < 2e-16
## mountain_rangeShoshone           1.842e+01  4.067e+01   0.453  0.65053
## mountain_rangeToiyabe            9.473e+01  3.346e+01   2.832  0.00465
## mountain_rangeToquima            1.798e+01  4.192e+01   0.429  0.66795
## elevation:mountain_rangeShoshone -3.521e-03  1.747e-02  -0.202  0.84028
## elevation:mountain_rangeToiyabe  -3.089e-02  1.435e-02  -2.152  0.03145
## elevation:mountain_rangeToquima  -2.771e-03  1.767e-02  -0.157  0.87538
##
## (Intercept)                      ***
## elevation                        ***
## mountain_rangeShoshone
## mountain_rangeToiyabe            **
## mountain_rangeToquima
## elevation:mountain_rangeShoshone
```

MULTIPLE PREDICTORS: INTERACTIONS

- difficult to interpret coefficients
  - effect of one depends on value of the other
  - particularly hard if both are continuous

- it is possible to include higher-order interactions
  - even more difficult to interpret