SCIENCE MEETS LIFE

Introduction to Linear Regression

With a continuous or categorical predictor

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

21/03/2018

Veronique Storme

Aims

- Understand linear regression with one predictor
- Understand how we assess the fit of a regression model
  - Total sum of squares
  - Model sum of squares
  - Residual sum of squares
  - *F*
  - R^2
- Know how to do regression using **R/SAS**
- Interpret a regression model

OLS with continuous predictor

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

SCIENCE MEETS LIFE

What is Regression?

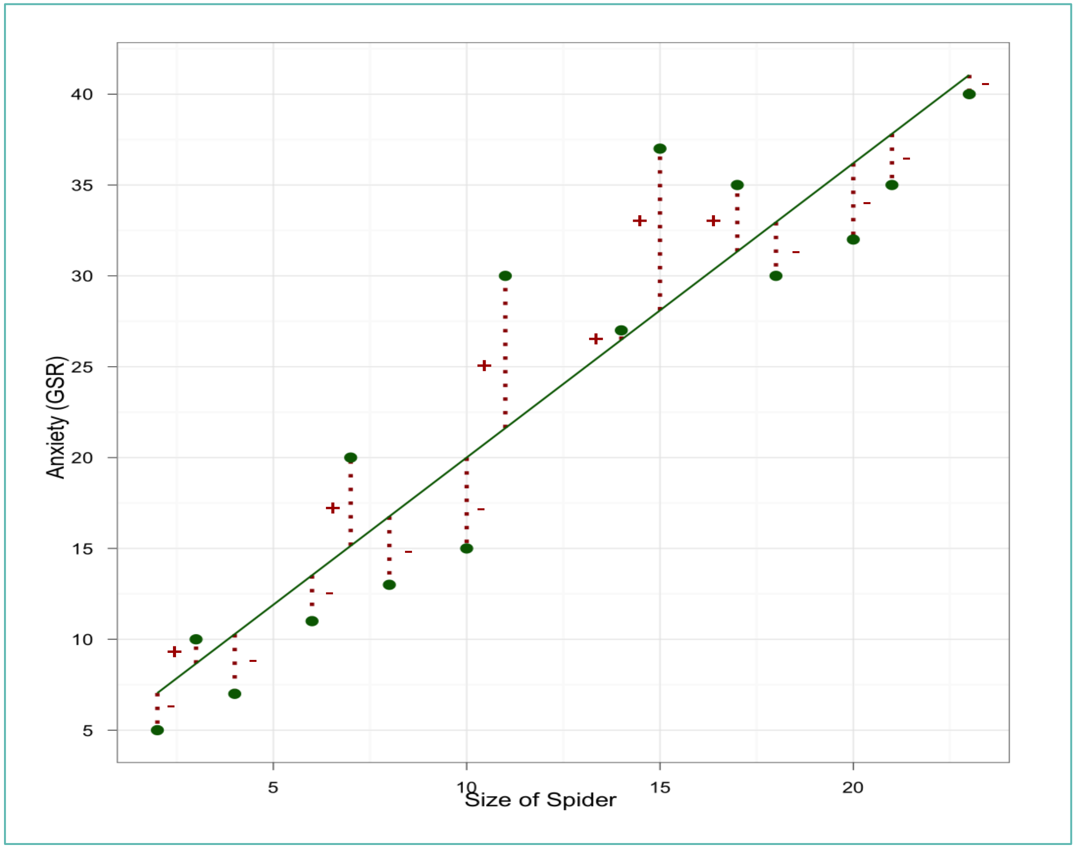
- A way of predicting the value of one variable from another.
  - It is a hypothetical model of the relationship between two variables.
  - The model used is a linear one.
  - Therefore, we describe the relationship using the equation of a straight line.

Describing a Straight Line

$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

- b_1
  - Regression coefficient for the predictor
  - Gradient (slope) of the regression line
  - Direction/strength of relationship
- b_0
  - Intercept (value of *Y* when *X* = 0)
  - Point at which the regression line crosses the *Y*-axis (ordinate)

The Method of Least Squares

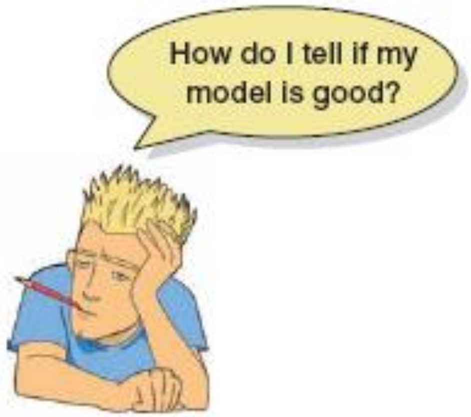


This graph shows a scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

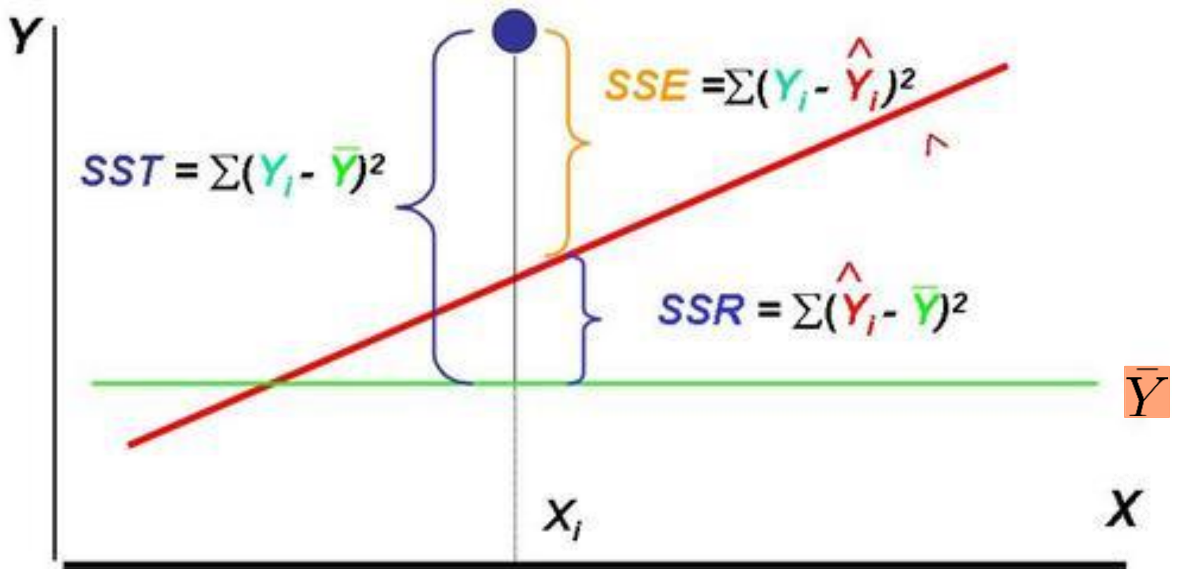
SCIENCE MEETS LIFE

How Good Is the Model?

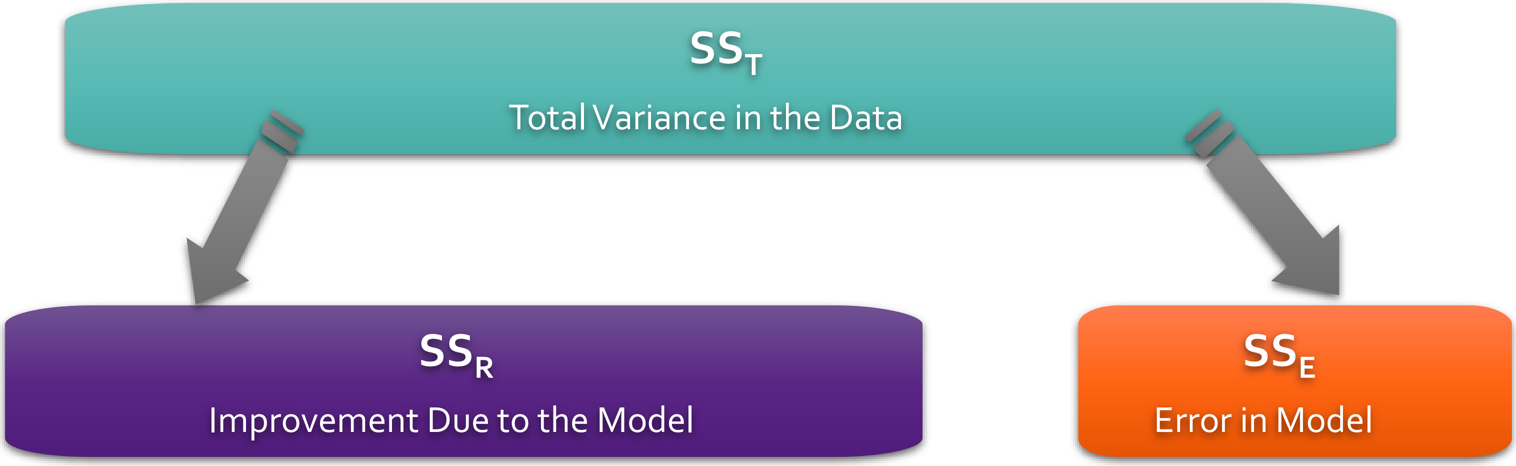
- The regression line is only a model based on the data.
- This model might not reflect reality.
  - We need some way of testing how well the model fits the observed data.
  - How?



Sums of Squares


Y
SST = $\sum(Y_i - \bar{Y})^2$
SSE = $\sum(Y_i - \hat{Y}_i)^2$
SSR = $\sum(\hat{Y}_i - \bar{Y})^2$
$\bar{Y}$
X_i
X

Testing the Model



- If the model results in better prediction than using the mean, then we expect SS_R to be much greater than SS_E

Testing the Model: *R*²

- *R*²
  - The proportion of variance accounted for by the regression model.
  - The Pearson Correlation Coefficient Squared

$$\boxed{R^2=SS_R/SS_T}$$

Testing the Model

- Mean squared error
  - Sums of squares are total values.
  - They can be expressed as averages.
  - These are called mean squares, MS.

$F=MS_R/MS_E$

with $MSR=SSR/1$

$MSE=SSE/(n-2)$

Assessing individual predictors

- Interpretation b_1
  - Change in average predicted outcome resulting from a unit change in the predictor
- Significance of b_1
  - H_0: b_1=0, tested with t-test
  - $t_{df=N-p-1} = \frac{b_{1observed}-b_{1expected}}{SE_{b1}} = \frac{b_{1observed}}{SE_{b1}}$

(p is the number of predictors in the model, thus p=1)

Graphics in R with ggplot2

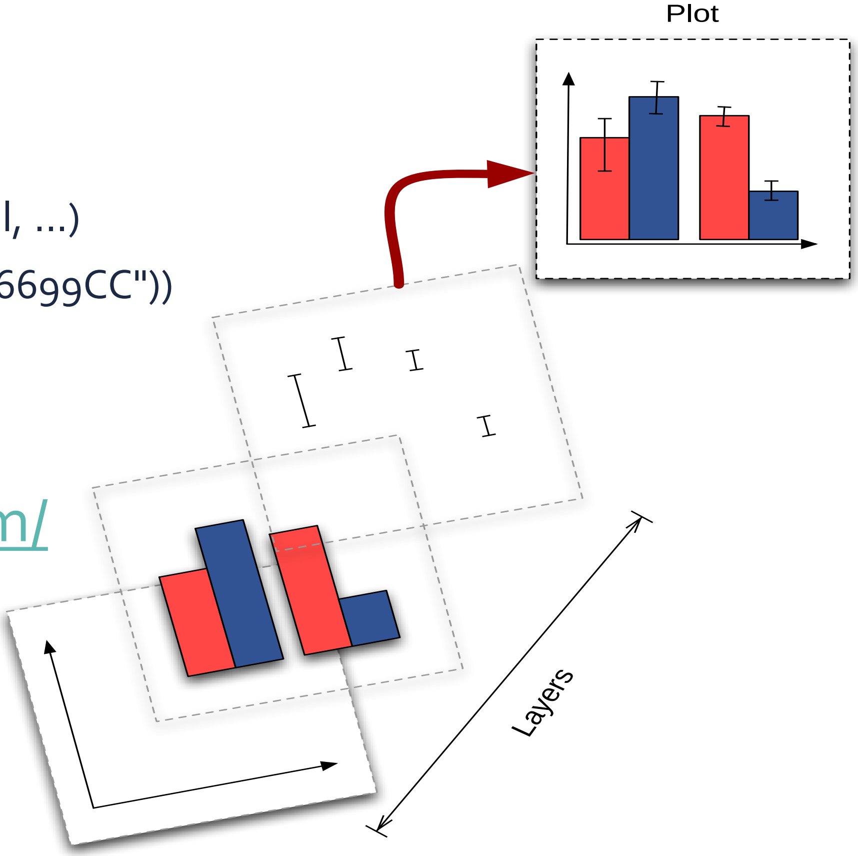
- In **ggplot2** a plot is made up of layers.

Eg:

```
bar <- ggplot(chickFlick, aes(x, y, fill = z))
bar + stat_summary(fun.y = mean, ...)
    + stat_summary(fun.data = mean_cl_normal, ...)
    + scale_fill_manual(values=c("#339966", "#6699CC"))
    + labs(...)
```

- More information on:

https://www.r-graph-gallery.com/

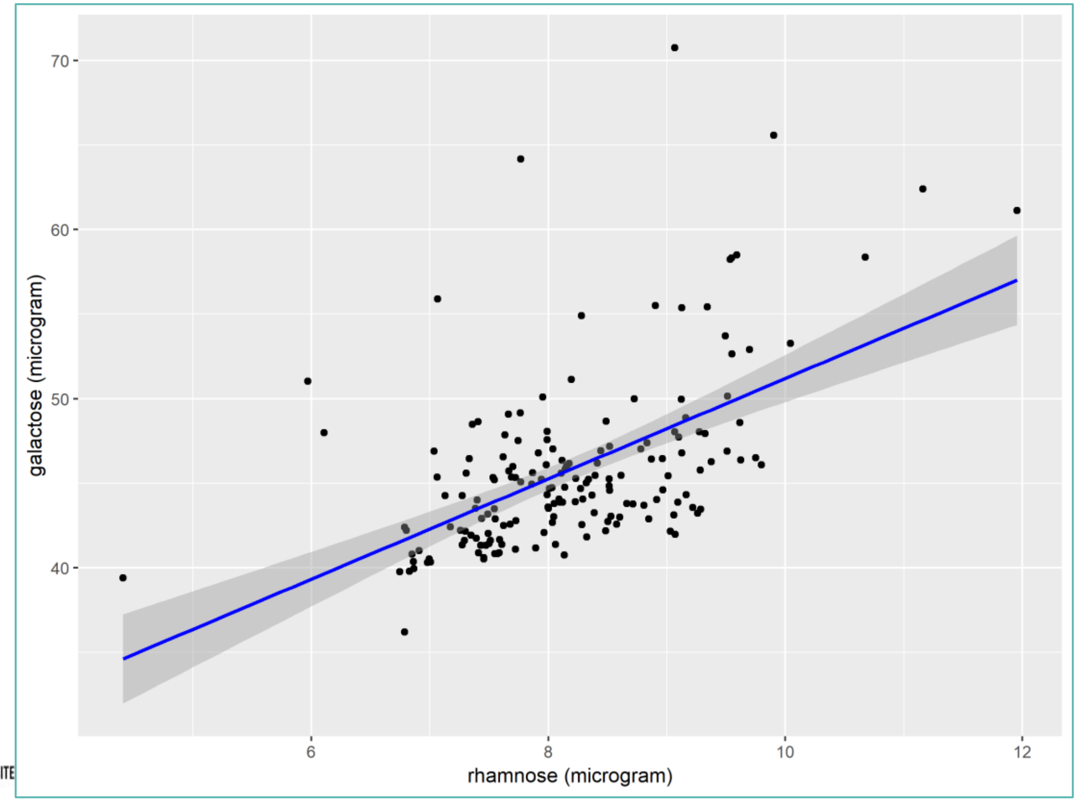


VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

example

- Investigate the relationship between rhamnose and galactose in Arabidopsis lignin mutants

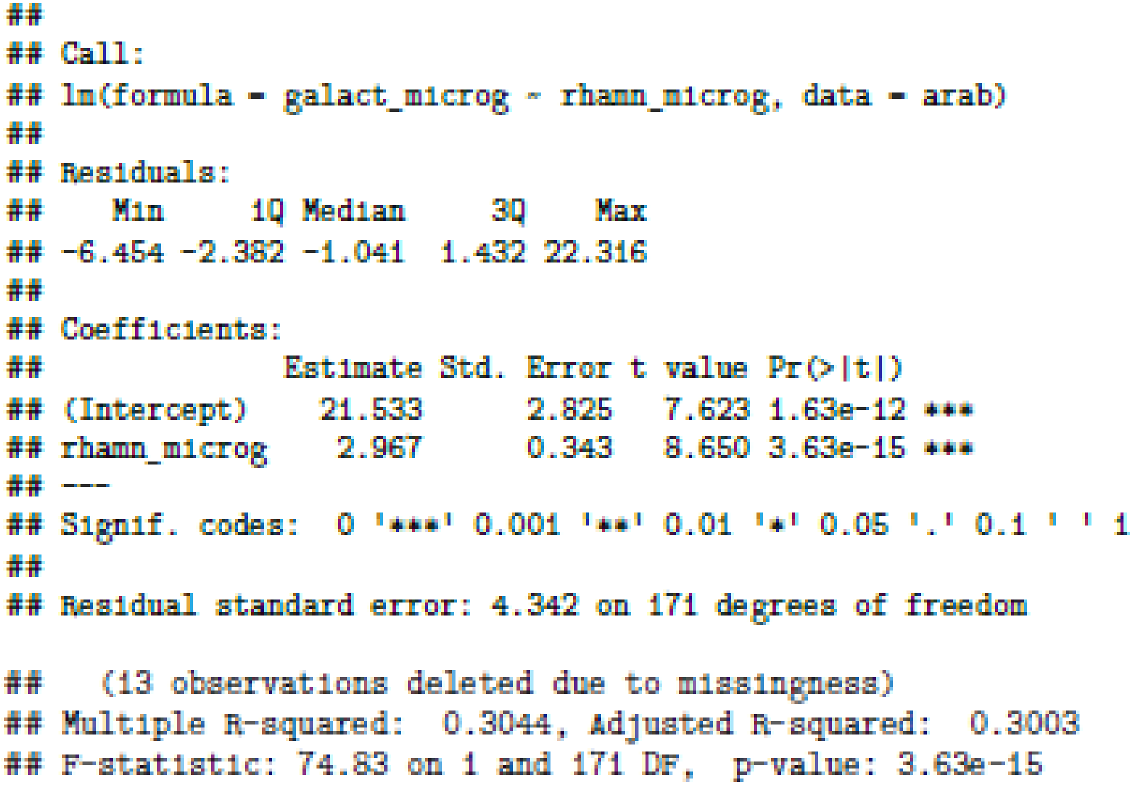


Example (continued)

- **lm**(outcome ~ predictor, data=dsn)

- lm.fit <- **lm**(galact_microg ~ rhamn_microg, data = arab)
- summary(lm.fit)

Output of a Simple Regression



```
##
## Call:
## lm(formula = galact_microg ~ rhamn_microg, data = arab)
##
## Residuals:
##    Min      1Q Median     3Q    Max
## -6.454 -2.382 -1.041  1.432 22.316
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    21.533      2.825   7.623 1.63e-12 ***
## rhamn_microg    2.967      0.343   8.650 3.63e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.342 on 171 degrees of freedom

##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.3044, Adjusted R-squared:  0.3003
## F-statistic: 74.83 on 1 and 171 DF,  p-value: 3.63e-15
```

Using the Model

$$
\begin{aligned}
galactose_i &= b_0 + b_1 rhamnose_i \\
&= 21.53 + (2.97 * rhamnose_i) \\
&= 21.53 + (2.97 * 1) \\
&= 24.5
\end{aligned}
$$

Checking Assumptions

- Variable type:
  - Pred cont or cat with non-zero variance
  - Outcome var: continuous or interval
- Linearity:
  - Linear in the parameters
- Normally distributed errors
- Homoscedasticity:
  - At each level of the predictor(s), the variance of the residuals should be the same
- Independent errors

Fitted Values and Residuals

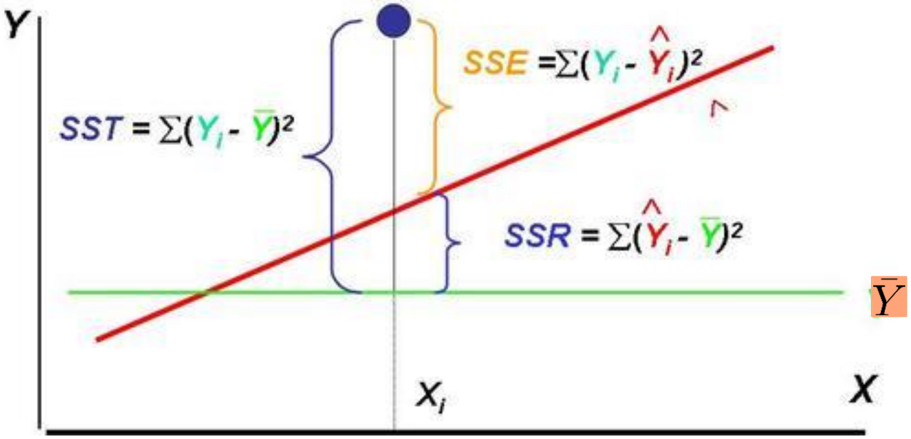
- Fitted values are the estimates of Y as determined by the regression equation.
- Residuals are the differences between each observed value and the corresponding fitted value.

$$y_i = b_0 + b_1 x_i + e_i$$

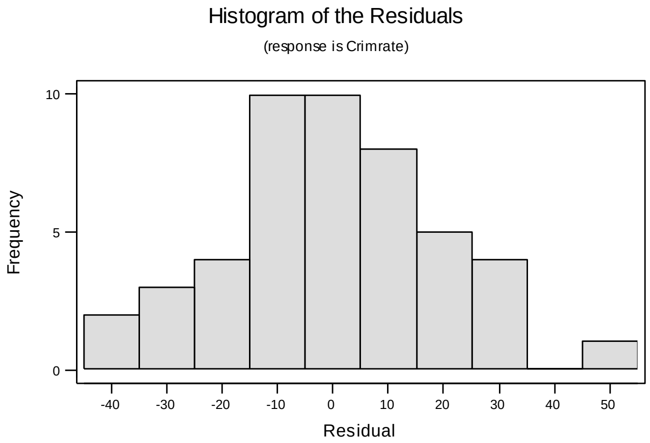
$$\hat{y}_i = b_0 + b_1 x_i$$

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

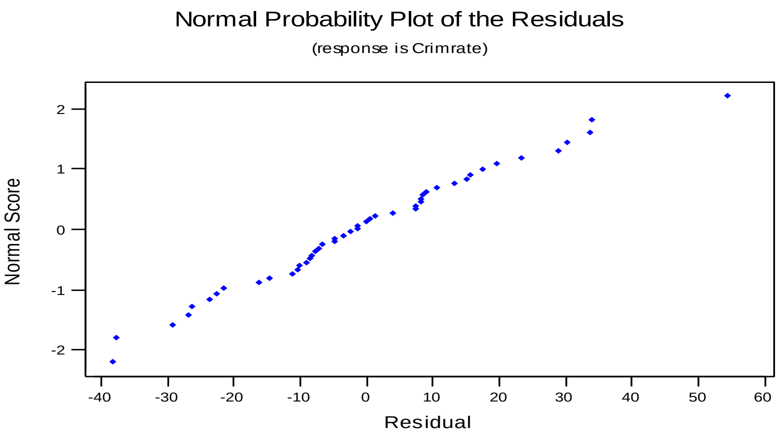
$$y_i - \hat{y}_i = e_i$$



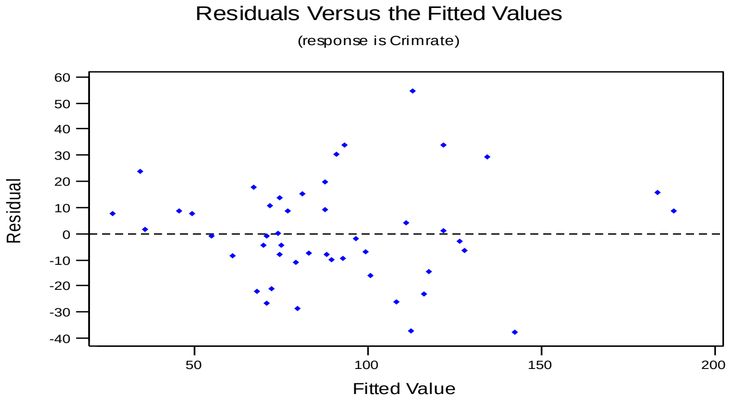
Residual Plots


Histogram of the Residuals

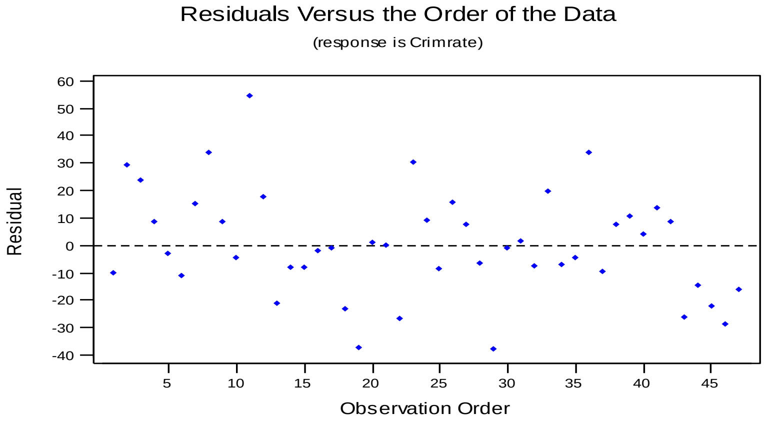
(response is Crimrate)


Normal Probability Plot of the Residuals

(response is Crimrate)


Residuals Versus the Fitted Values

(response is Crimrate)


Residuals Versus the Order of the Data

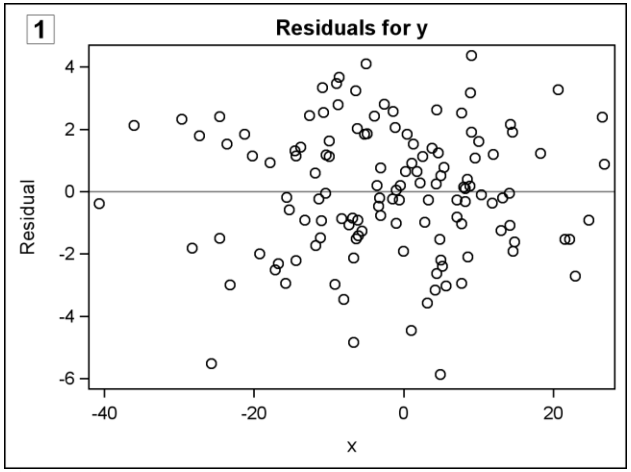
(response is Crimrate)

Studentized Residuals

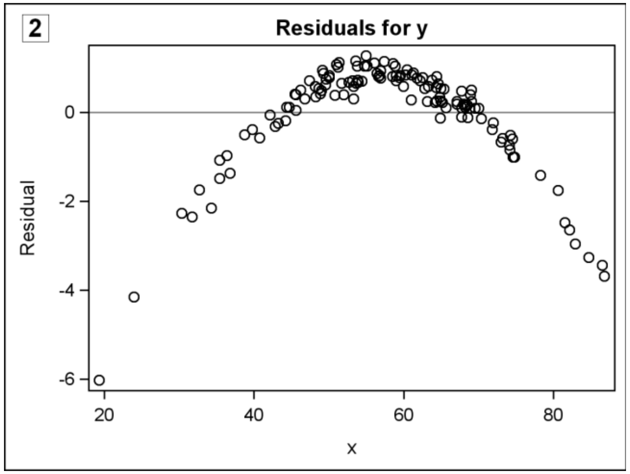
- Disadvantage raw residuals:
  - In same unit as observation (what is small/large?)
- Studentized residual:
  - Residual divided by the estimated standard deviation of the residuals (95% within -2 and +2)
- Suggested cutoffs are as follows:
  - |SR| > 2 for data sets with a relatively small number of observations
  - |SR| > 3 for data sets with a relatively large number of observations

Examining Residual Plots

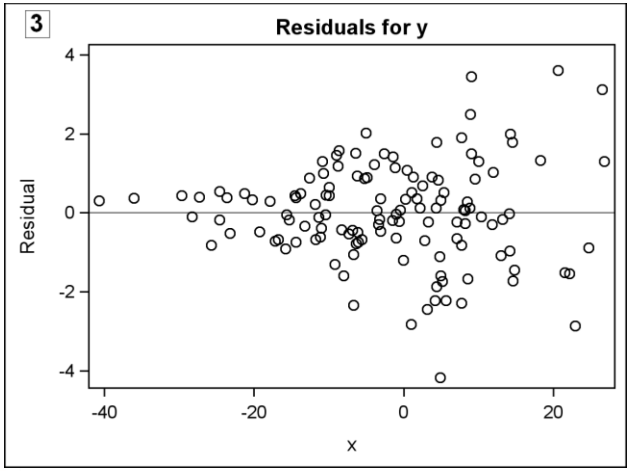
1



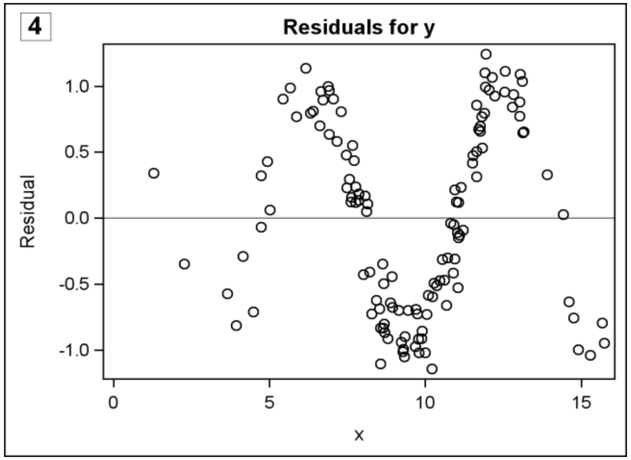
2



3

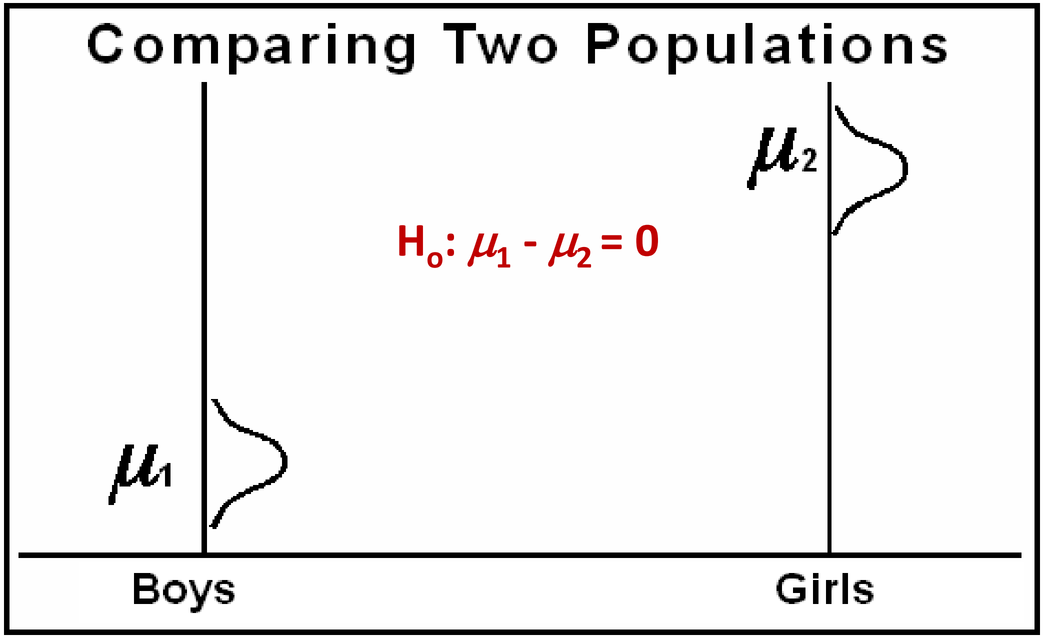


4



OLS with a binary predictor

Comparing Two Population Means: T-test



- Statistical Assumptions:
  - independent observations
  - normally distributed population means
  - equal population variances (Folded *F* Test)

The *T*-test as a GLM

$$\boxed{Y_i = b_0 + b_1 Group_i + \epsilon_i}$$

Dummy coding

- Provides a way of using categorical predictors in linear regression
- Uses zeros and ones to convey group membership
- For k groups, we can create k-1 dummies
- For 2 groups, only one dummy variable
    - X=1 when an observation belongs to group 2 and 0 otherwise

Design matrix

- Assume treatment: control and treated

$b_0*1 \quad + \quad b_1*\text{treatmenttreated}$

obs	sample	treatment
1	C1	control
2	C2	control
3	C3	control
4	T1	treated
5	T2	treated
6	T3	treated

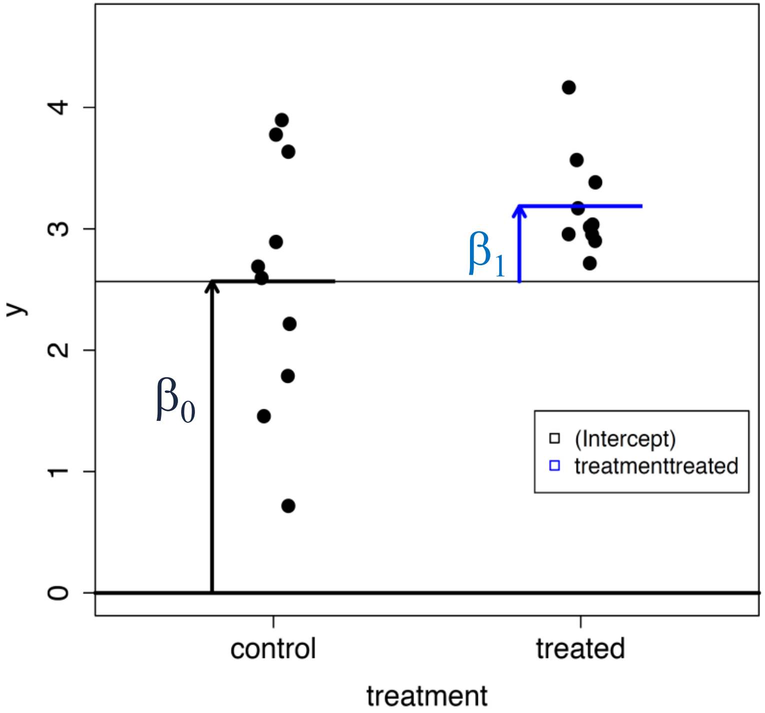
obs	Intercept	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Design matrix

$$E[Y|X = control] = \beta_0$$

$$E[Y|X = treated] = \beta_0 + \beta_1$$

$$H_0 : E[Y|X = treated] - E[Y|X = control] = \beta_1 = 0$$



OLS with a categorical predictor

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

SCIENCE MEETS LIFE

ANOVA as Regression

Consider a control group, a low dose treatment group and a high dose treatment group and some outcome.

The regression model is:

$$outcome_i = b_0 + b_1 X_1 + b_2 X_2 + e_i$$

With X_1 = 1 if observation belongs to the Low Dose group and 0 otw

X_2 = 1 if observation belongs to the High Dose group and 0 otw

Example with 3 treatment levels

observation	Intercept	Low Dose	High dose
Control	1	0	0
Control	1	0	0
Control	1	0	0
Low Dose	1	1	0
Low Dose	1	1	0
Low Dose	1	1	0
High Dose	1	0	1
High Dose	1	0	1
High Dose	1	0	1

Conditional models

- Control group
  - X_1=X_2=0

$$E[outcome|Control] = b_0 = \overline{x_{Control}}$$

- Low dose group
  - X1=1

$$E[outcome|LowDose] = b_0 + b_1 = \overline{x_{LowDose}}$$

$$b_1 = \overline{x_{LowDose}} - \overline{x_{Control}}$$

- High dose group
  - X2=1

$$E[outcome|HighDose] = b_0 + b_2 = \overline{x_{HighDose}}$$

$$b_2 = \overline{x_{HighDose}} - \overline{x_{Control}}$$

Performing Simple Linear Regression in R and SAS

R code regression.R and SAS code regression.sas illustrate the concepts discussed previously.