7. Multivariable linear regression

Introduction to Medical Statistics

OUCRU, Ho Chi Minh City

March 23-27, 2026

Van Thuan Hoang

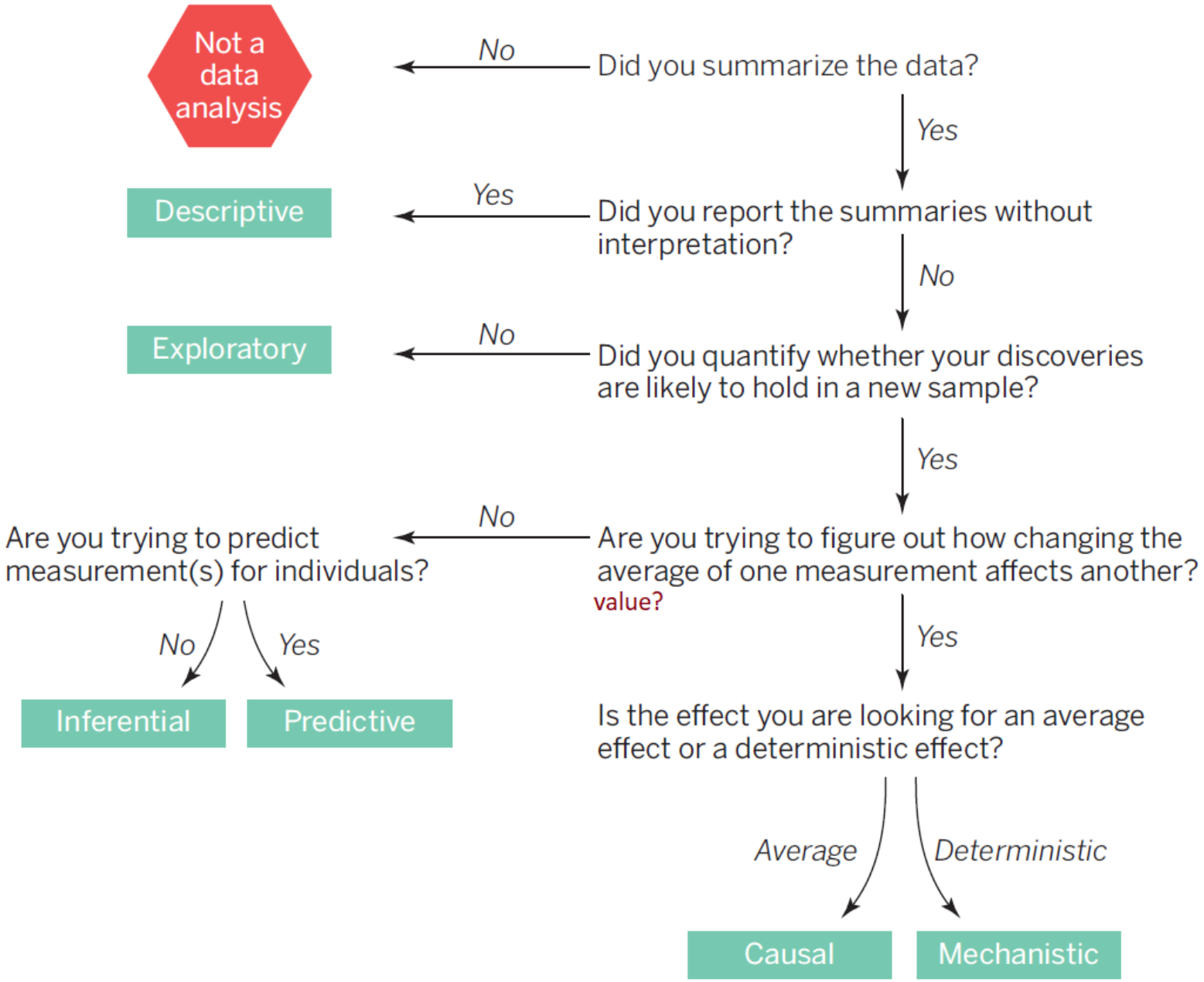
and the biostatistics crew

Model formulation, estimation and testing

Example: Birth weight of babies

- Data: 96 children from Baystate Medical Center, Springfield, Massachusetts during 1986
- Variables:
  - Weight of child at birth (g) - outcome/response
  - Characteristics of the mother: covariables/predictors
    - Age at birth (years)
    - Weight at last menstrual period (kg)
    - Smoking status during pregnancy (yes/no)

Data analysis flowchart

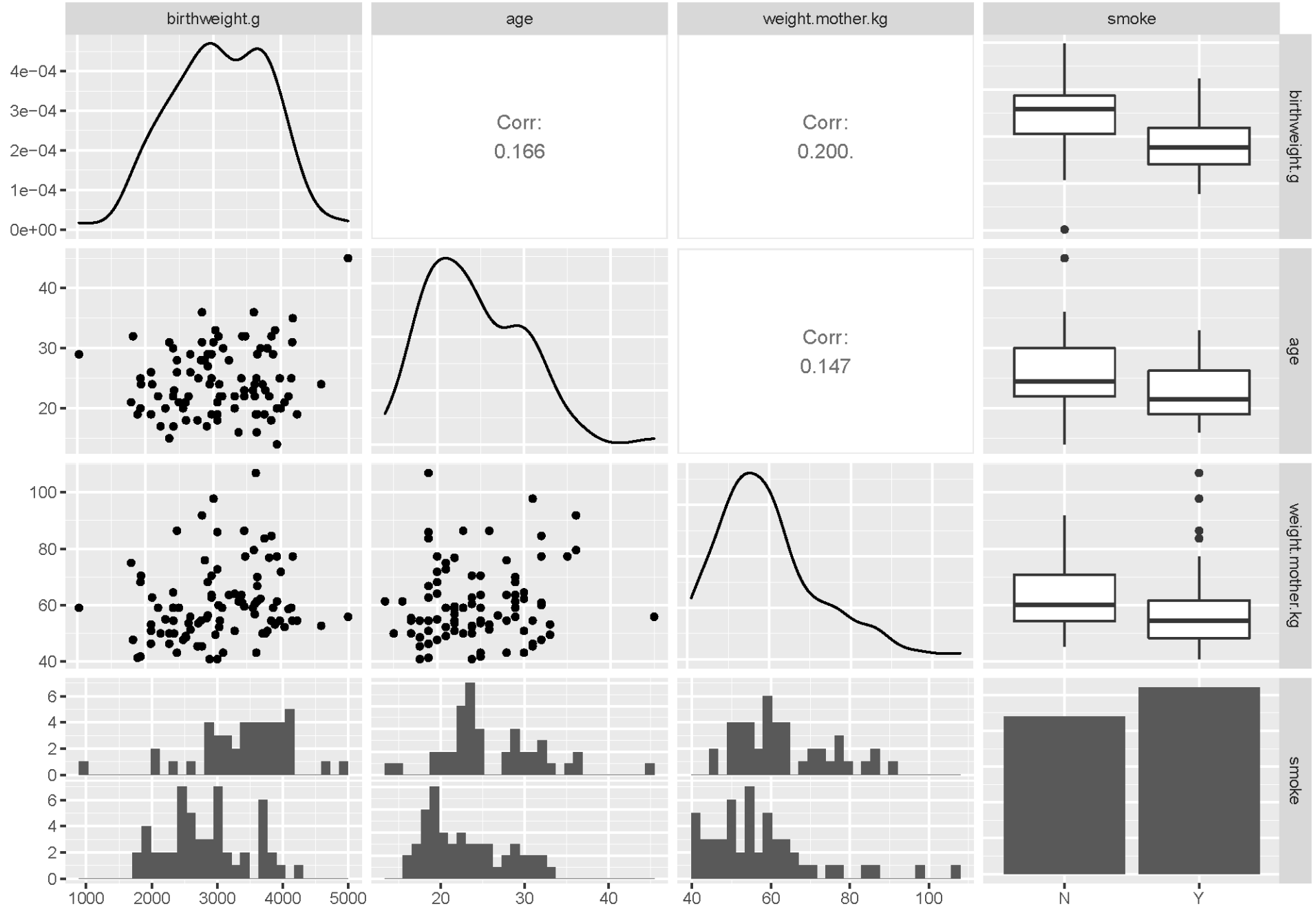


Example: Birth weight of babies

→ How do i) mother's age at birth, ii) weight at last menstrual period and iii) smoking status during pregnancy relate to weight of child at birth, *when considered together*?

→ What do we mean with *when considered together?* How do we interpret the results from the model.

→ Can we predict the child's weight based on these covariables?


birthweight.g
age
weight.mother.kg
smoke
Corr:
0.166
Corr:
0.200.
Corr:
0.147
4e−04
3e−04
2e−04
1e−04
0e+00
40
30
20
100
80
60
40
6
4
2
0
1000
2000
3000
4000
5000
20
30
40
40
60
80
100
N
Y

Separate linear regressions

```
> summary(lm(birthweight.g~age,data=birthweight.data))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2583.54     326.44   7.914 4.83e-12
age            21.37      13.09   1.632    0.106

> summary(lm(birthweight.g~weight.m.kg,data=birthweight.data))
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2442.418    341.806   7.146 1.89e-10
weight.m.kg       11.001      5.562   1.978   0.0509

> summary(lm(birthweight.g~smoke,data=birthweight.data))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3428.8      100.4  34.148  < 2e-16
smokeTRUE     -601.9      136.4  -4.412 2.73e-05
```

What do we miss with separate linear regressions?

- Smoking mothers on average have lower weight, hence it is not clear whether the positive association between mother weight and baby weight is biological or *confounded* by the effect of smoking on baby weight
- The effect of smoking on baby weight may be different depending on age. E.g. older women on average have smoked for a longer period. *Is the effect of smoking modified by age?*

The multiple regression model (with 3 covariables)

- Regression model with 3 covariables

$$E(Y) = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$$

- In our example

$$E(\text{Birthweight}) = a + b_1 \cdot \text{age.mother} + b_2 \cdot \text{weight.mother} + b_3 \cdot \text{smoke.mother}$$

  - b1: effect of age.mother on birthweight, other two covariables fixed
  - b2: effect of weight_mother on birthweight, other two covariables fixed
  - b3: effect of smoke_mother on birthweight, other two covariables fixed
  - a: the average birthweight for “null” value of other covariables

- Categorical covariable: one level is the baseline category and the effects of other levels on outcome compare to the baseline category will be estimated by using dummy covariables (1 for a specific level, 0 for other levels)
  - In general: for categorical covariable with k levels → need k-1 dummy covariables (the intercept refers to the baseline category of the categorical covariable)

Estimates and interpretation

- Estimation technique basically the same as for univariable linear regression
- Interpretation of coefficients different
  - b_2 is difference in average birthweight if the mothers differ by 1kg but have same age and smoking status

    →b_2 is the relation between woman's weight and baby's weight after "controlling for" age and smoking status
  - If covariables are correlated, coefficients of a multivariable linear regression are different from separate simple linear regression analyses

Multivariable regression output

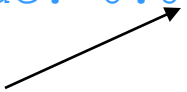
1. “Overall” p-value of test for null hypothesis that none of the proposed covariables is related to outcome
2. For each covariable: relation with outcome, adjusted for all other covariables in the model (point estimates and their corresponding 95% confidence intervals)
3. For each covariable: p-value of test for null hypothesis that there is no relation with outcome, after adjustment for other covariables
4. A measure that indicates how much of the observed variation in the outcome can be explained by the model

R: output multiple regression (abbreviated)

```
> fit.bwt <- lm(birthweight.g~age+weight.m.kg+smoke,
                data=birthweight.data)
> summary(fit.bwt)

                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2896.687    460.009   6.297 1.02e-08 ***
age                5.487     12.677   0.433 0.666138
weight.m.kg        6.168      5.324   1.158 0.249670
smokeTRUE       -549.228    144.990  -3.788 0.000271 ***

Multiple R-squared: 0.186
F-statistic: 7.008 on 3 and 92 DF,  p-value: 0.0002689
```

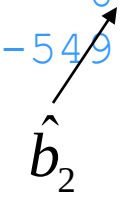


p - value of "global F - test", i.e. of the null hypothesis that all covariates jointly do not affect birthweight.

R: output multiple regression (abbreviated)

```
> fit.bwt <- lm(birthweight.g~age+weight.m.kg+smoke,
                data=birthweight.data)
> summary(fit.bwt)

                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2896.687    460.009   6.297 1.02e-08 ***
age                5.487     12.677   0.433 0.666138
weight.m.kg        6.168      5.324   1.158 0.249670
smokeTRUE       -549.228    144.990  -3.788 0.000271 ***
```




```
Multiple R-squared: 0.186
F-statistic: 7.008 on 3 and 92 DF,  p-value: 0.0002689
```

R: output multiple regression (abbreviated)

```
> fit.bwt <- lm(birthweight.g~age+weight.m.kg+smoke,
                data=birthweight.data)
> summary(fit.bwt)

                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2896.687    460.009   6.297 1.02e-08 ***
age                5.487     12.677   0.433 0.666138
weight.m.kg        6.168      5.324   1.158 0.249670
smokeTRUE       -549.228    144.990  -3.788 0.000271 ***
```

p - value of null hypothesis that the mother's weight does not affect birthweight after controlling for age and smoking status

```
Multiple R-squared: 0.186
F-statistic: 7.008 on 3 and 92 DF,  p-value: 0.0002689
```

R: output multiple regression (abbreviated)

```
> fit.bwt <- lm(birthweight.g~age+weight.m.kg+smoke,
                data=birthweight.data)
> summary(fit.bwt)

                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2896.687    460.009   6.297 1.02e-08 ***
age                5.487     12.677   0.433 0.666138
weight.m.kg        6.168      5.324   1.158 0.249670
smokeTRUE       -549.228    144.990  -3.788 0.000271 ***
```

R^2, coefficient of explained variation

```
Multiple R-squared: 0.186
F-statistic: 7.008 on 3 and 92 DF,  p-value: 0.0002689
```

R^2 (coefficient of explained variation)

- Multiple R-squared (R^2)
  - In simple linear regression: Squared correlation between x and y
  - In general: Fraction of the variance of y that can jointly be explained by all covariables
  - Between 0 and 1
    - 0: no (linear) association between the x's and y
    - 1: y can be perfectly explained by the x's (perfect fit)

R: output multiple regression – confidence intervals for coefficients

```
> fit.bwt <- lm(birthweight.g~age+weight.m.kg+smoke,
                data=birthweight.data)
> confint(fit.bwt)
                        2.5 %     97.5 %
(Intercept)      1983.069547 3810.30399
age               -19.690274   30.66464
weight.mother.kg   -4.406563   16.74296
smokeTRUE        -837.190230 -261.26615
```

Tables with regression outputs in publications

- Report regression coefficients, 95% confidence intervals and p-values
- Use sensible scaling of variables
  - Below we use `I(age/10)` instead of `age` → Report age coefficient as "age (by +10 years)"
- Note: "not smoking" is reference level. We change TRUE to "-Yes"

Variable	Effect	95% CI	p-value
Mother's age (by +10 years)	54.87	(-196.90, 306.65)	0.666
Mother's weight (by +1 kg)	6.17	(-4.41, 16.74)	0.249
Smoking status			
-No	0		
-Yes	-549.23	(-837.19, -261.27)	<0.001

Prediction

- Birth weight of a baby (average, variation) if her/his mother is 35 years old (at birth), weighs 55 kgs and smoked during pregnancy?

"Create" such a mother:

```
> newdata <- data.frame(age = 35,
                        weight.mother.kg = 55,
                        smoke = TRUE)
```

Compute the prediction

```
> predict(fit.bwt, newdata=newdata, interval="prediction")
```

Three values are reported: average, and range that covers 95% of all birth weights

```
        fit         lwr         upr
1  2878.761   1505.294    4252.227
```

Assessing model assumptions

Model assumptions

- The simplest multiple regression model with 3 covariables

$$E(Y) = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$$

- Additivity: the total effect of all covariables on outcome is the sum of each covariable's effect. No interaction between covariables.

- Other assumptions as for simple linear regression
  - Linearity: the population mean of Y depends linearly on x_1, x_2, x_3
  - Distribution of random errors/residuals (the differences between observed and fitted responses)
    - are normally distributed
    - have constant standard deviation σ (i.e. the same for all combinations of x_1, x_2, x_3)
    - are independent of each other

Model assumptions

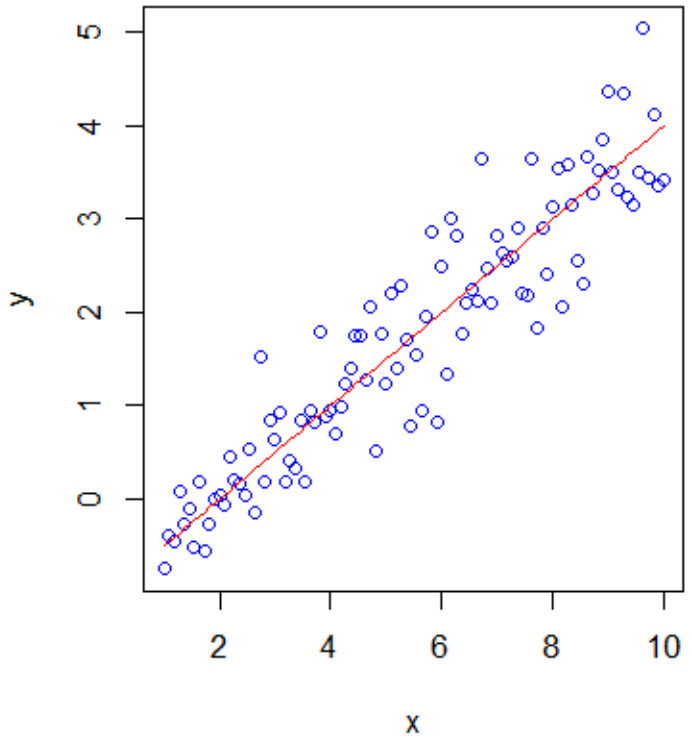
→If assumptions are violated, the estimates, confidence intervals and tests for regression coefficients are biased

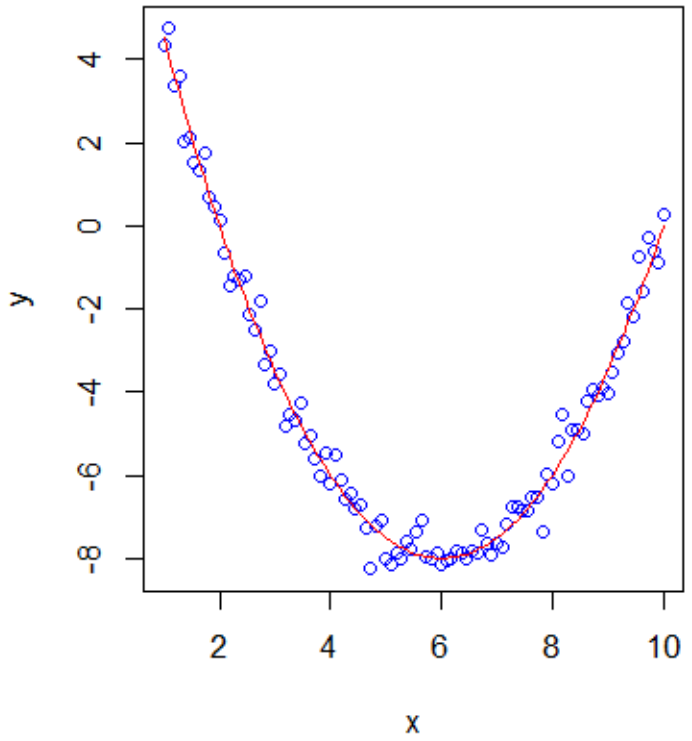
→Check assumptions, using

- Visualization
  - Diagnostic plots
- Hypothesis testing
  - Assessing the sufficiency of a simple model by comparing it to more complex model

Non-linear effects

- A covariable may not affect the outcome linearly but in a more complex non-linear way





Non-linear effects

- Sometimes a transformation of the covariable (or the outcome) is sufficient to transform the effect into a linear one
- However, non-linear effects can also be modelled explicitly
- Example: Quadratic effect
  - Model: $y=a+b_1x+b_2x^2$

    → model is still linear in the parameters a, b_1, and b_2

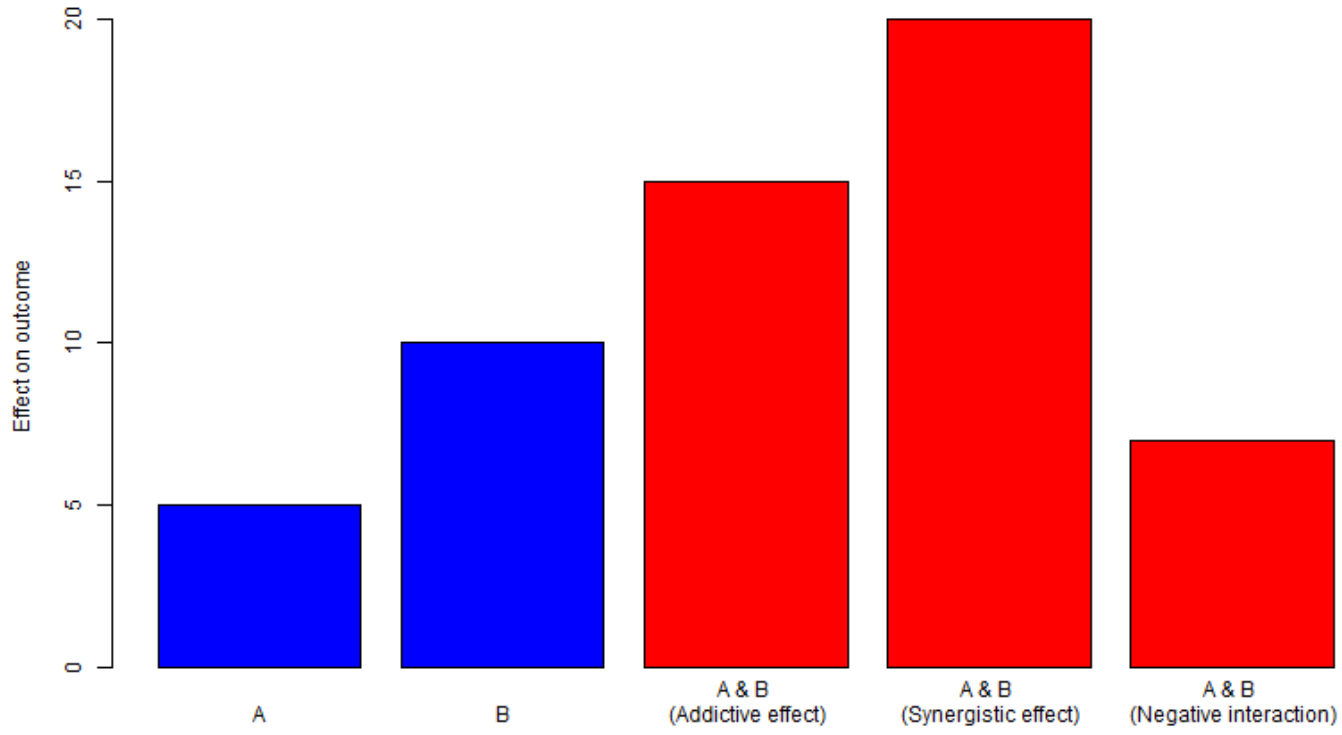
    → all we need to do is define a "new covariable" x^2 and then fit a multiple linear regression which includes both x and x^2

    → in R, we do not even have to explicitly define the "new covariable" x^2 , instead we can use y~poly(x,2) in the model formula

    → also more flexible and powerful functions than quadratic polynomials available (e.g. spline functions and fractional polynomials)

Interactions

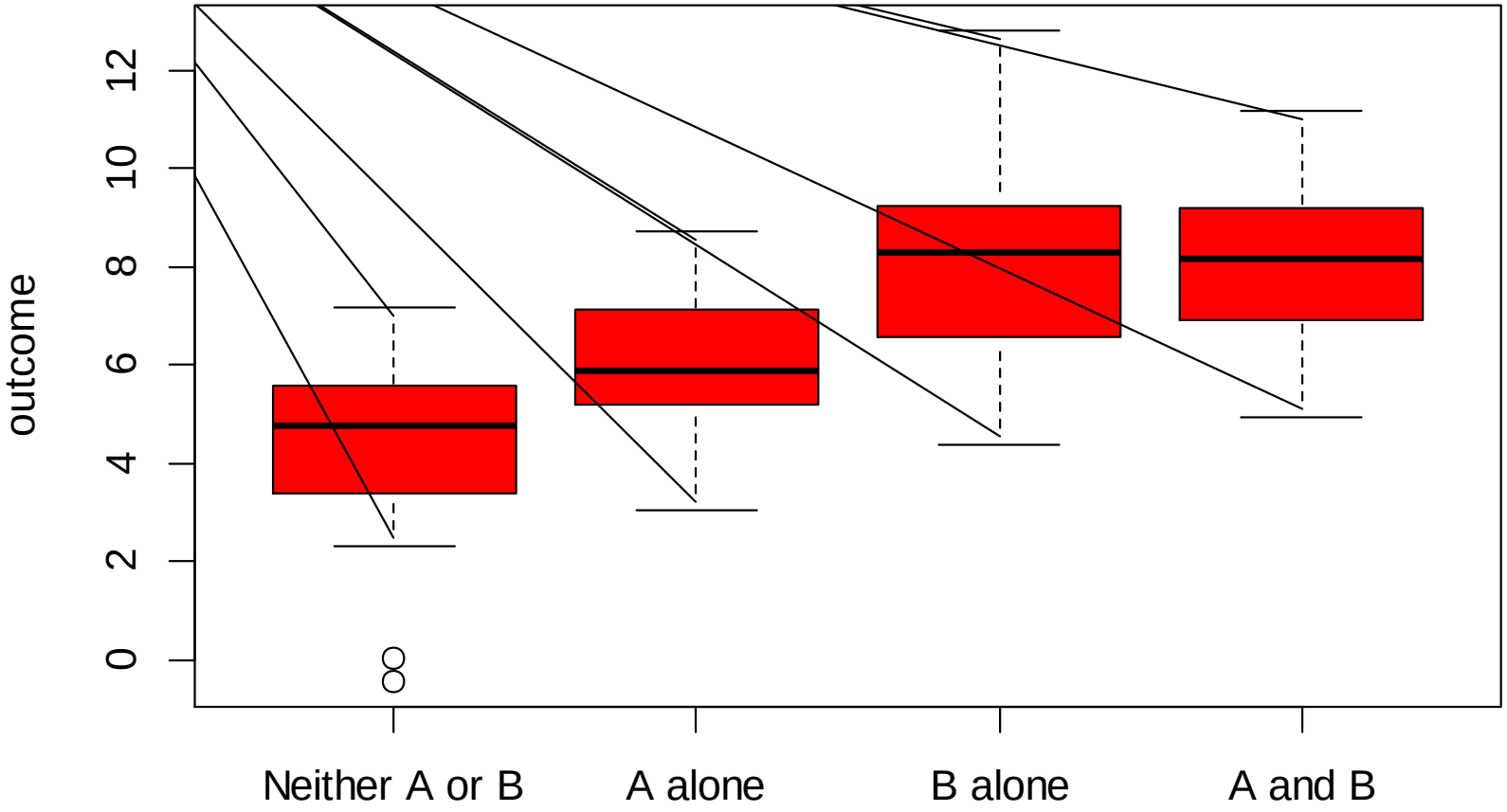
- Two covariables may not affect the outcome “independently” in an additive way but there is an interaction/effect modification
  - Example: Two drugs to lower blood pressure (A and B) can be given jointly or alone



Interactions

- Interactions can be explicitly modelled
- For the example above:
  - Instead of modelling separate effects of drug A and B, we create a categorical variable with 4 levels: “no drug”, “A alone”, “B alone”, “A and B”.
  - We can even test whether the coefficient for “A and B” is different from the sum of the effects of “A alone” and “B alone”

Example: Interactions



Example: Interactions (continued)

```
> summary(lm(outcome~trt.A+trt.B+trt.A:trt.B))
```

Main effect of treatment A.

Use ":" for interactions (between treatment A and B)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.3373     0.3640  11.915  < 2e-16 ***
trt.A          1.7271     0.5148   3.355  0.00114 **
trt.B          3.9935     0.5148   7.757 9.25e-12 ***
trt.A:trt.B   -1.9195     0.7281  -2.636  0.00977 **
```

Strong suggestion for an interaction
→ Effects of treatments A and B not additive.

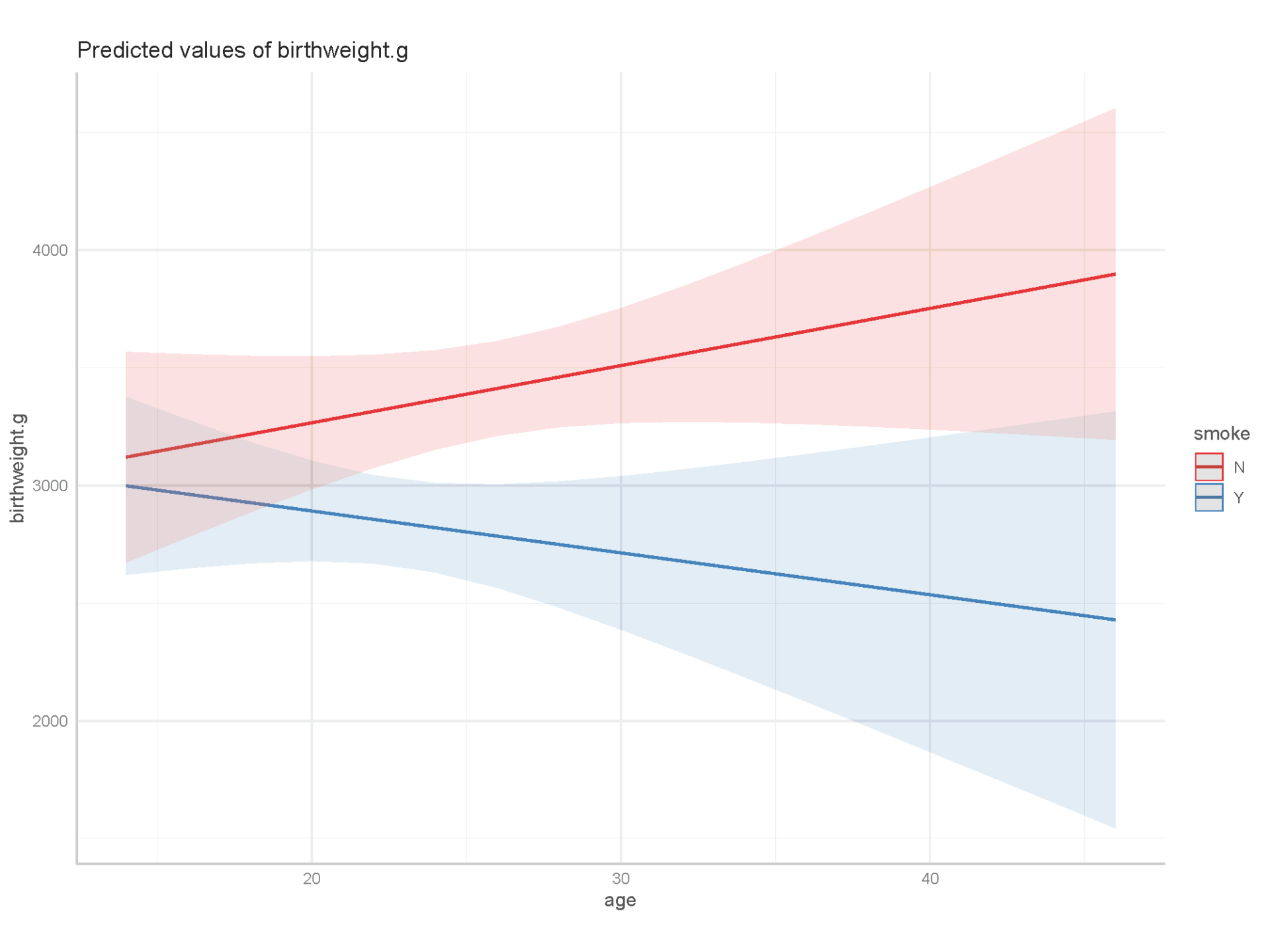
Note: R has a shorthand notation for model with interaction:

```
> lm(outcome~trt.A*trt.B)
```

Birthweight: interaction smoking-age

```
> fit.bwt.1 <- lm(birthweight.g~age*smoke+weight.m.kg,
                  data=birthweight.data)
> summary(fit.bwt.1)

                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2457.346    526.485   4.667 1.05e-05 ***
age               24.303     16.891   1.439   0.1536
smokeY           467.353    627.028   0.745   0.4580
weight.m.kg        5.371      5.295   1.014   0.3131
age:smokeY       -42.099     25.276  -1.666   0.0992 .
```


Predicted values of birthweight.g
4000
3000
2000
birthweight.g
20
30
40
age
smoke
N
Y

Birthweight: interaction smoking-age

```
> fit.bwt.1 <- lm(birthweight.g~I(age-23)*smoke+
                    weight.m.kg, data=birthweight.data)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3016.315    346.197   8.713 1.26e-13
I(age - 23)          24.303     16.891   1.439 0.153622
smokeY             -500.923    146.511  -3.419 0.000942
weight.m.kg           5.371      5.295   1.014 0.313130
I(age - 23):smokeY  -42.099     25.276  -1.666 0.099244
> anova(fit.bwt.1,fit.bwt.2)
Analysis of Variance Table

Model 1: birthweight.g ~ I(age - 23) * smoke + weight.m.kg
Model 2: birthweight.g ~ I(age - 23) + weight.m.kg
  Res.Df      RSS Df Sum of Sq   F    Pr(>F)
1     91 39758252
2     93 47360403 -2  -7602151 8.7 0.0003488
```

Distribution of residuals

- residuals from a linear regression model

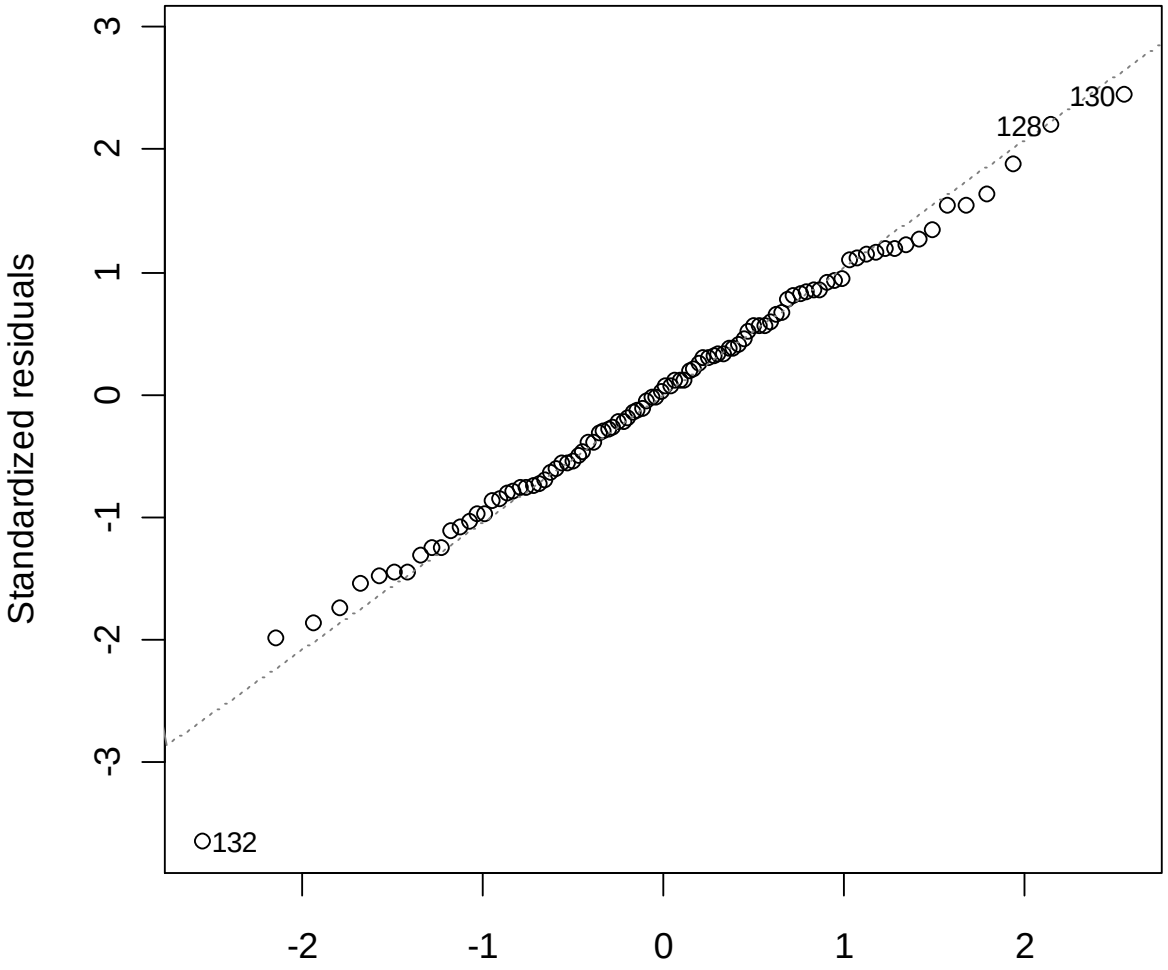
  (1) are normally distributed

  (2) have constant standard deviation σ (i.e. independent of the x's)

→ (1) can be assessed by the normal plot of residuals

→ (2) can be assessed by plotting fitted values versus residuals (Tukey-Anscombe plot)

Normal plot for the birthweight example

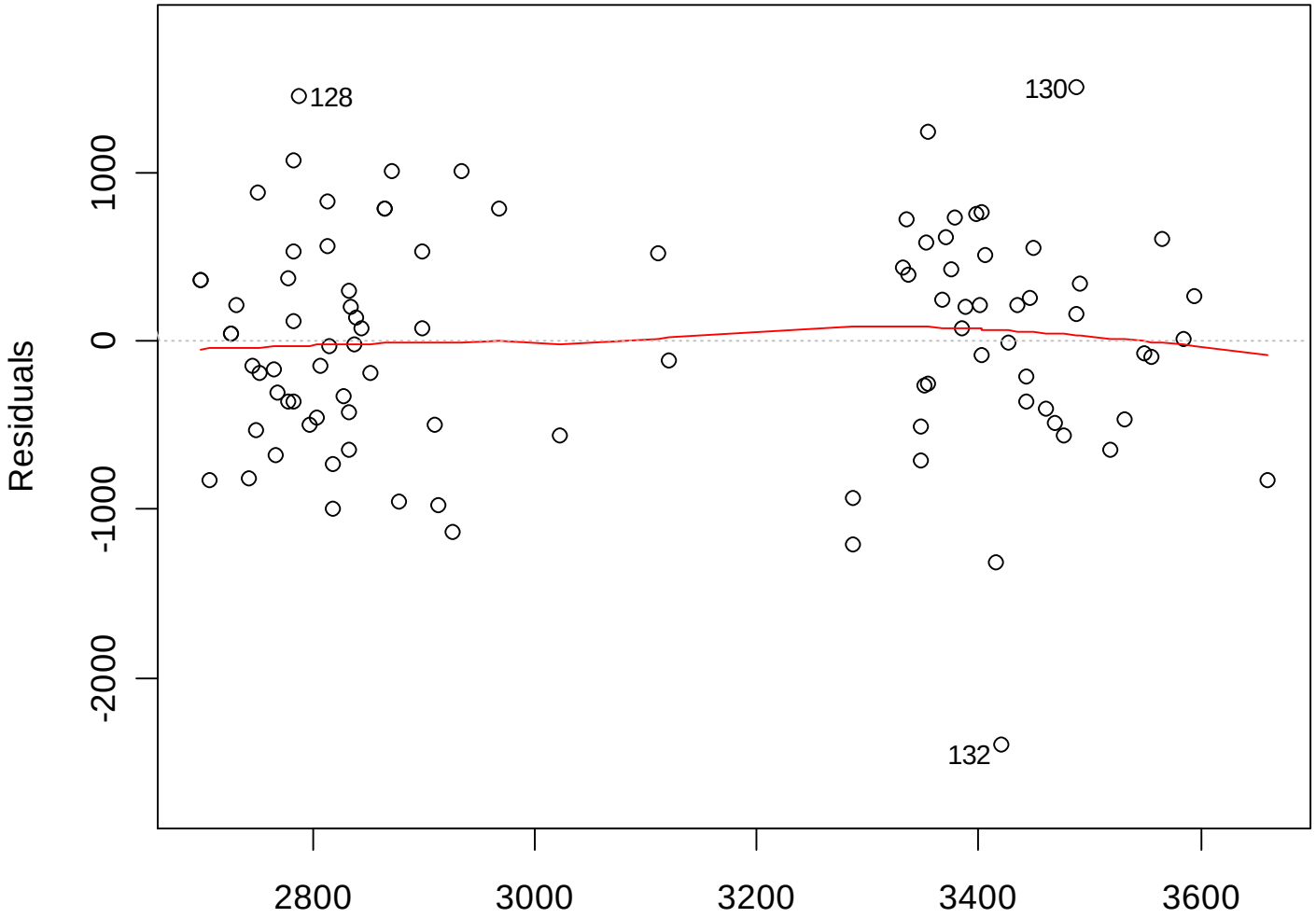


Tukey-Anscombe plot

- Plot fitted values versus residuals
  - Fitted value for observation i:
    $a+b_1 \cdot x_{2i}+b_2 \cdot x_{2i}+b_3 \cdot x_{3i}$
  - Residual for observation i: $y_i – (a+b_1 \cdot x_{2i}+b_2 \cdot x_{2i}+b_3 \cdot x_{3i})$

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

Tukey-Anscombe plot for the birthweight example



Selection of covariables in multiple linear regression

Why regression analysis?

- Inferential (90% of observational studies):
  - When we have little background knowledge
  - Model as simple as possible; “parsimony”
  - “X is associated with Y”
- Prediction:
  - Accurate individual prediction
  - “X is a predictor of Y”
- Explanation:
  - Causal; “X has an effect on Y”
  - RCT gold standard, but observational studies sometimes only possible option

Which covariables to include?

- Inferential:
  - Any possible risk factor of outcome
  - Often too many potential candidates
- Prediction:
  - Any variable that improves prediction, preferably easy to measure
  - Often too many potential candidates
- Explanation:
  - RCT: Important variables that affect outcome; improve precision
    - Note: they can never be confounders
  - Observational studies: confounders, i.e. common cause of both outcome and exposure

Limits on the number of predictors

- If sample size *n* small and number of parameters *p* large, there is a risk that the regression fit adapts too much to the examined dataset (but not the population)

→ Over-fitting:

Results will not replicate in future validation studies

- Empirical evidence and simulations show that to avoid over-fitting (ideally) : $p \leq n/10$
  - If sample size 100, one should not consider >10 parameters in model

→ Carefully select your covariables!

Stepwise backward model selection

- Start with the model with all covariables
- If covariable with largest p-value has p-value>0.15 (say), then remove it and refit the model (without this covariable) → model reduced by one covariable
- If covariable with largest p-value in the reduced model has p-value>0.15, then remove it and refit the model → model reduced by two covariables
- Repeat process until all covariables have p-value ≤0.15

Stepwise backward model selection for the birthweight example

1. Model with all covariables → remove age

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2896.687    460.009   6.297 1.02e-08 ***
age                  5.487     12.677   0.433 0.666138
weight.mother.kg     6.168      5.324   1.158 0.249670
smokeTRUE         -549.228    144.990  -3.788 0.000271 ***
```

2. Model without age → remove weight of mother

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3026.193    347.883   8.699 1.15e-13 ***
weight.mother.kg     6.379      5.279   1.208 0.229985
smokeTRUE         -565.565    139.378  -4.058 0.000103 ***
```

3. Model without age and weight of mother → final model

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3428.8      100.4  34.148  < 2e-16 ***
smokeTRUE      -601.9      136.4  -4.412 2.73e-05 ***
```

Problems with stepwise regression

- Statistical tests were not invented for model selection
- Which variables are selected in final model can be quite arbitrary (especially if covariables are highly correlated)
- In final model (after stepwise regression): estimates inflated, p-values too small and confidence intervals too narrow
- No strong conclusions possible. But in inferential studies it may serve as basis for replications of same study
- There are several more recent and advanced model selection methods, especially in prediction framework (this is an active area of current statistics research)

Model selection: do and don't

- DO
  - Use background knowledge before model building
    Think about causal pathways
  - Report all candidate covariables in the manuscript methods section, not just those selected in the final model
  - If possible: Validate final model in an independent data set

- DON'T
  - Select variables based on p-value in univariable analysis
    - Variable significant in univariable analysis might be non-significant in adjusted analysis
    - Variable non-significant in univariable analysis might be significant in adjusted analysis

MODELING STRATEGIES AND OUTLOOK

Modeling strategies

- Determine type of study (inferential, predictive, causal)
- Define covariables (and outcomes) a-priori
  - Chose carefully based on subject matter knowledge
  - Pre-define in protocol or statistical analysis plan
  - Limit number of covariables
- Explore univariable associations between covariables and outcome (simple linear regression) and full model with all covariables (multivariable linear regression)
- Explore transformations (covariables and response)
  - Log-transformation for laboratory data
  - Log transformation (or square-root for discrete/count data)

Modeling strategies (continued)

- Try to interpret the fitted model – does it make sense?
  - If results are implausible, this might be due to non-linear effects or interactions → explore these
- Always assess model assumptions
- No or very limited stepwise model selection
- Get the help of a statistician for the development of complex models.

Outlook: Other regression models

- Linear regression is only appropriate if the outcome is a continuous variable
  - For binary outcomes → logistic regression
  - For survival outcomes → Cox regression
- Understanding linear regression is the basis for understanding other regression models as many of the principles are the same.

Summary

- Multivariable (=multiple) linear regression
  - Model formulation, estimation, confidence intervals and tests
  - Interpretation of regression outputs
  - Model assumptions assessment

    (non-linearity, interaction, residuals diagnostic)
  - Model selection
  - General modeling strategies