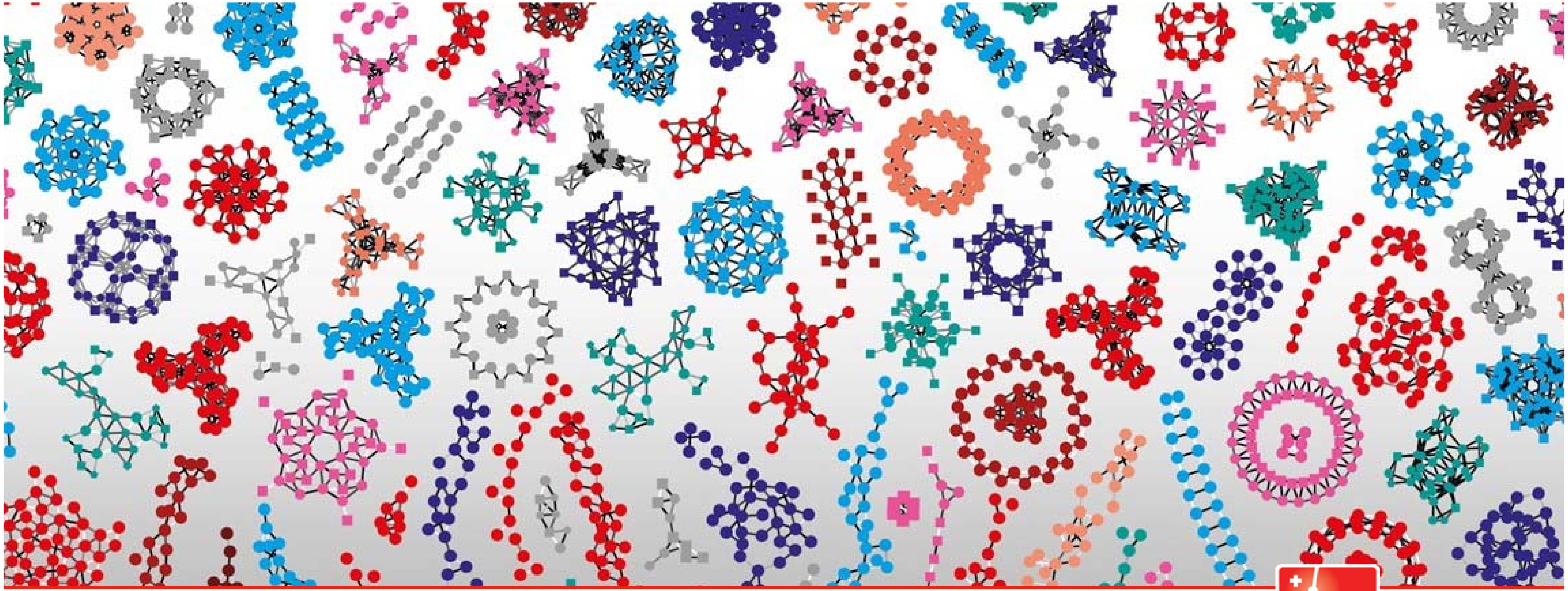


SIB
Swiss Institute of Bioinformatics

Regression

BCF

Isabelle Dupanloup

www.sib.swiss

birthwt {MASS} Risk Factors Associated with Low Infant Birth Weight

The birthwt data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

This data frame contains the following columns:

- low: indicator of birth weight less than 2.5 kg
- age: mother's age in years
- lwt: mother's weight in pounds at last menstrual period
- race: mother's race (1 = white, 2 = black, 3 = other)
- smoke: smoking status during pregnancy
- ptl: number of previous premature labours
- ht: history of hypertension
- ui: presence of uterine irritability
- ftv: number of physician visits during the first trimester
- bwt: birth weight in grams

Apply several linear regression models to the birthwt dataset

1. Inspect the dataset
2. Does the mother's age allow us to predict the birth weight ?
3. Check the assumptions of the model you used in 2.
4. Can you better predict birth weight using other variables in the dataset, in addition to mother’s age ?
5. Which set of variables in the dataset predict best birth weight ?

Apply several linear regression models to the birthwt dataset

1. Inspect the dataset
2. Apply a simple linear regression model to predict birth weight in grams using mother’s age
3. Check the assumptions of the linear regression model
4. Apply a multiple linear regression model to predict birth weight in grams using multiple predictors
5. Perform model selection

(Simple) Linear Regression

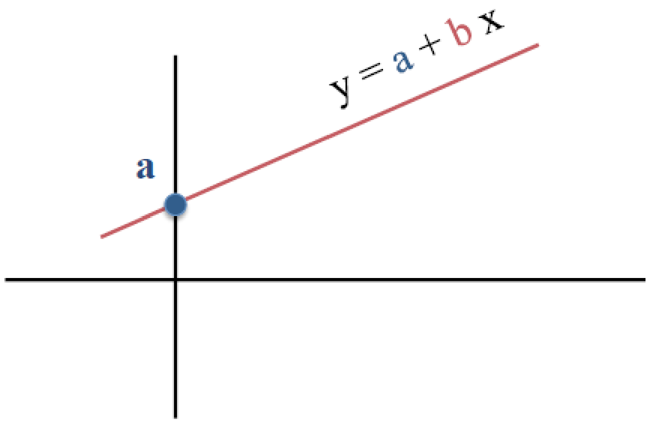
Simple linear regression refers to drawing a (particular, special) line through a scatterplot

It is used for 2 broad purposes: **explanation** and **prediction**.

The equation for a line to predict y knowing x (in slope- intercept form) looks like

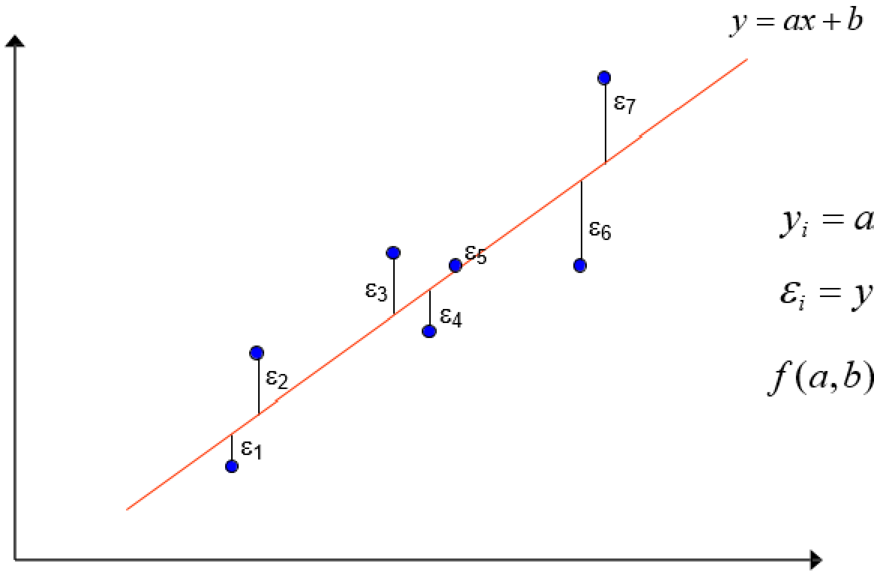
$$y = a + b\,x$$

where a is called the intercept and b is the slope.



Linear regression: least-squares fitting

Least-square fitting



Regression line such that:

$$\sum_i \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \ldots$$

minimum

$$y_i = ax_i + b + \varepsilon_i$$

$$\varepsilon_i = y_i - (ax_i + b)$$

$$f(a,b) = \sum_i \varepsilon_i^2 = \sum_i \left[y_i - (ax_i + b)\right]^2$$

$$\boxed{\begin{aligned} \partial f(a,b)/\partial a = 0 \\ \partial f(a,b)/\partial b = 0 \end{aligned}}$$

The least-squares procedure finds the straight line with the **smallest sum of squares of vertical errors**.

Linear regression: least-squares fitting

Formalization and extension of linear regression

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

$i = 1, \cdots, n$

Y_i : response (known)

β_0, β_1 : model parameters (estimated)

X_i : predictor (known)

ε_i : error term $\sim N(0, \sigma^2)$ (estimated)

Y represents **one** data point

Minimizing $\sum_i \varepsilon_i^2$ yields b_0 and b_1 estimators of β_0 and β_1

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \qquad b_0 = \overline{Y} - b_1 \overline{X}$$

(Simple) Linear Regression: interpretation of parameters

The regression line has two parameters: the slope and the intercept

The regression slope is the average change in Y when X increases by 1 unit

The intercept is the predicted value for Y when X = 0

If the slope = 0, then X does not help in predicting Y (linearly)

(Simple) Linear Regression: residuals

There is an error in making a regression prediction:

error = observed Y – predicted Y = y – (a + bX)

These errors are called residuals

The regression equation is calculated so that the sum (and mean) of the residuals is 0 (« in average, the model is correct »).

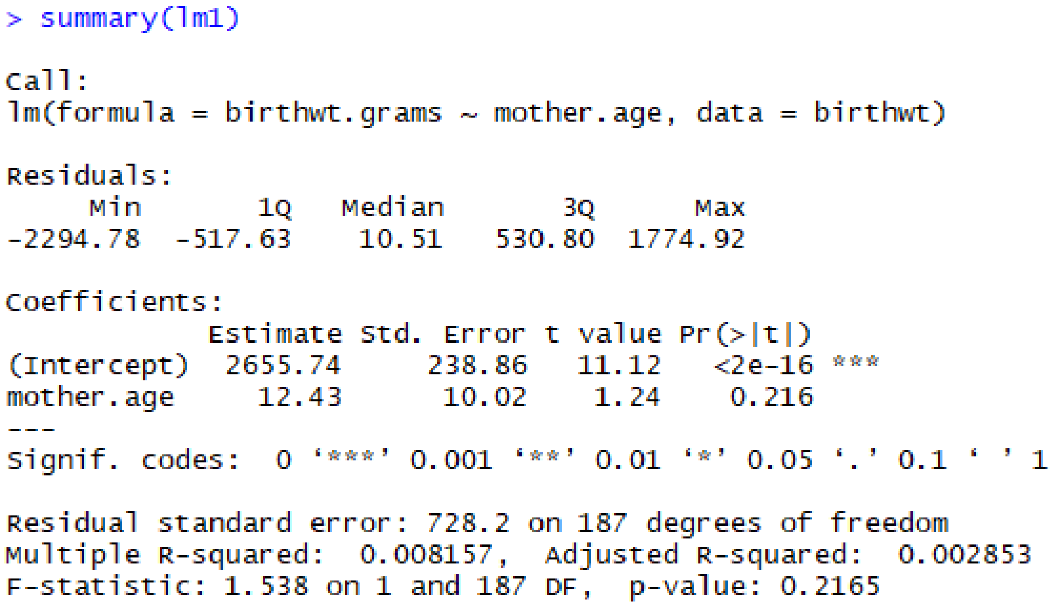
Ideally, we want the regression to include all the predictable variance, so that the distribution of the residuals is random and does not depend on X or on the predicted Y.

Apply several linear regression models to the birthwt dataset

These statistical tests tell us if the parameters are significantly different from 0. **It is not interesting for the intercept, but usually interesting for the slope.

Estimate and Std. Error are used for hypothesis testing
T-value = Estimate / Std. Error

This assumes that the residuals follow a normal distribution !



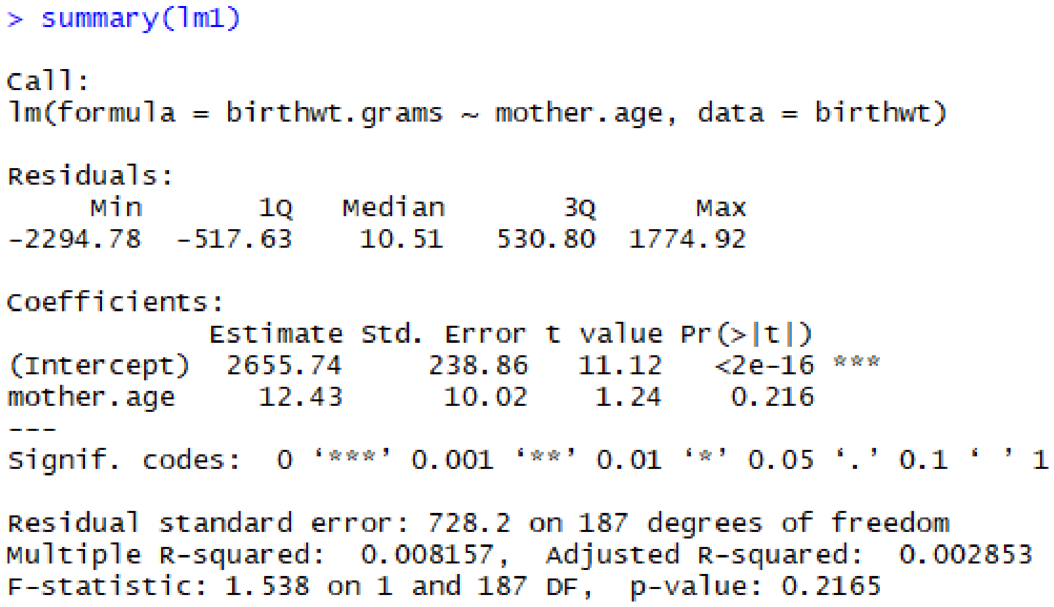
```
> summary(lm1)

Call:
lm(formula = birthwt.grams ~ mother.age, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max 
-2294.78  -517.63    10.51   530.80  1774.92 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2655.74     238.86   11.12   <2e-16 ***
mother.age     12.43      10.02    1.24    0.216    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 728.2 on 187 degrees of freedom
Multiple R-squared:  0.008157,  Adjusted R-squared:  0.002853 
F-statistic: 1.538 on 1 and 187 DF,  p-value: 0.2165
```

Apply several linear regression models to the birthwt dataset

number of DF =
total observations – number of parameters estimated

residual standard error = standard deviation of the residuals

R^2 = proportion of the total variance in the response data that is explained by the model



```
> summary(lm1)

Call:
lm(formula = birthwt.grams ~ mother.age, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max 
-2294.78  -517.63    10.51   530.80  1774.92 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2655.74     238.86   11.12   <2e-16 ***
mother.age     12.43      10.02    1.24    0.216    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 728.2 on 187 degrees of freedom
Multiple R-squared:  0.008157,	Adjusted R-squared:  0.002853 
F-statistic: 1.538 on 1 and 187 DF,  p-value: 0.2165
```

The F-statistic allows us to test if the whole regression (adding all variables *vs* having only the intercept in) is significant. With only one variable, it provides *exactly* the same result as the t-test for the significance of the coefficient of this variable.

Assumptions of a linear regression model

- Linearity of the data. The relationship between the predictor (x) and the outcome (y) is assumed to be linear.
- Normality of residuals. The residual errors are assumed to be normally distributed: $Ei \sim N(0,V)$
- Homogeneity of residuals variance. The residuals are assumed to have a constant variance (homoscedasticity): $V(Ei) = V$
- Independence of residuals error terms: Ei are independent from Xi and mutually independent

Model selection

- The regsubsets() function (from the leaps library) performs best subset selection by identifying the best model that contains a given number of predictors, where best is quantified using the residual sum of squares for each model. The syntax is the same as for lm(). The summary() command outputs the best set of variables for each model size.

- Validation set approach:
  1. split the observations into a training set and a test set
  2. apply regsubsets() to the training set in order to perform best subset selection
  3. compute the validation set error for the best model of each model size on the test set
  4. estimate the parameters of the best model on the full data set