6. Linear regression

Introduction to Medical Statistics

OUCRU, Ho Chi Minh City

March 23-27, 2026

Nguyen Lam Vuong
and the biostatistics crew

Learning Objectives

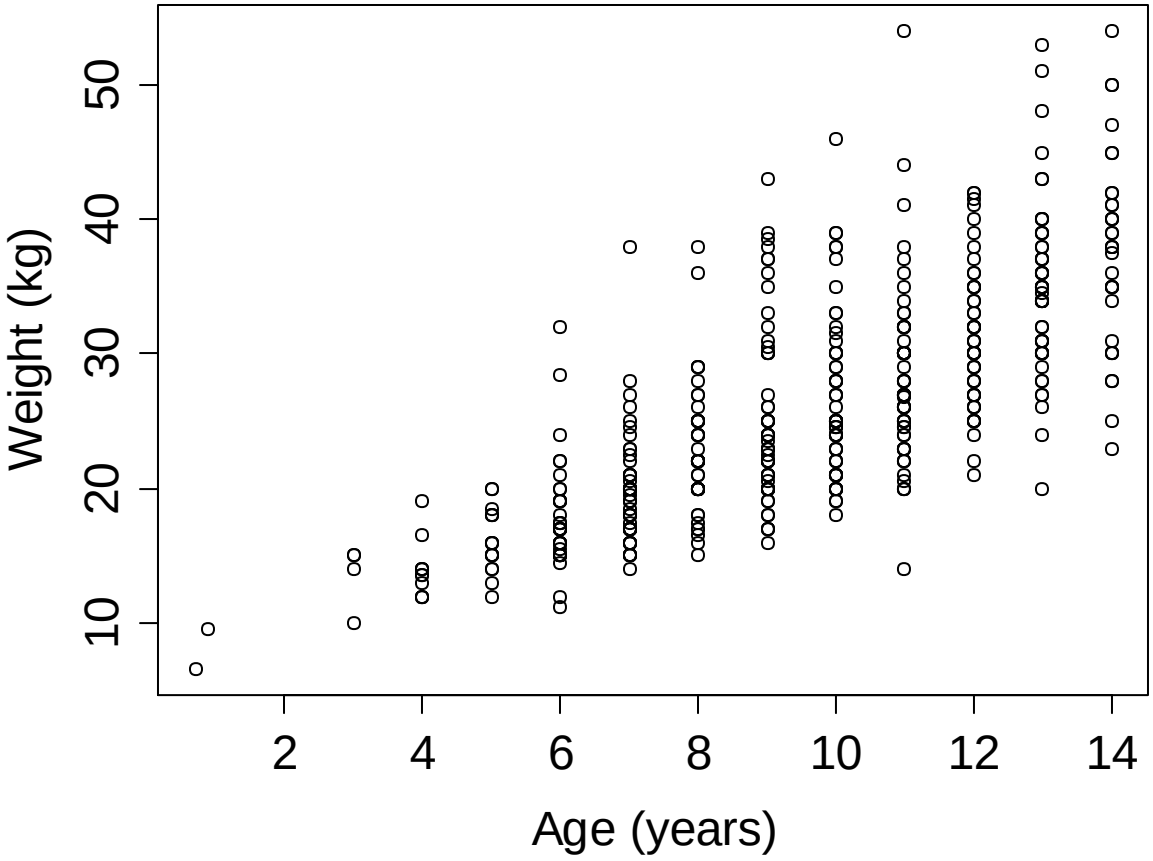
- Know the graphical display for the association between 2 continuous variables
- Understand linear regression model
- Interpret the results from linear regression model
- Understand linear regression diagnostics

Summarizing the association between two continuous variables

Graphical display: scatterplot

- First step: plot the data – scatterplot
  - x – independent variable (predictor, covariable)
  - y – dependent variable (outcome, response)
  - Sometimes not clear which variable is x or y - then any is good for the plot (not for the analysis)
  - Each observation is represented by one point
  - Pattern of the points gives an idea whether there is a relationship between variables and clarifies the type of relationship

Example: Scatter plot of age vs. weight for dengue shock dataset



- Roughly linear increase of weight with age
- A lot of variability, especially for higher age

Model formulation, assumptions and estimation

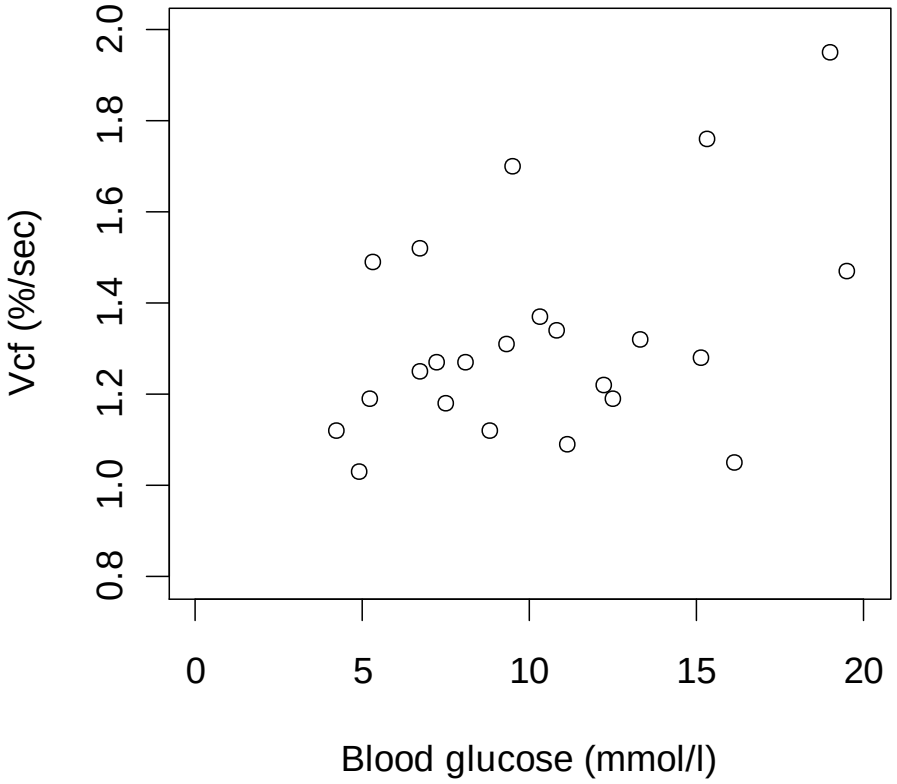
Correlation and regression

- (Pearson) correlation
  - Strength of a linear association between two numeric (say x and y) variables on a scale between -1 and +1
  - Variables are treated symmetrically
  - Descriptive
- T-test:
  - Mean value of (numeric) outcome y differs between groups?
- Linear regression
  - Quantifies how (numeric) y depends on covariable x (categorical/ numeric): Does x have effect on y? Can we predict y if we know x?
  - Variables not treated symmetrically
    - x: covariate, covariable, predictor or independent variable
    - y: response, outcome or dependent variable
  - Multiple/multivariable linear regression: 2 or more covariables

Example: Glucose

- Measurements in 23 type I diabetic patients
  - Fasting blood glucose (x)
  - Velocity of circumferential shortening of the left ventricle (Y)

  → Is Y related to x?



Altman, p.299f

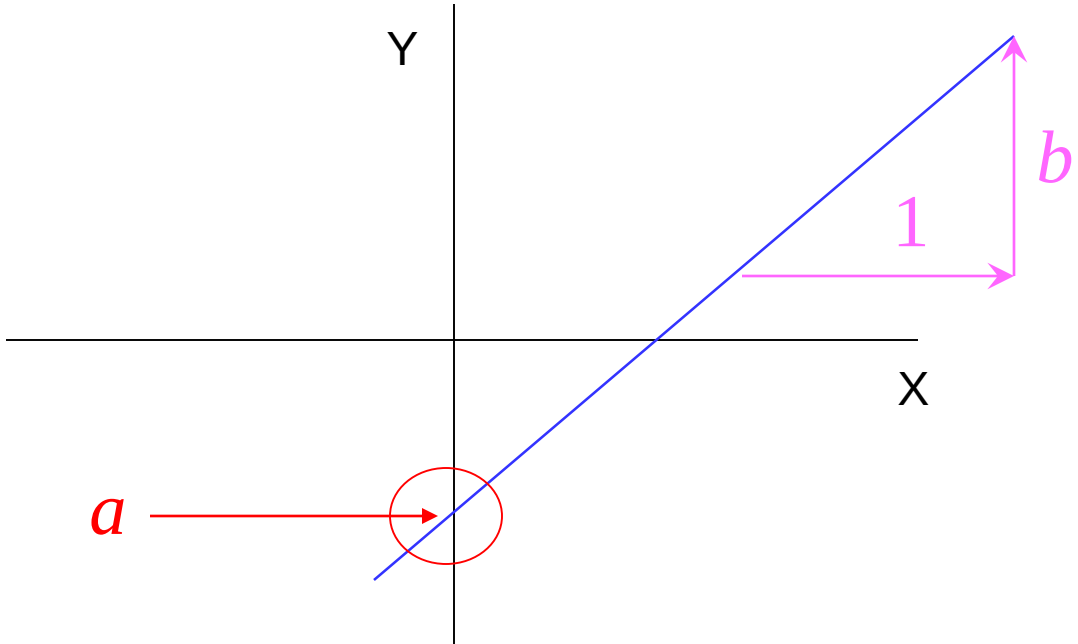
Linear regression model

- Assumes that the average value ("expected value") of Y depends on x via

$$E(Y)=a+b\cdot x$$

a and **b** are called *parameters*

- a: Intercept of the line, i.e. expected value of Y if x=0
- b: Slope of the line, i.e. expected change in Y if x increases by 1



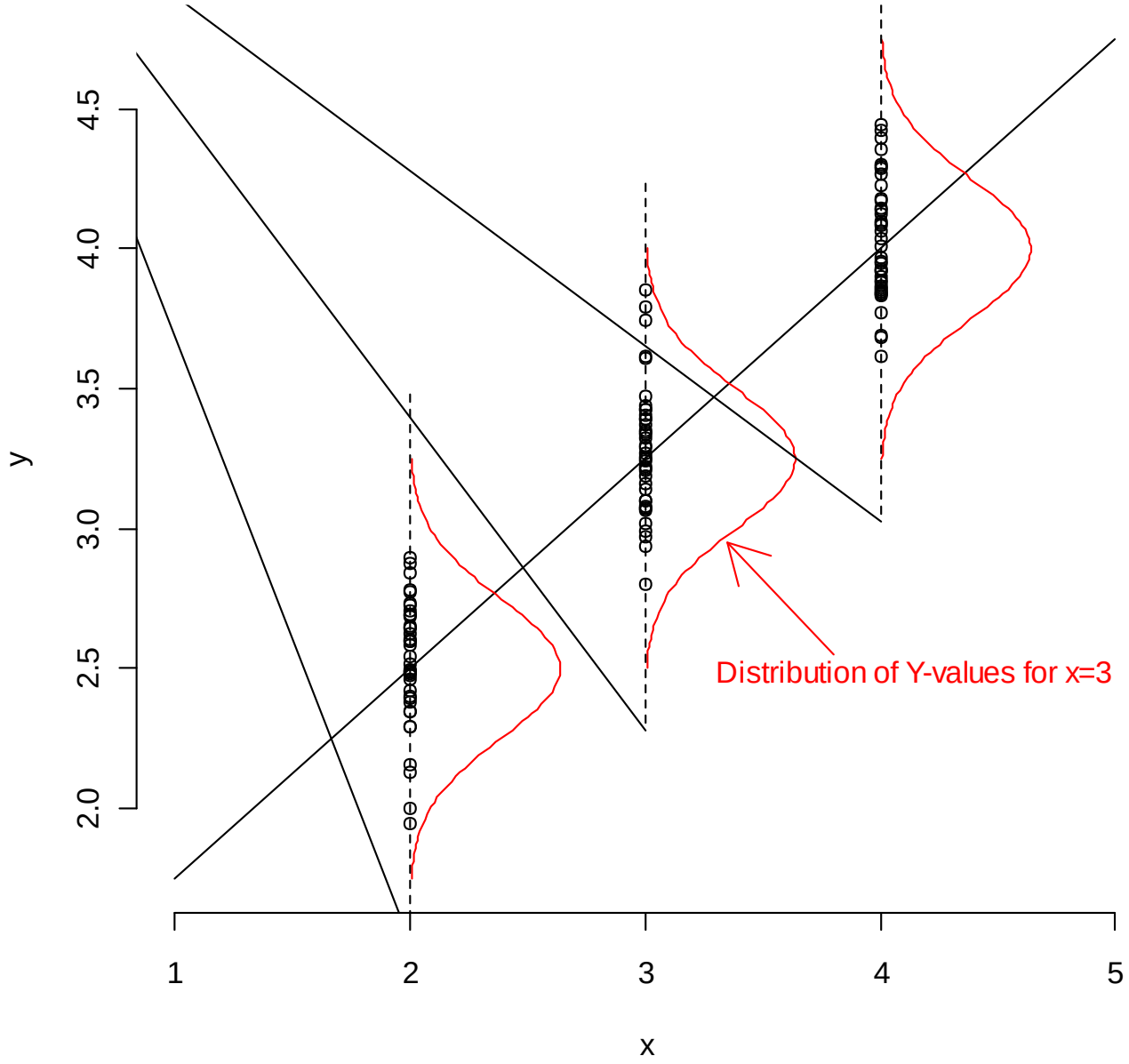
Linear regression data model

- Mathematical formulation:

$$Y_i = a + b \cdot x_i + \varepsilon_i$$

  - x_i and Y_i: Independent and dependent variable for observation I "Systematic part"
  - $\varepsilon_i \sim N(0,\sigma^2)$: "Unexplained variation" - independent of each other
- Assumptions (all can be relaxed, not covered in course)
  - Population mean of Y depends on x via $a+b \cdot x$
  - Deviations of Y-observations from the assumed line
    - are normally distributed
    - have constant standard deviation σ (i.e. the same for all x)
    - are independent of each other

The linear regression model (illustration)



Least squares estimation of the regression line

- Find the line that minimizes the sum of the squared vertical distances of the observations from the line. → Least squares estimation of the regression line.
- For an observation i define:
  - Fitted value: Predicted y-value on the regression line based on x_i:

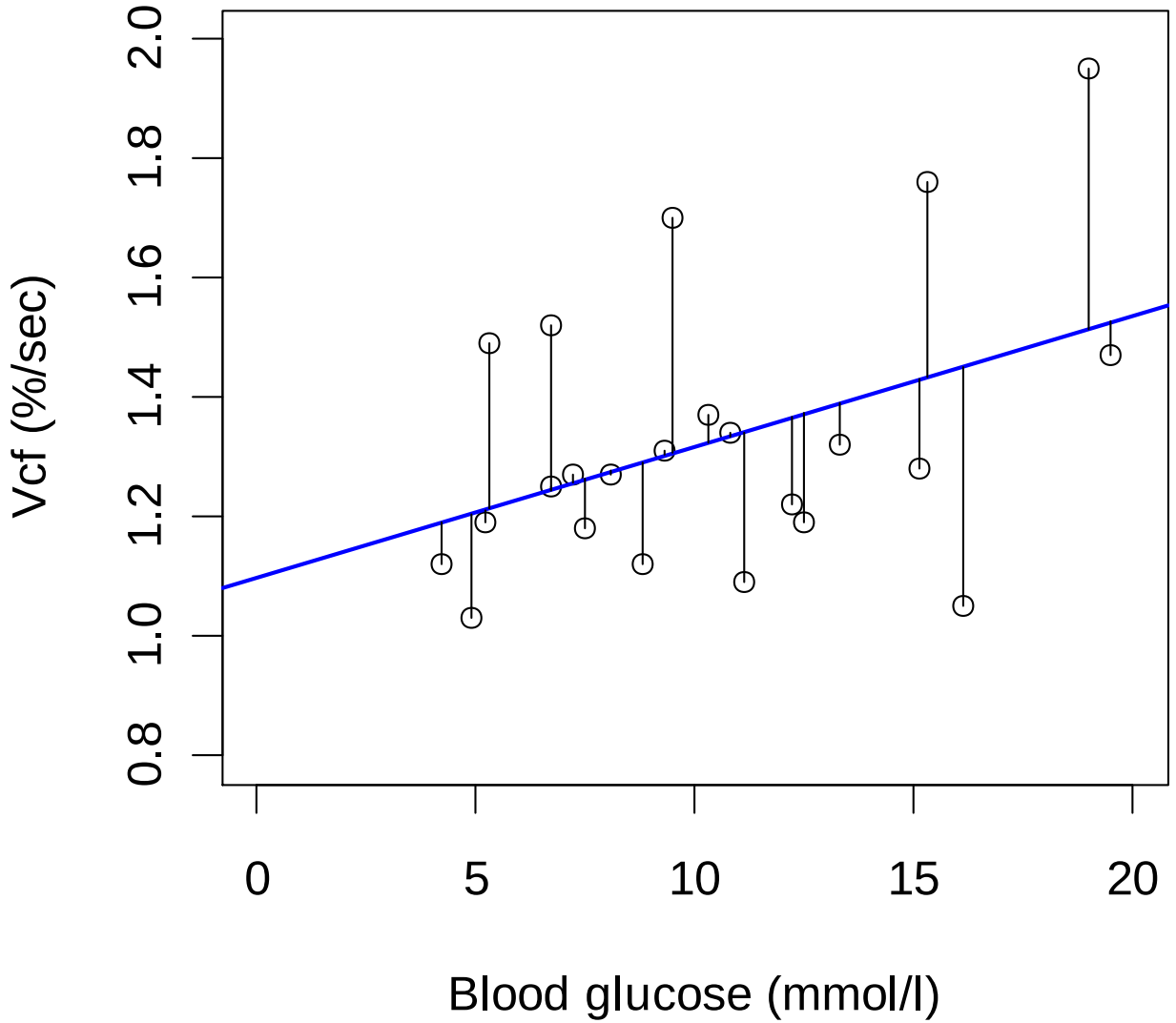
$$\text{Fitted value: } \hat{y}_i = \hat{a} + \hat{b}x_i$$

  - Residual: Vertical distance between the observed y-value and the value on the regression line:

$$\text{Residual} = y_i - (\hat{a} + \hat{b}x_i)$$

→ The linear regression line minimizes the sum of the squared residuals.

Least squares line and residuals for the glucose example



Regression line, fitted values and residuals for the glucose example

- The regression line has the form

  $Vcf = 1.10 + 0.0220 \cdot Glucose$

- For a patient with a glucose of 5 mmol/l, the linear regression predicts an average Vcf of
  $1.10 + 0.0220 \cdot 5 = 1.21$ %/sec

- A patient has a glucose of 5 mmol/l and an observed Vcf of 1.40 %/sec:
  - → Fitted value: 1.21
  - → Residual: 1.40 - 1.21=0.19

Statistical tests and confidence intervals

Goals

- Find methods for the 3 basic questions of statistical inference
  - Which values of the parameters are most plausible in the light of the data? → Estimation
  - Which parameter values are plausible given the data? → Confidence interval
  - Is a certain, predetermined parameter value plausible? → Test

Solutions

- Estimation
  - Least-squares estimation as discussed

- Confidence intervals and tests
  - Determine standard errors (s.e.) of parameter estimates

$$se\left(\hat{b}\right) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

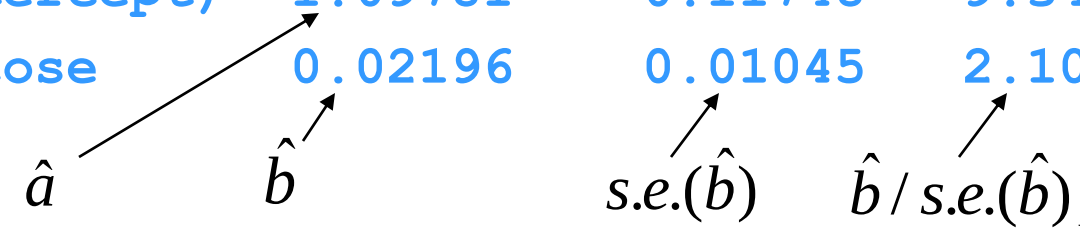
  - Use general rules for testing and confidence intervals
  - Exact methods are based on the t-distribution with n-2 degrees of freedom

→ R will do all this for you

R-output, glucose example (abbreviated)

```
> regr.glucose <- lm(vcf~glucose,data=glucose.vcf)
> summary(regr.glucose)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.09781    0.11748   9.345   <0.001
glucose      0.02196    0.01045   2.101   0.0479
```



p-value of null hypothesis that population slope is 0, i.e. that glucose does not linearly affect vcf

```
Residual standard error: 0.2167
```

$\hat{\sigma}$, i.e. estimated s.d. of random error around regression line.

```
Multiple R-squared: 0.1737
```

Squared Pearson correlation between vcf and glucose.

Conclusions

Regression equation: Exp(Vcf) = 1.10 + 0.022·Glucose

In addition, we get s.e.(b) = 0.010

Statistical test of H0: True slope b = 0, i.e.,
"Glucose does not (linearly) influence Vcf"
p-value = 0.048 → Glucose may influence Vcf

Approximate 95% confidence interval for true slope b:

$b \pm 1.96 \cdot s.e.(b) = (0.0024, 0.0416)$

Better 95% CI for parameters

Via `confint` function in R:

```
> regr.glucose <- lm(vcf~glucose,data=glucose.vcf)
> confint(regr.glucose)
                    2.5 %      97.5 %
(Intercept) 0.8534993816 1.34213037
glucose     0.0002231077 0.04370194
```

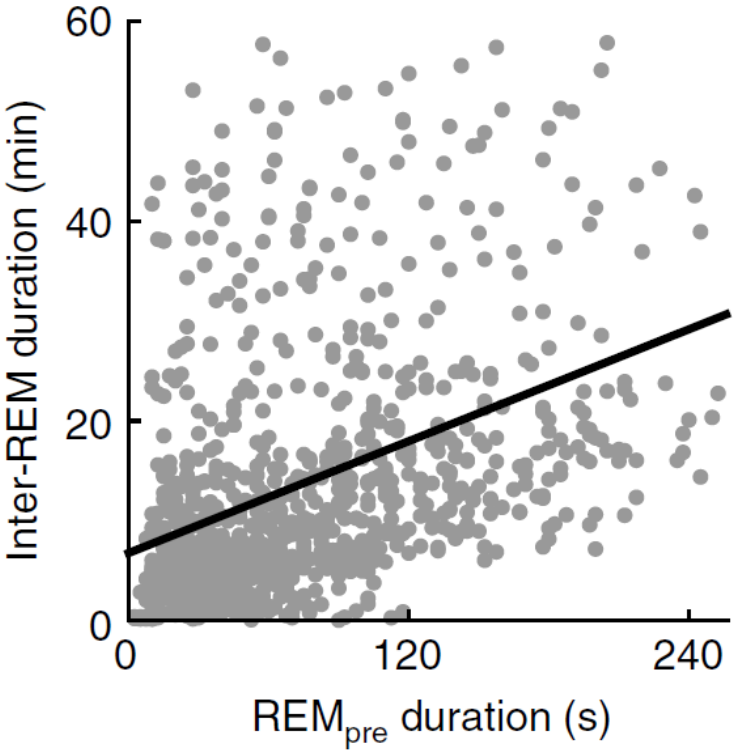
Transformation and scaling of continuous covariables, binary covariables

Transformations?

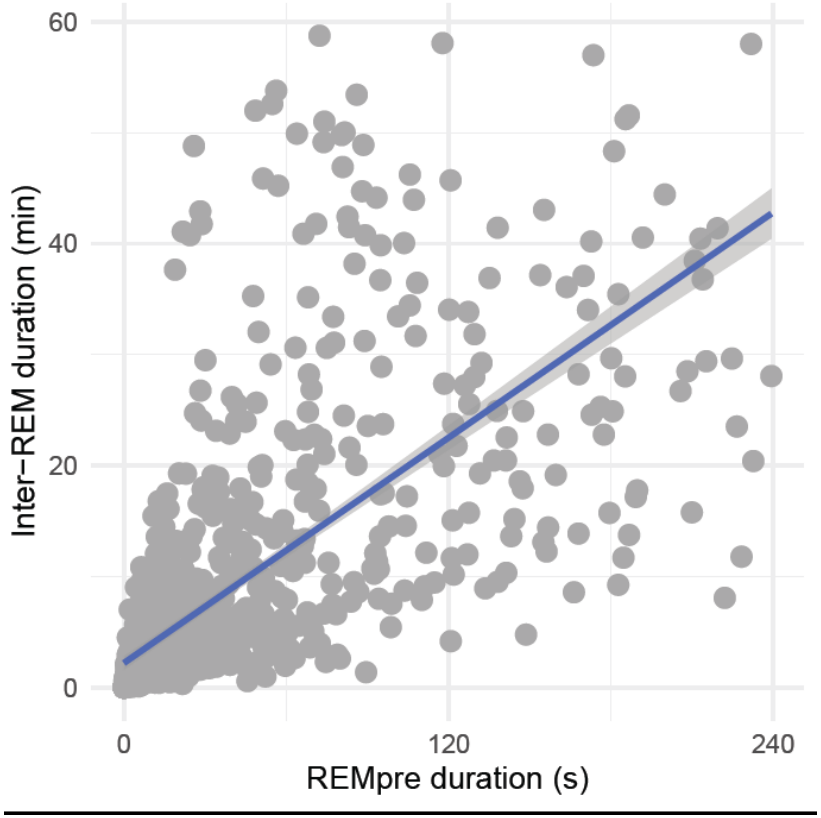
- No assumption on distribution of covariables made
  Doesn't have to be normal, can even be categorical!
- Yet, if variable has highly skewed distribution, relation with the outcome often more linear after transformation.
  - Change in DENV-viral load from 1000 to 100 is expected to influence severe dengue outcome much more than a change from 1,000,000 to 999,900 copies/ml.
    Log-transformation (log10(x) for DENV-viral load)
    → assume that each log10 increase (i.e. 10-fold increase) has same effect, e.g. from 1000 to 100 same effect as from 1,000,000 to 100,000
- Don't categorize numeric variables; gives unrealistic model
- Check distribution of the residuals after the model has been fitted (regression diagnostics)

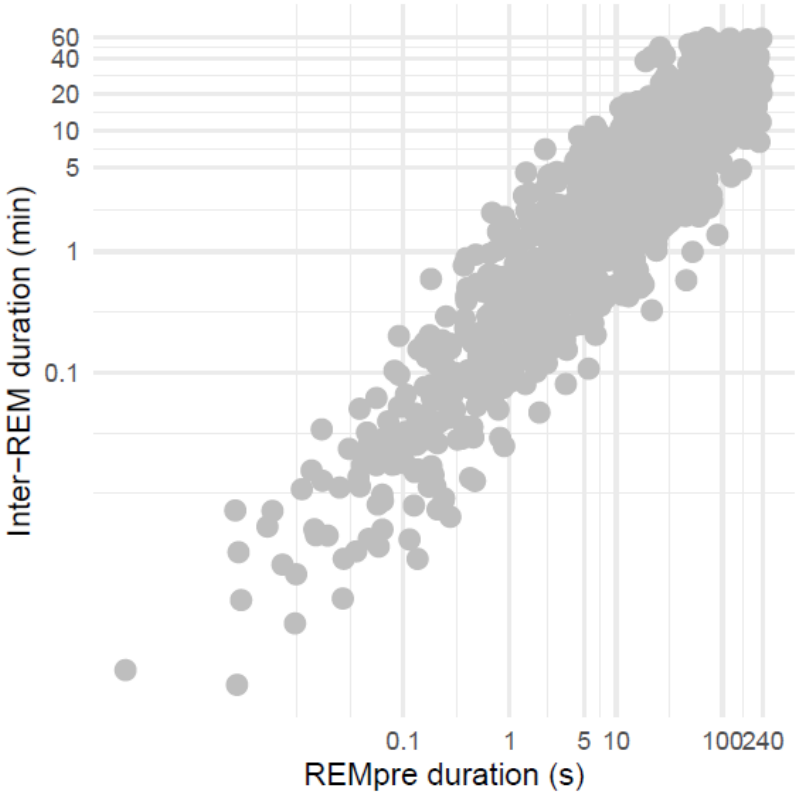
Example

- From Nature Communications (2018). DOI: 10.1038/s41467-017-02765-w
- Duration data often have a skewed distribution
- Assumption of normal distribution not justified and regression line not correct



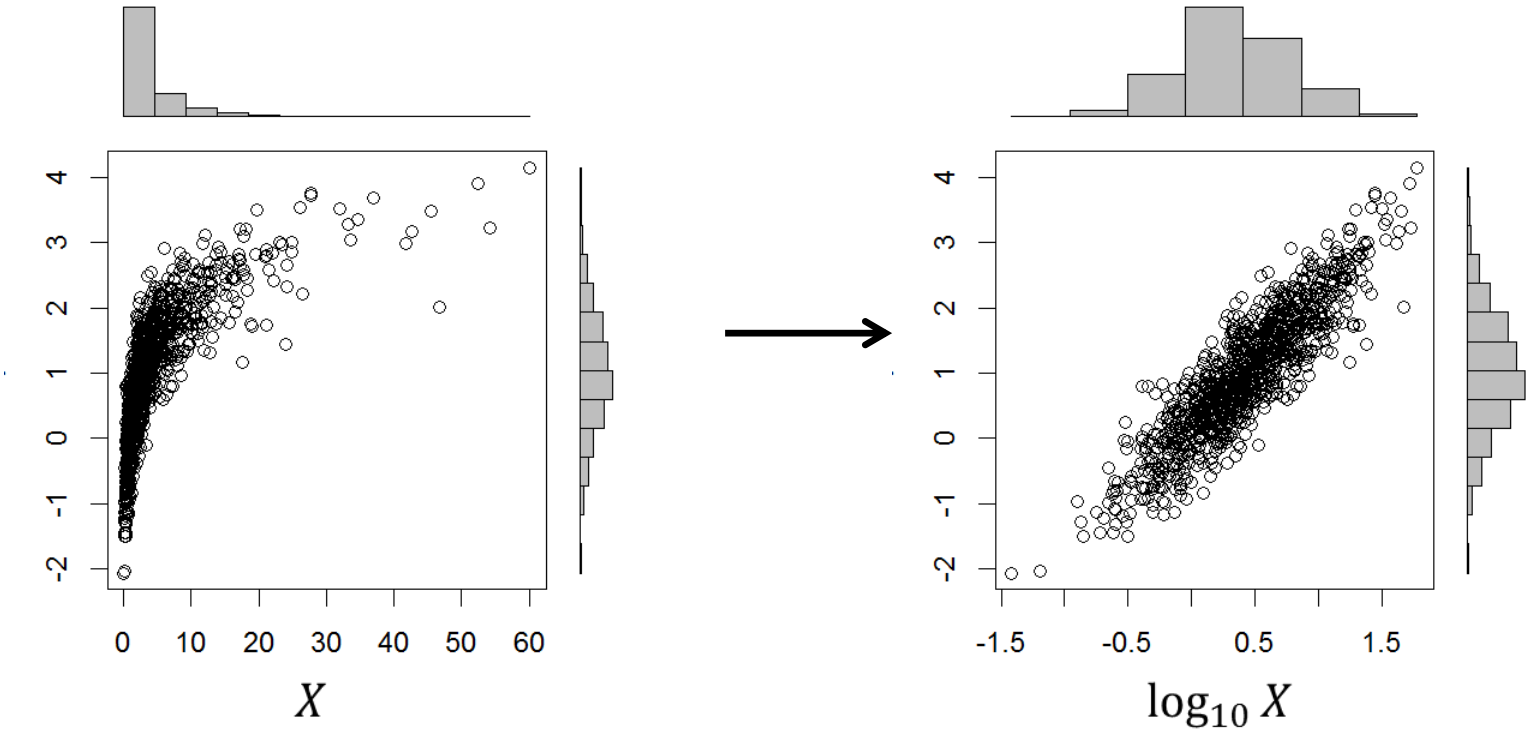
Log-transformed (artificial) data





Graphical display: Scatter plot

– Try to transform variable with skewed distribution before plotting



Scaling of covariables

- Sometimes, a change by +1 unit in the covariable is a negligible change
  - Example:
    - Age as a covariable in a study of adults aged 20-80→ a change by +1 year is usually expected to change the outcome only minimally
    - Effect if age changes by +10 years is much easier to interpret

  → Scale covariables prior to the analysis in a way that a one-unit change is interpretable
- Intercept corresponds to expected value if covariable has value 0
  - Often not meaningful, unless covariables are centered

Scaling of covariables in R - example

- covariable **`age`**
  - Use **`I(age/10)`** instead of **`age`** in the statistical model
    → regression slope corresponds to a 10-year increase in age
  - Use **`I(age-20)`** instead of **`age`** in the statistical model
    → regression intercept corresponds to a 20 year old
  - Use **`I((age-20)/10)`** instead of **`age`** in the statistical model
    → regression intercept corresponds to a 20 year old
    → regression slope corresponds to a 10-year increase in age

R-output for the glucose example (glucose slope by +10 mmol/l increase)

```
> regr.glucose <- lm(vcf~I((glucose-5)/10),data=glucose.vcf)
> summary(regr.glucose)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.20763    0.07209  16.751 1.26e-13 ***
I((glucose - 5)/10)  0.21963    0.10454   2.101   0.0479 *
```

$\hat{a}$, corresponds to expected value for individuals with glucose 5

$\hat{b}$, corresponds to expected change in vcf if glucose increases by 10 (coefficient is 10-fold higher than before, but p-value unaffected)

Categorical covariables

- Categorical covariables can easily be included in a linear regression model but they need to be coded.
- Easiest coding: "dummy coding"
  - Binary variable: in computations one level is 0 (reference level) and the other is 1
  - E.g. covariable gender:
    - "female" → 0
    - "male" → 1
- Regression with one binary covariable
  - a (intercept) corresponds to the reference level
  - b (slope) corresponds to the difference between the two groups

  → Testing the null hypothesis H_0: b=0 is equivalent to a two sample t-test (which assumes equal variances in both groups)
- R will automatically do the coding to dummy for you

R-example: Binary covariables (abbreviated output)

Regression

```
> regr.hct1 <- lm(hct1~sex,data=dengue)
> summary(regr.hct1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.1719     0.2576 187.025   <2e-16 ***
sexmale       0.4226     0.3650   1.158    0.247
```

Mean hct1 (hematocrit) in females

Estimated difference of hematocrit in males (compared to females)

Two sample t-test

```
> t.test(hct1~sex,data=dengue,var.equal=T)

t = -1.1579, df = 508, p-value = 0.2474
```

Confidence intervals for the regression line and prediction intervals

Glucose example

- Assume a new patient with a glucose of 5 mmol/l arrives → we'd like to predict his Vcf
- Estimate
  - Linear regression predicts a mean Vcf of 1.10 + 0.0220·5 = 1.21
- Confidence interval: uncertainty in mean value
  - What are plausible population mean values of Vcf for patients with a glucose of 5 mmol/l? → Confidence interval around regression line.
- Prediction interval: variation in measured value
  - What are plausible observed Vcf values in patients with a glucose of 5 mmol/l → Prediction interval around regression line
  - Takes into account that patient values vary around the mean value

$$\mathrm{Vcf}_i = 1.10 + 0.022 \cdot 5 + \varepsilon_i$$

- 95% prediction interval will be wider than the 95% confidence interval

Glucose example – confidence and prediction interval with R

- Assume a patient with a glucose of 5 mmol/l

```
> newdata <- data.frame(glucose=5)
```

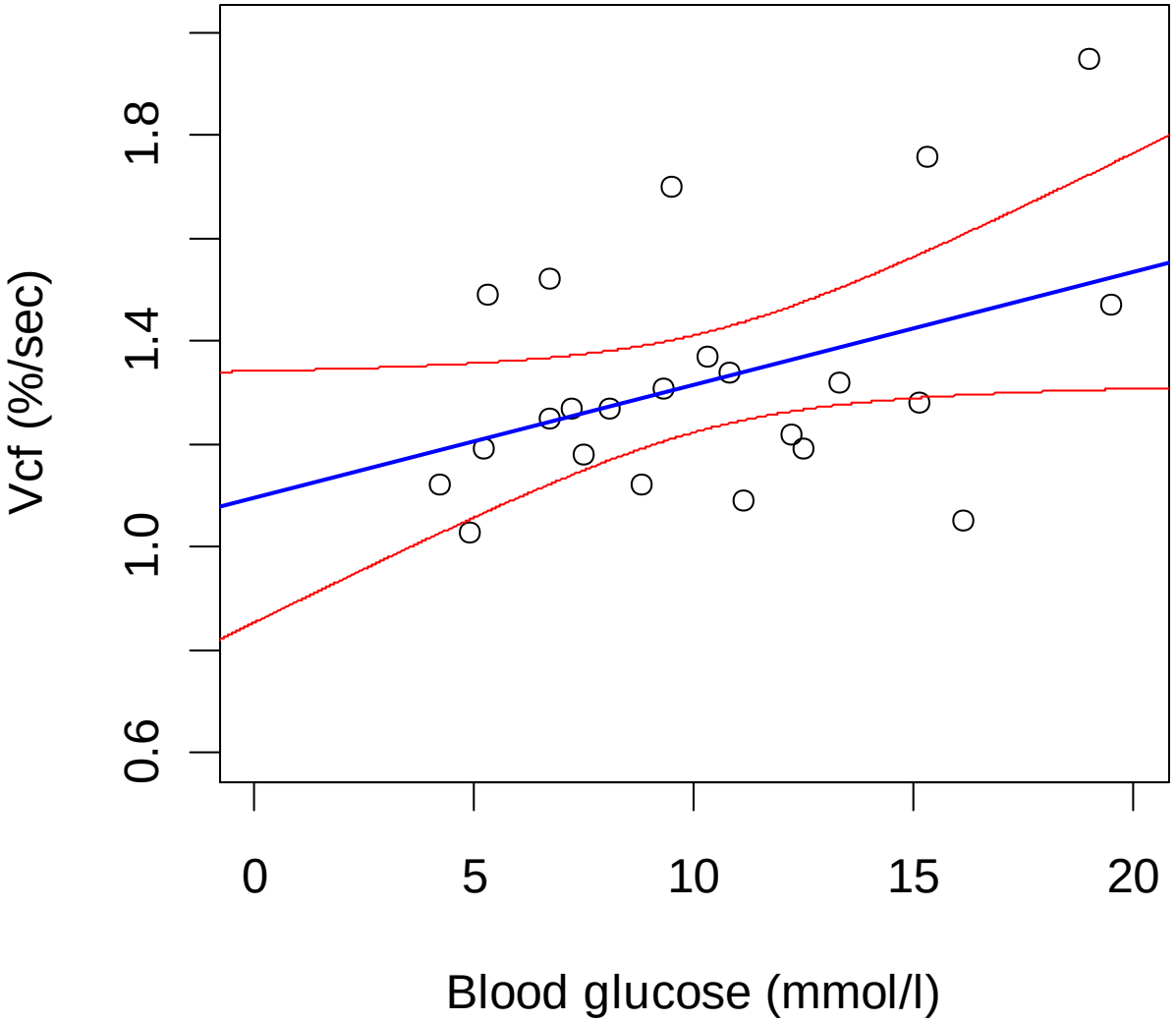
- 95% confidence interval for population mean in patients with glucose 5 mmol/l

```
> predict(regr.glucose,newdata,interval="confidence")
       fit      lwr      upr
1.207627 1.057702 1.357553
```

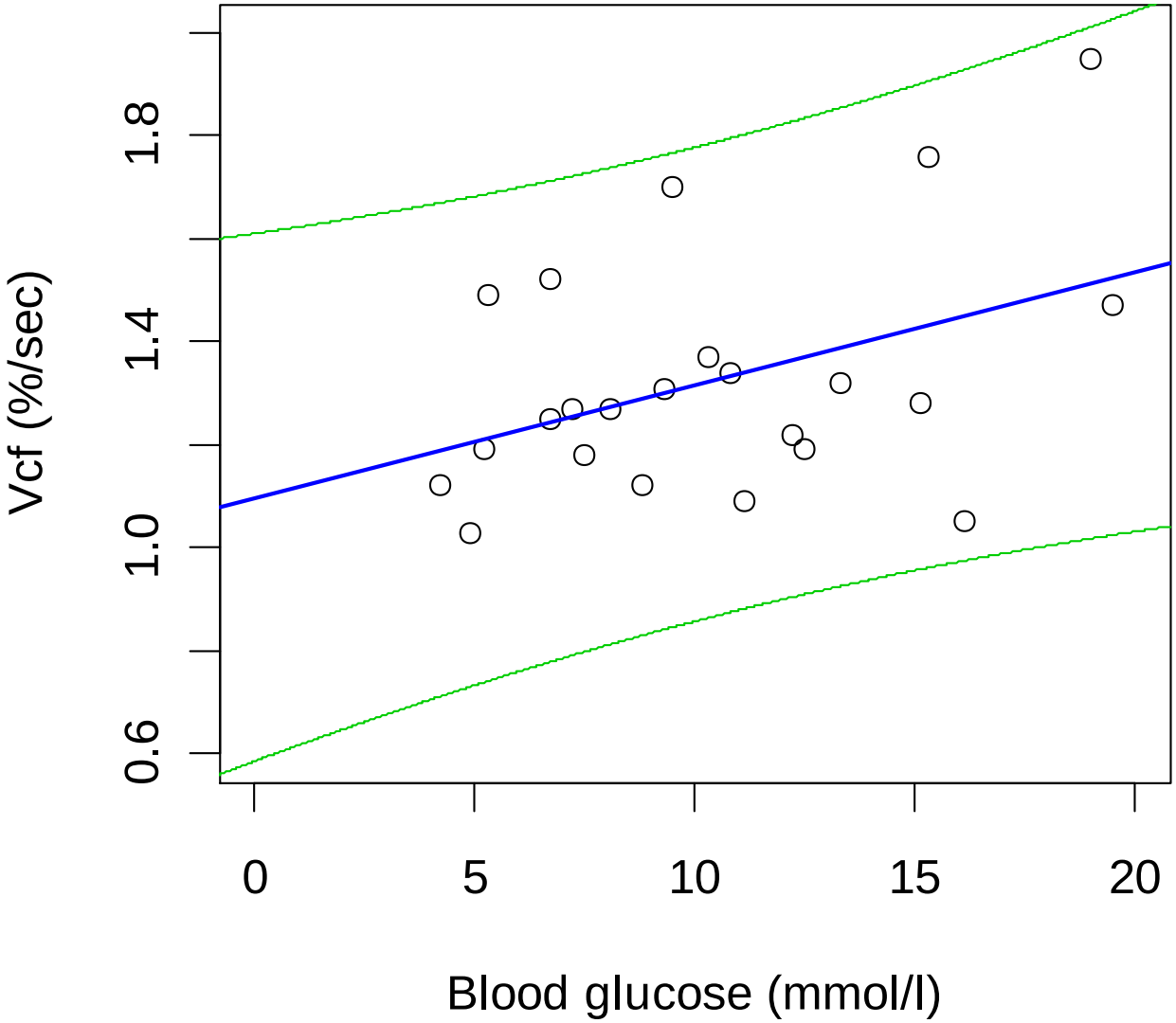
- 95% prediction interval for future patients with glucose 5 mmol/l

```
> predict(regr.glucose,newdata,interval="prediction")
       fit       lwr      upr
1.207627 0.7326991 1.682556
```

Glucose example – 95% confidence intervals for the regression line (conditional population mean)

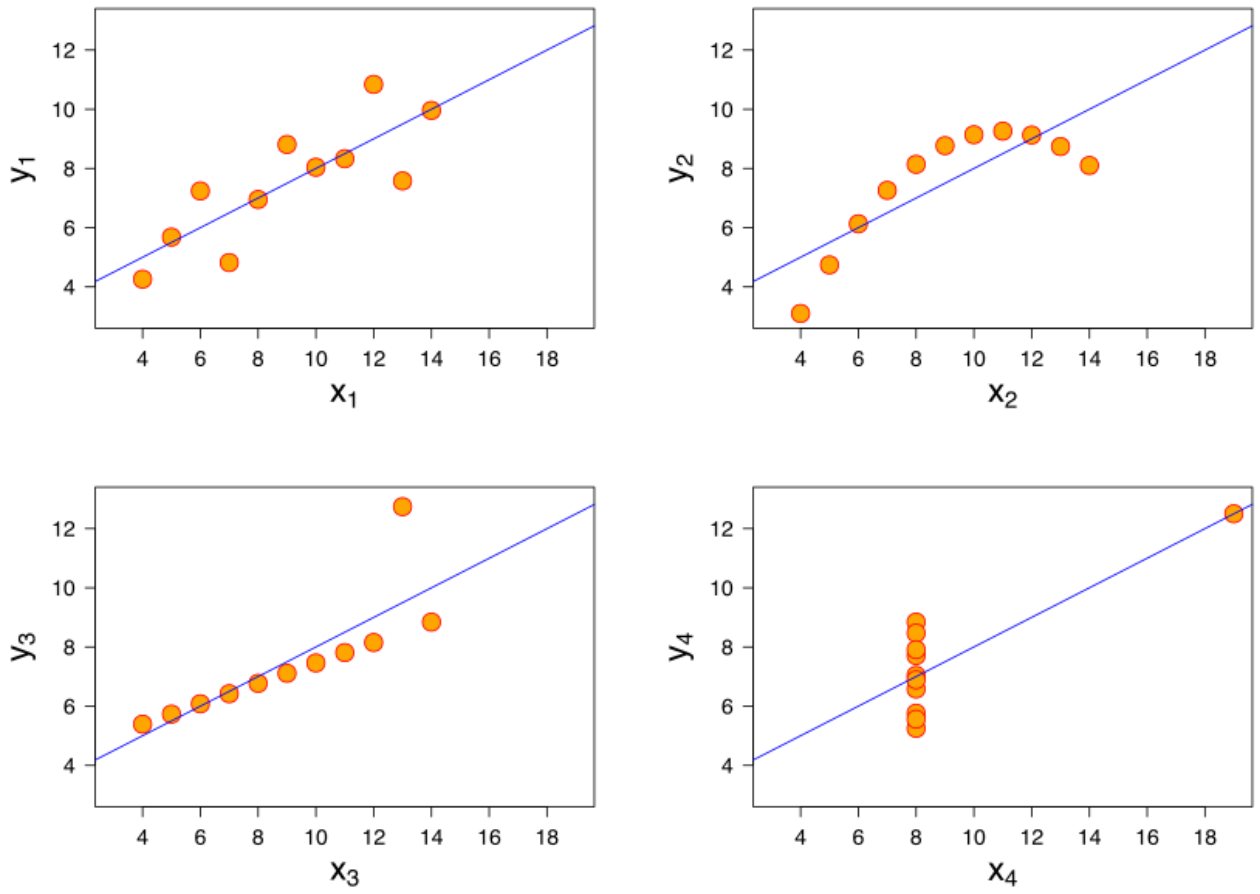


Glucose example – 95% prediction intervals for future patients



REGRESSION DIAGNOSTICS

The Anscombe's quartet



Linear regression line: Y = 3.0 + 0.5 * X

Anscombe FJ (1973) Am Stat, 27(1): 17-21 (Figure from Wikipedia)

Assumptions in linear regression

- The population mean of Y depends linearly on x
- The deviations of Y-observations from the (conditional) population mean, i.e. the residuals
  - are normally distributed
  - have constant standard deviation σ (i.e. independent of the x's)
  - are independent of each other

→ If these assumptions are violated, the estimates, confidence intervals and tests for regression coefficients are biased

→ Need to check assumptions

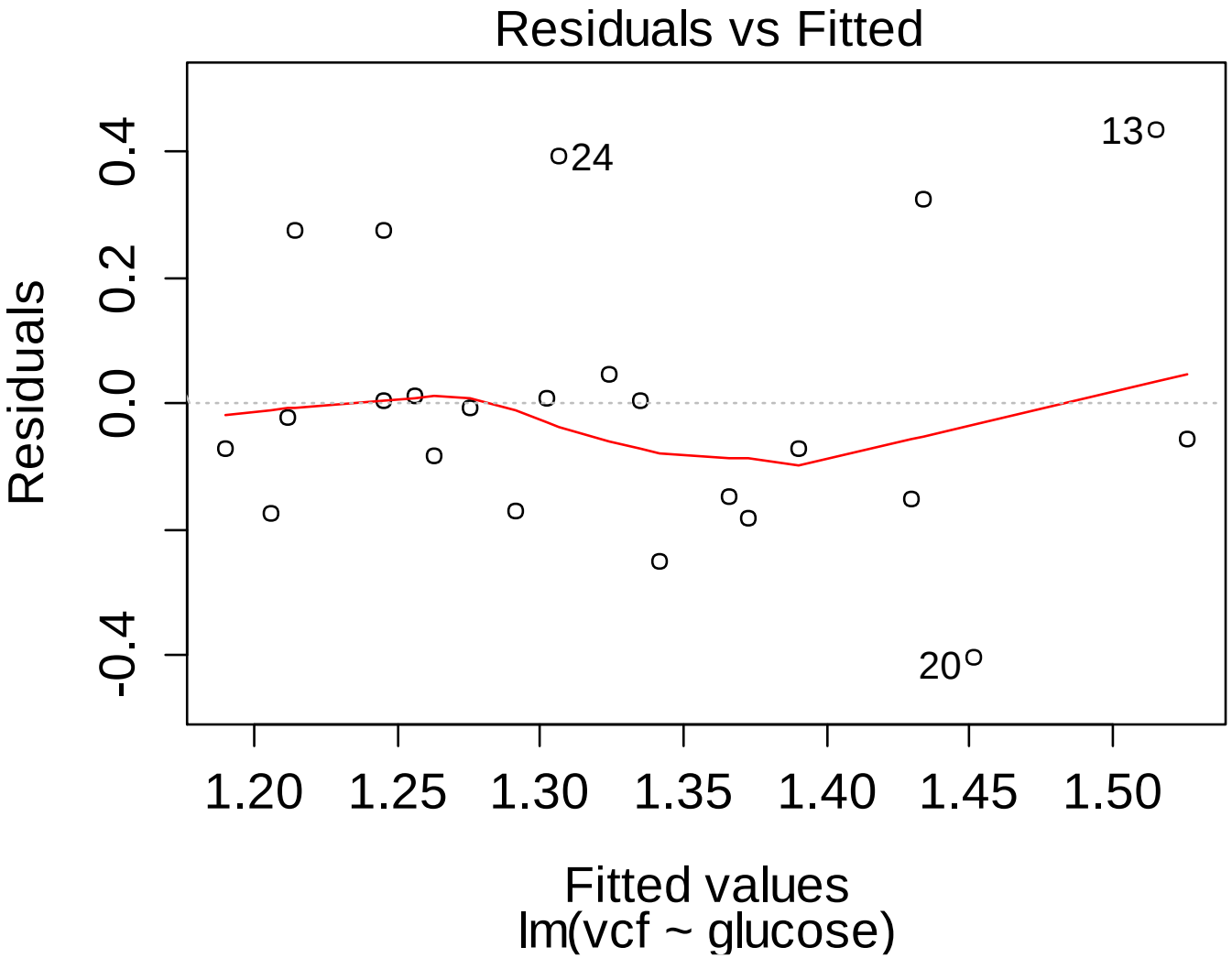
→ Usually, checks are done by graphical examination of the residuals (diagnostic plots)

Diagnostic plot 1: residual plot

- Also called Tukey-Anscombe plot
- Plot of fitted values (predicted values) versus residuals
  - Fitted value for observation i: $a+b \cdot x_i$
  - Residual for observation i: $y_i - (a+b \cdot x_i)$

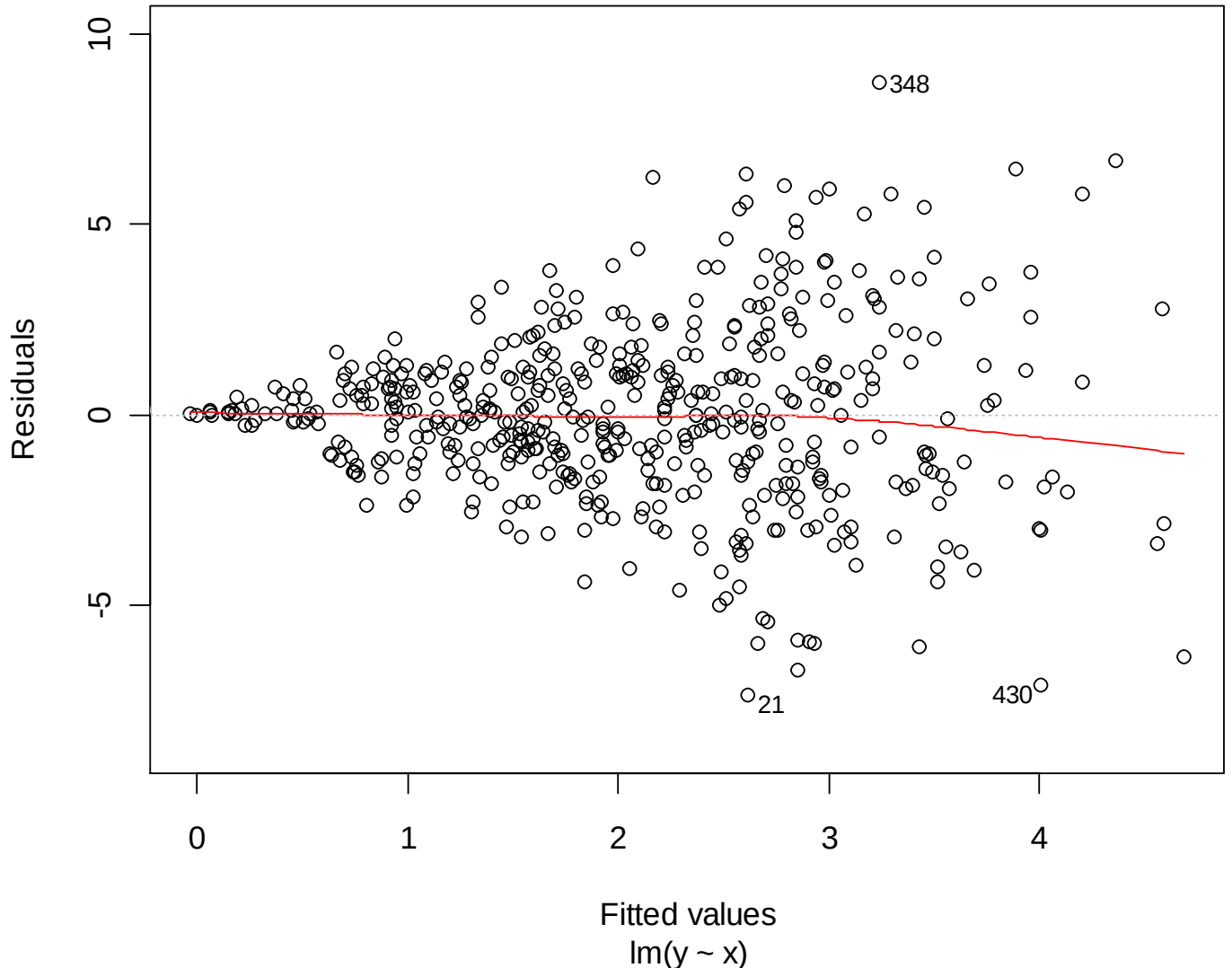
- Check if
  - Residuals have similar spread (standard deviation) regardless of the fitted values
  - No systematic patterns visible (they could indicate a non-linear association)

Tukey-Anscombe plot for the example



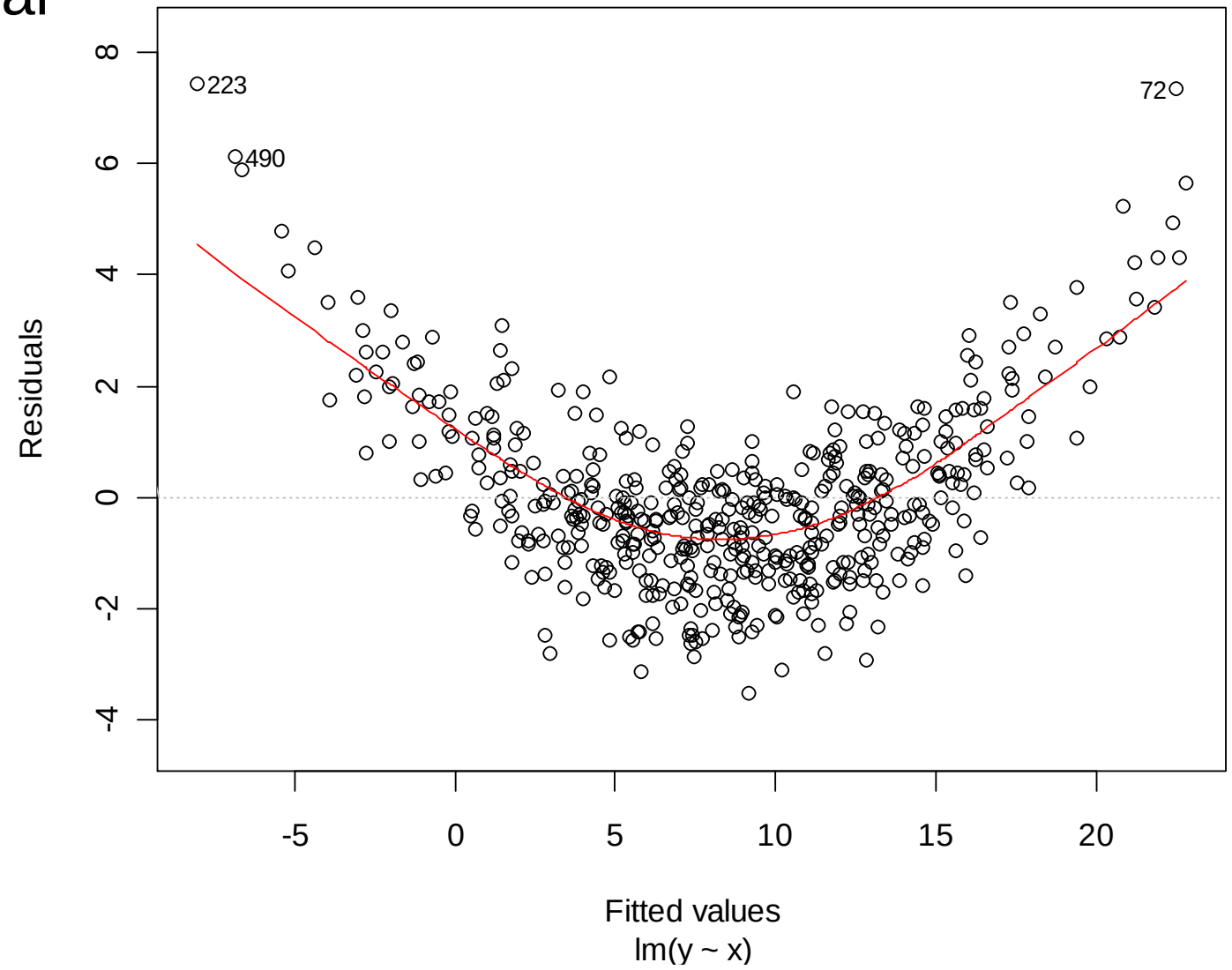
Typical violation of assumptions in Tukey-Anscombe plot I

Variance of residuals increases with fitted values →
often a log-transformation of the outcome variable will resolve this



Typical violation of assumptions in Tukey-Anscombe plot II

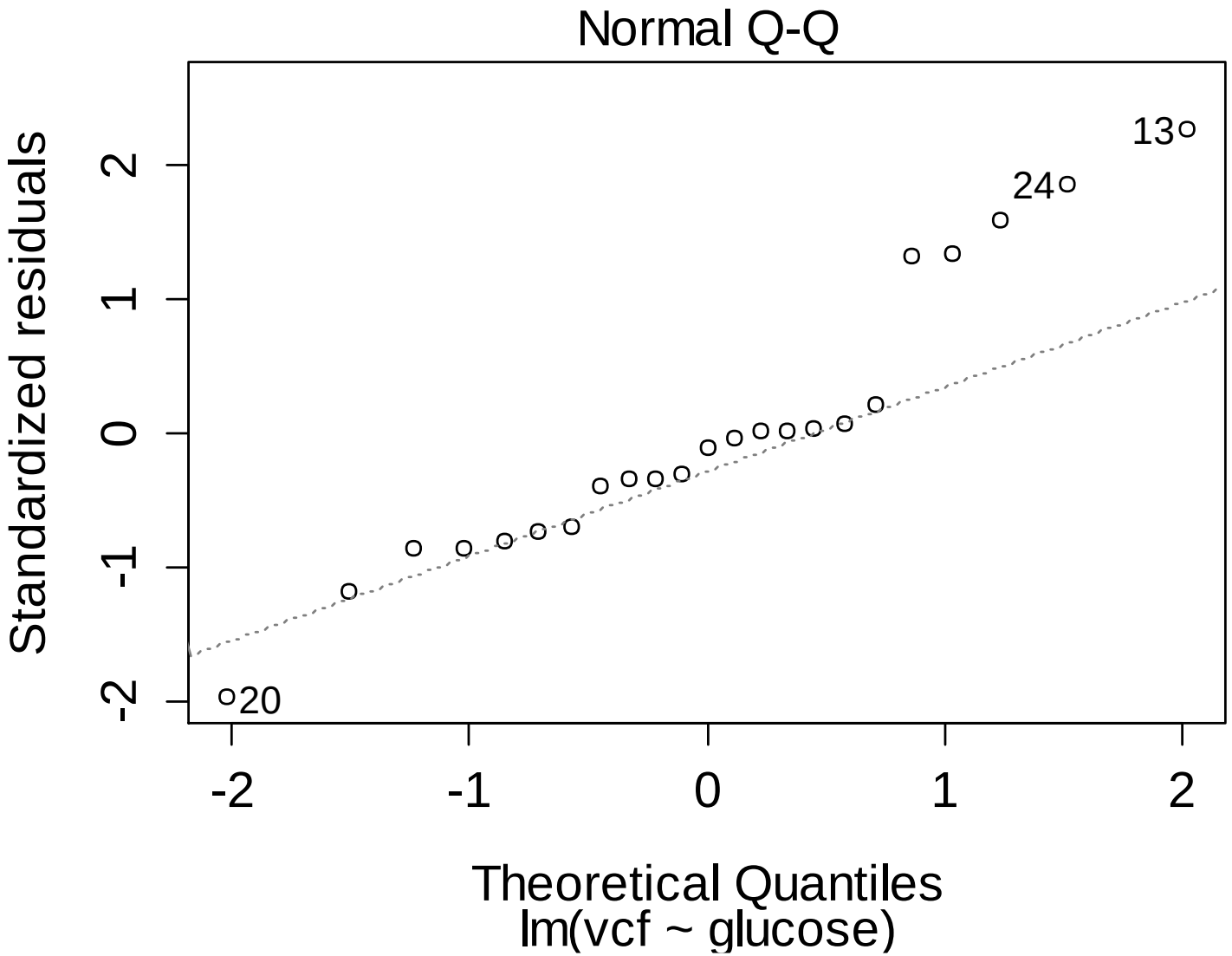
Systematic pattern, may indicate that true association is not linear



Diagnostic plot 2: normal plot of residuals

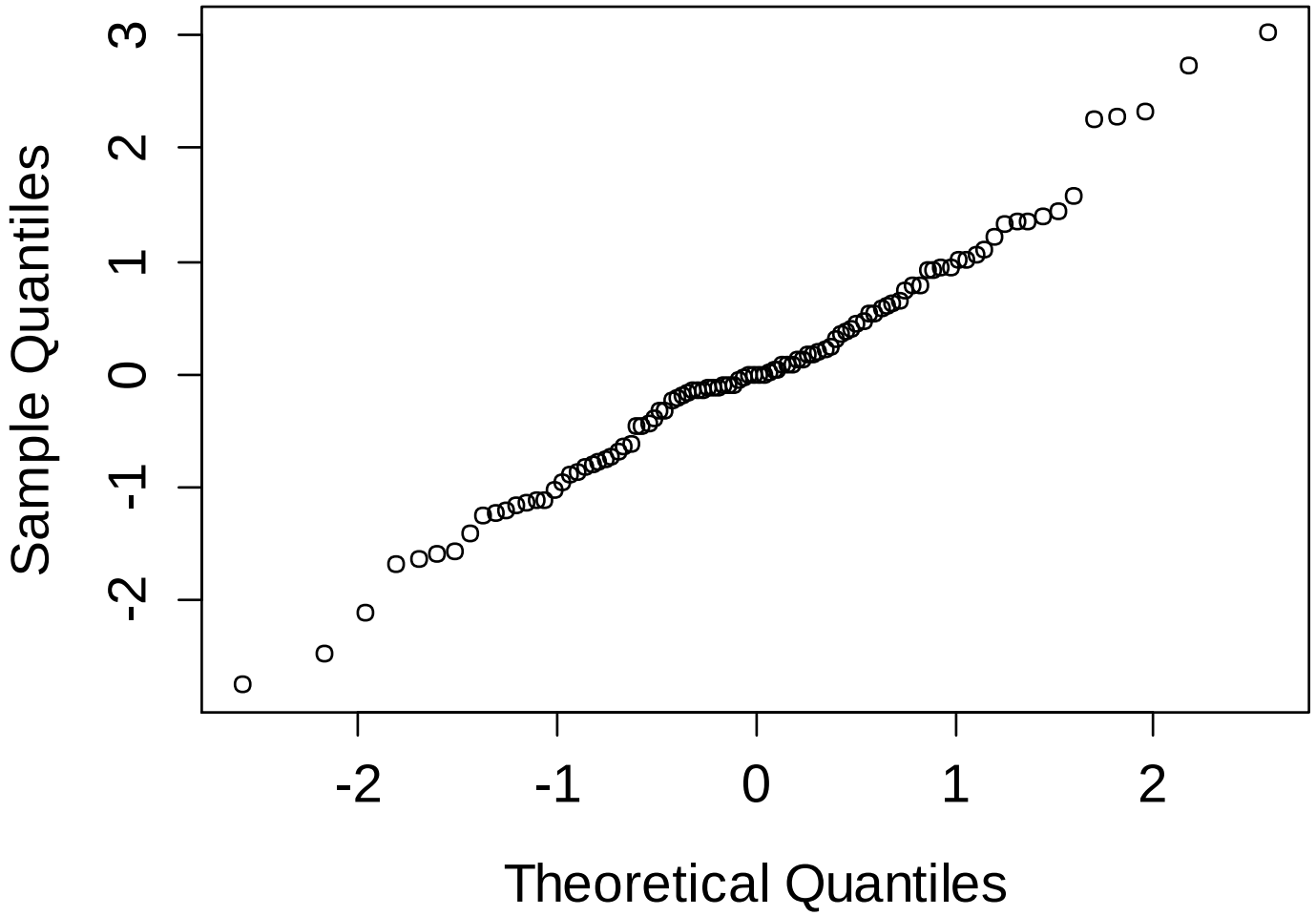
- The normal plot (or Q-Q plot) displays the ordered residuals (y-axis) versus their expected values assuming a normal distribution (x-axis)
- Strong deviations from a straight line indicate that the residuals are not from a normal distribution
- In a normal plot, it is much easier to see what's going on in the tails of a distribution than in a histogram
- For small sample size, normal plots are difficult to interpret

Normal plot for the example



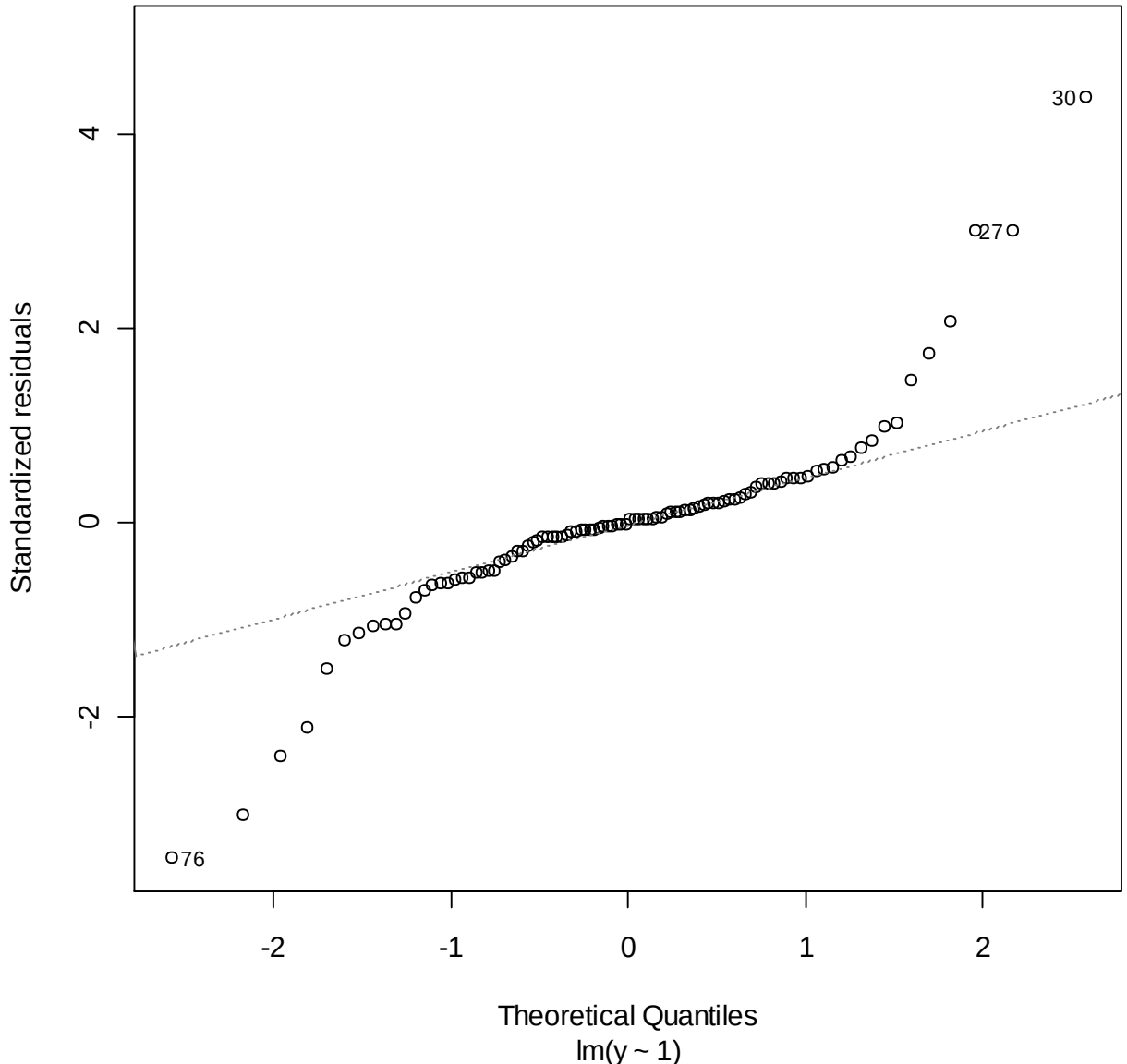
Normal plot for a larger dataset (n=100)

Residuals show good agreement with the normal distribution



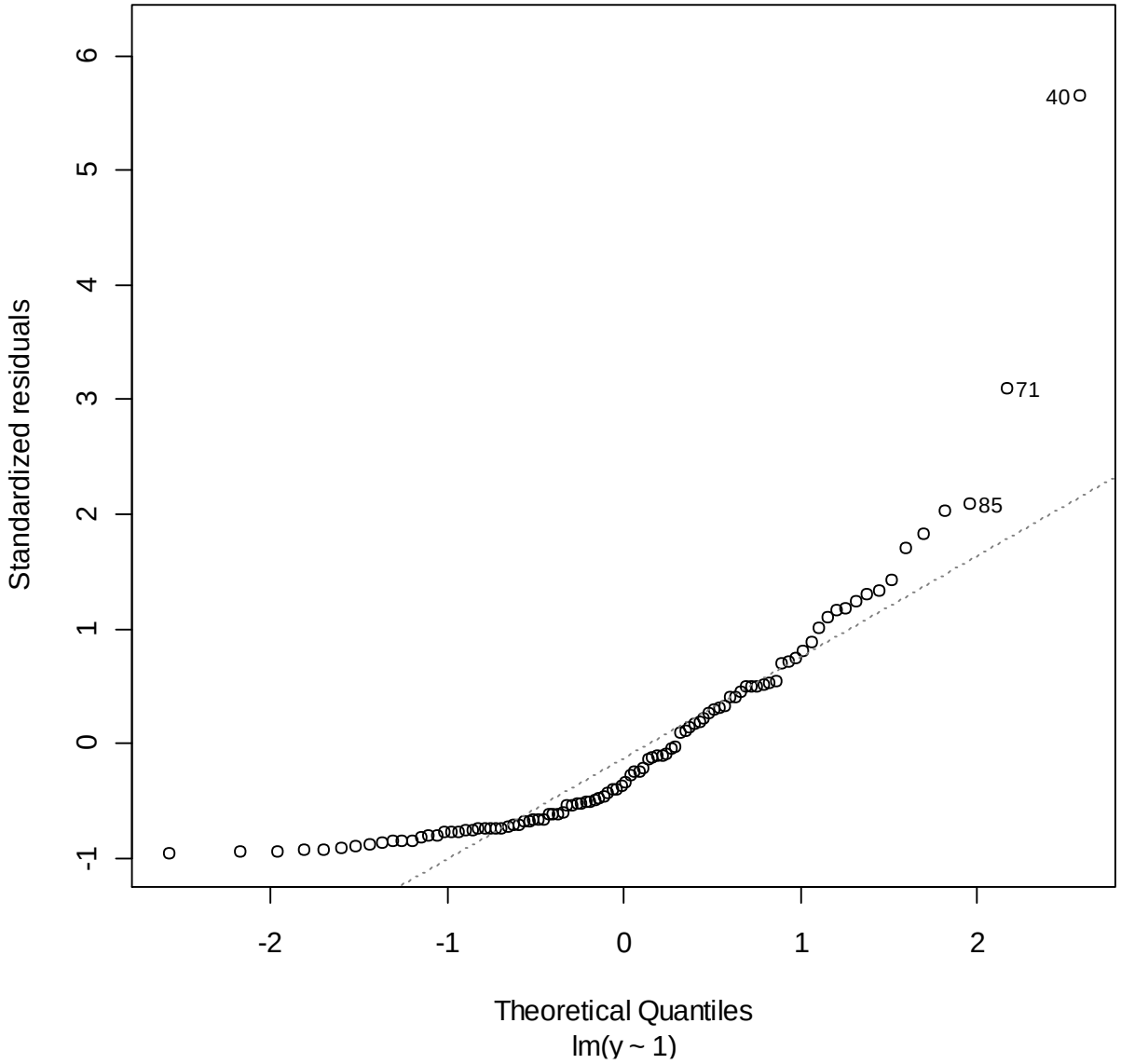
Typical violation of assumptions in normal plot I

Residuals have longer tails than a normal distribution

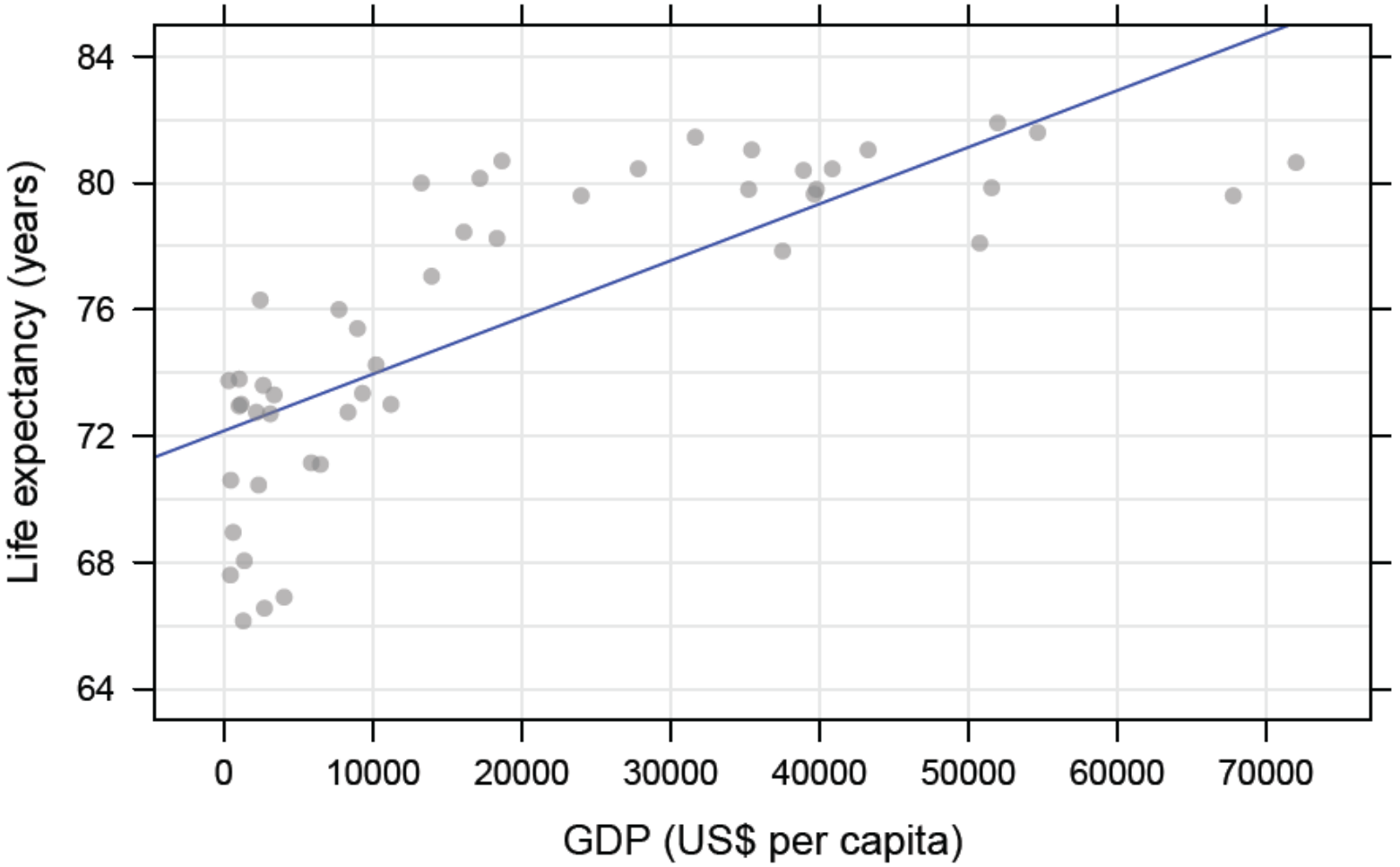


Typical violation of assumptions in normal plot II

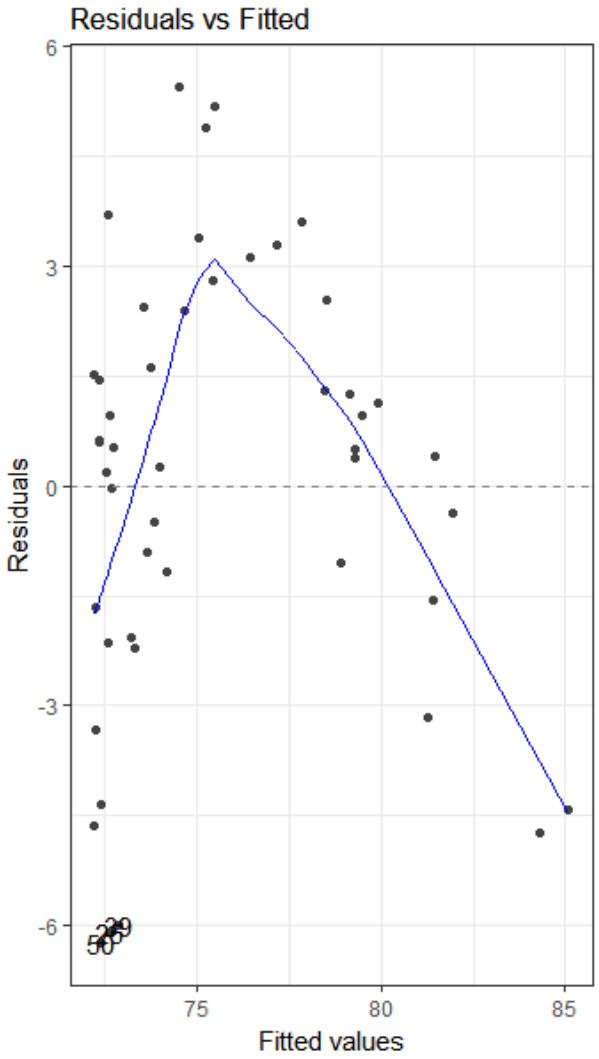
Distribution of residuals is skewed

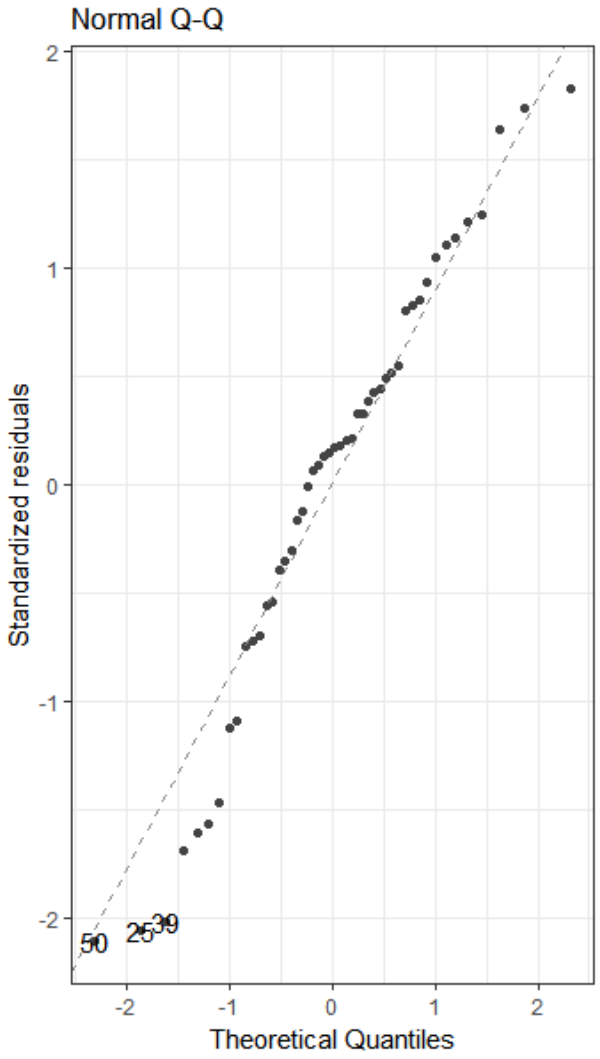


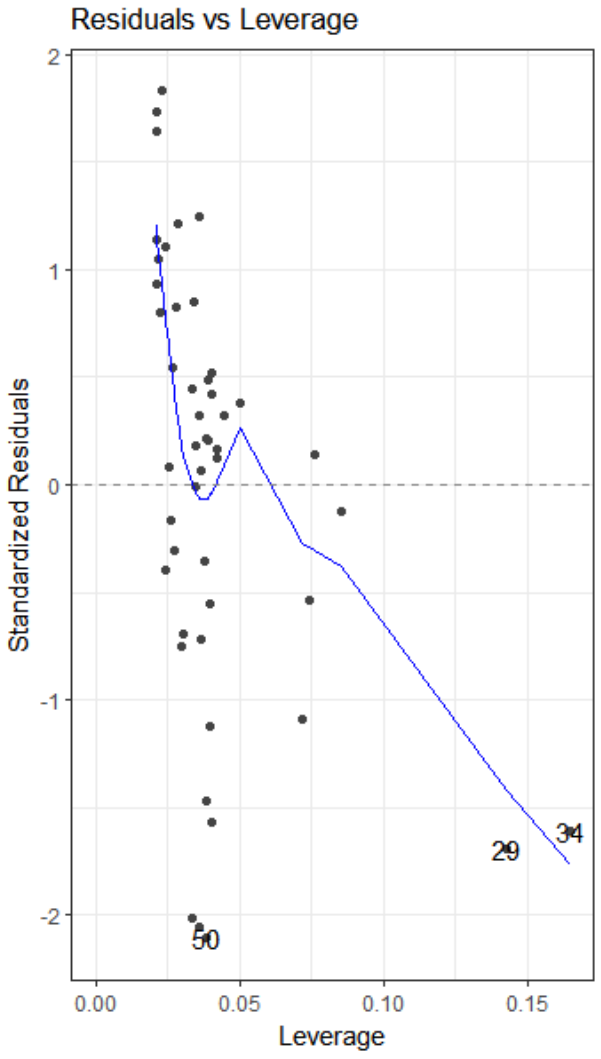
A badly fitted model (WHO)



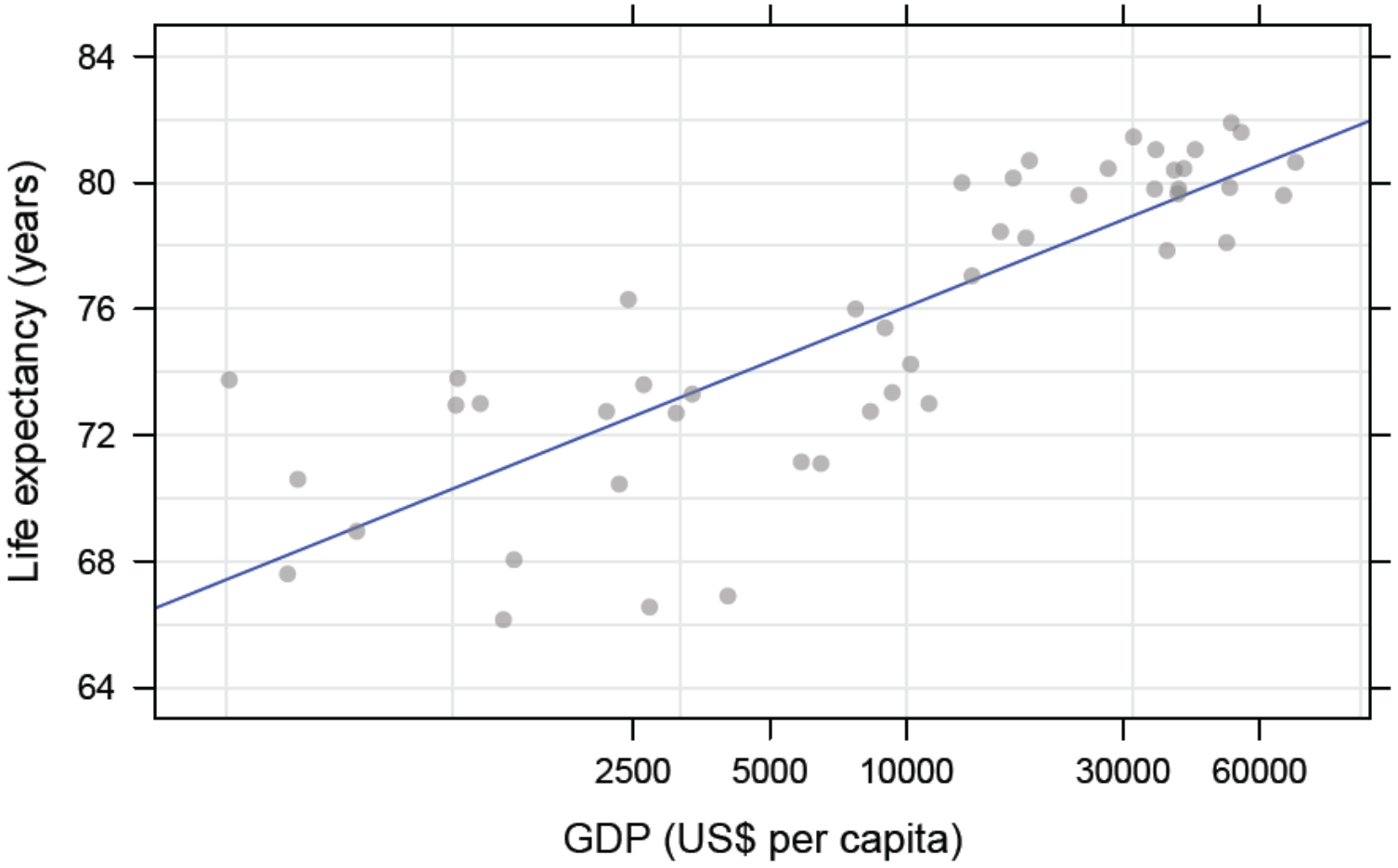
Three residual plots (WHO)



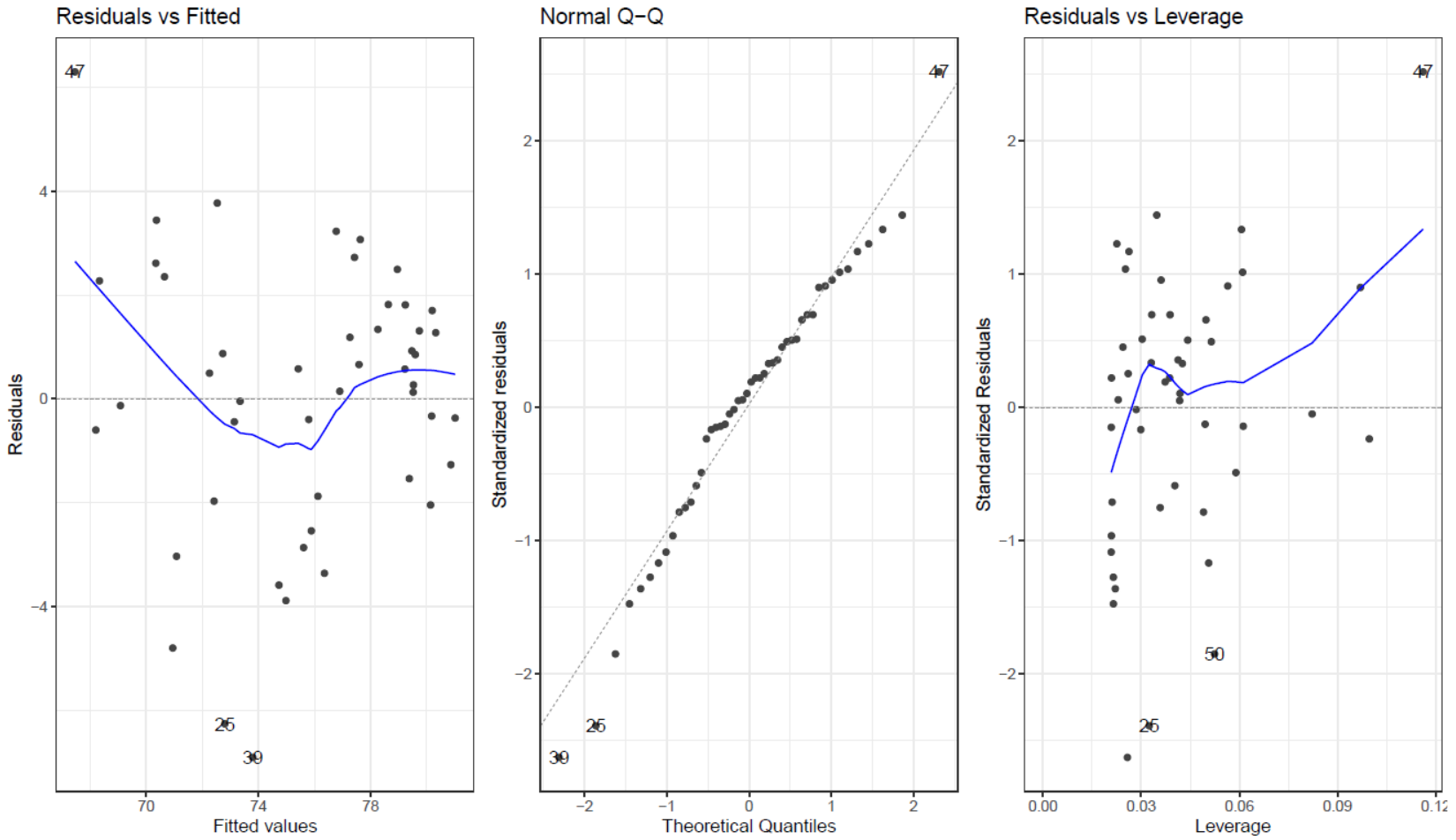




After transformation of GDP



Residual plots after log transformation



Residual diagnostics

- A number of other residual plots have been proposed
  - E.g.: residuals versus Leverage: are there observations that may have large impact on parameter estimates?
- Looking at residual plots is an integral part of every regression analysis
  - Always examine Tukey-Anscombe plot and normal plot
  - Detected violations of model assumptions may help to
    - Understand the data better
    - Find a better model
    - Interpret the results with the appropriate caution if results depend on model
    - Conclude on robustness of model if results of interest do not change much over different models

Summary

- First step to assess the association between 2 continuous variables: graphical display
- Linear regression model:
  - Y = a + bX
  - Fit a linear regression model in R with **lm()**
  - Model diagnostics using plots