Removal of confounding factors

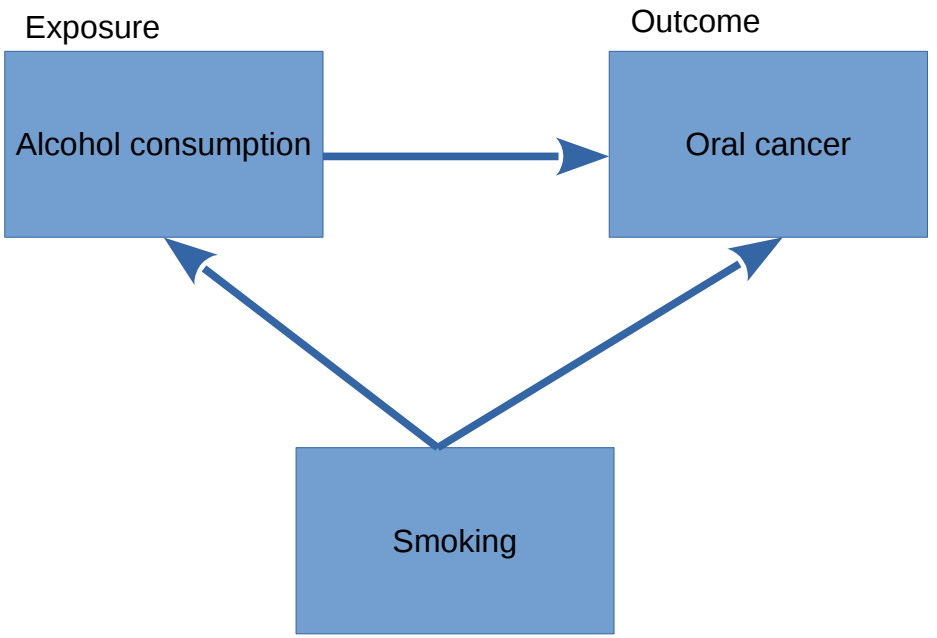
Bishwa Ghimire

Single cell RNA-seq data analysis with R
CSC, Espoo
27.05.2019

Confounding factors/confounders

- A third variable correlated with both the dependent variable and the independent variable.
- Distortion of true effect of exposure on a disease by third factor/variable.
- The factor can cause over/under estimation of true effect. In other words it biases our study.
- Example:
  - Lets say alcoholism has true effect on oral cancer.
  - There is different level of effect on oral cancer for people who smoke and who don't.
  - People who smoke are likely to drink alcohol.

Confounding factors



Confounder: distorts the relationship between alcohol consumption and oral cancer.

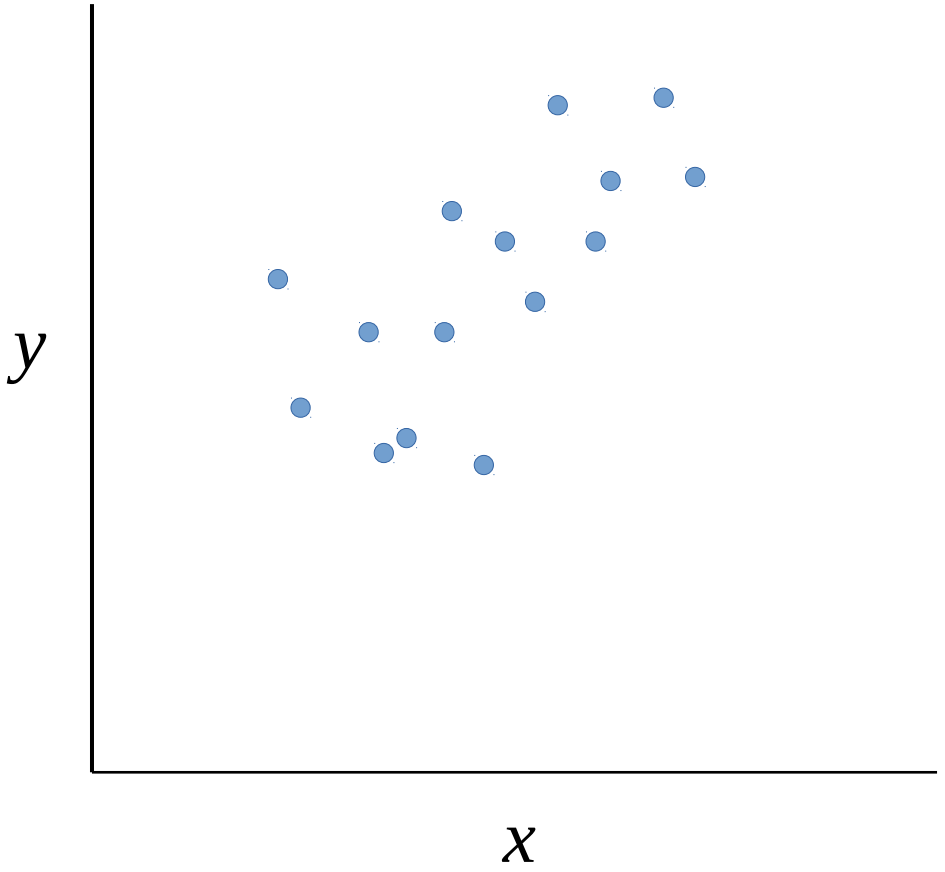
Adjusting confounding factors

- Experimental design
  - **Randomization** : Randomly assigning experimental units to treatment groups. It balances known and unknown confounders. Limited applicabality.
  - **Restriction** : Using only one category for confoundings. Eg. Taking only smokers. Study results can not be generalized.
  - **Matching** : Making a match of a experimental unit for different treatment groups.

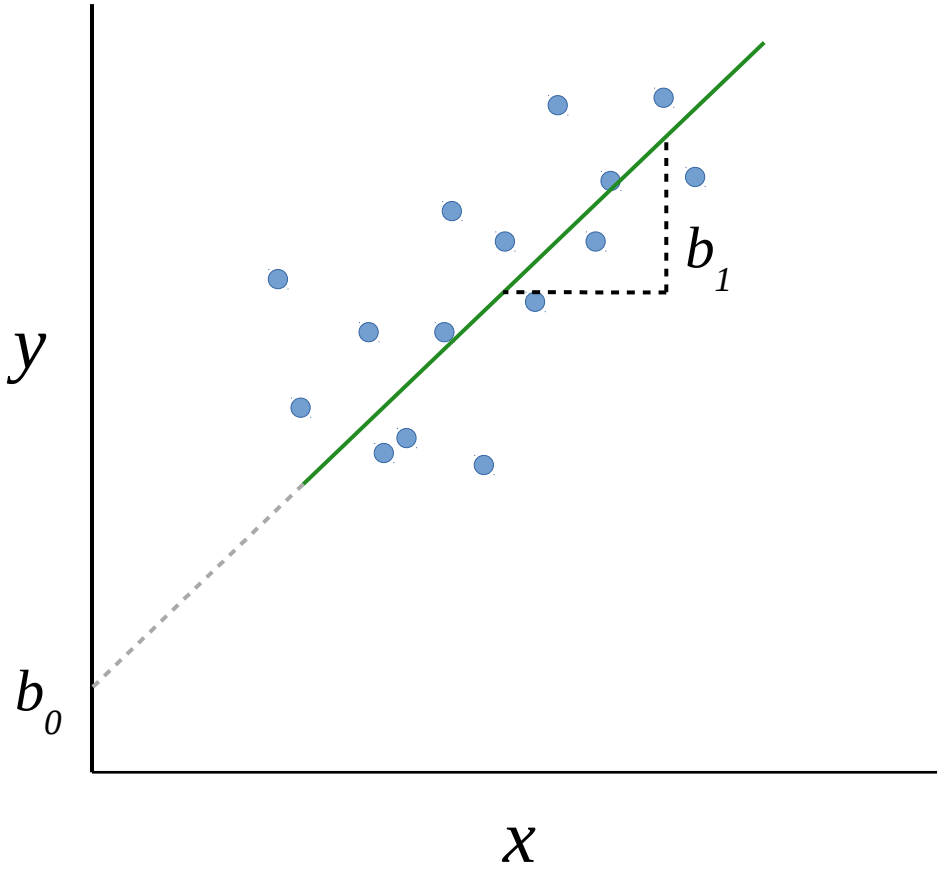
Adjusting confounding factors

- Data analysis
  - Multiple linear regression
  - Logistic regression
  - Analysis of Covariance

Simple linear regression



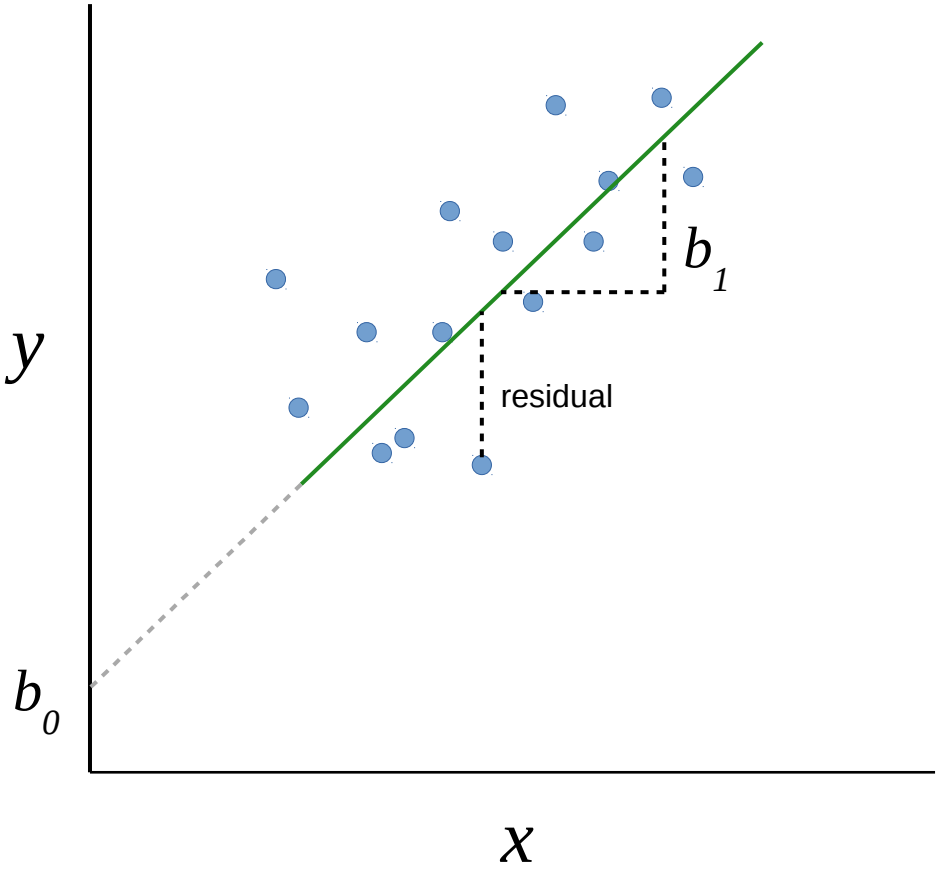
Simple linear regression



Equation of the line

$$y = b_0 + b_1 x$$

Simple linear regression



Equation of the line

$$y = b_0 + b_1 x$$

Constant

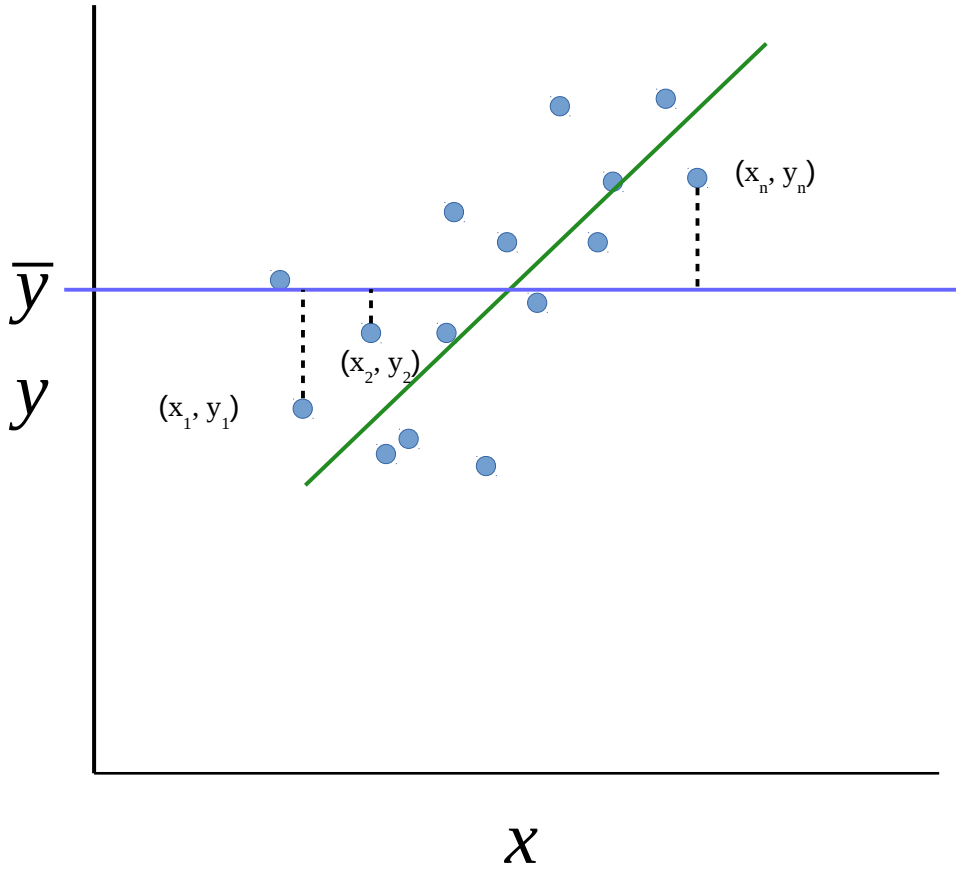
Coefficient

$$b_1 = r S_y / S_x$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Simple linear regression

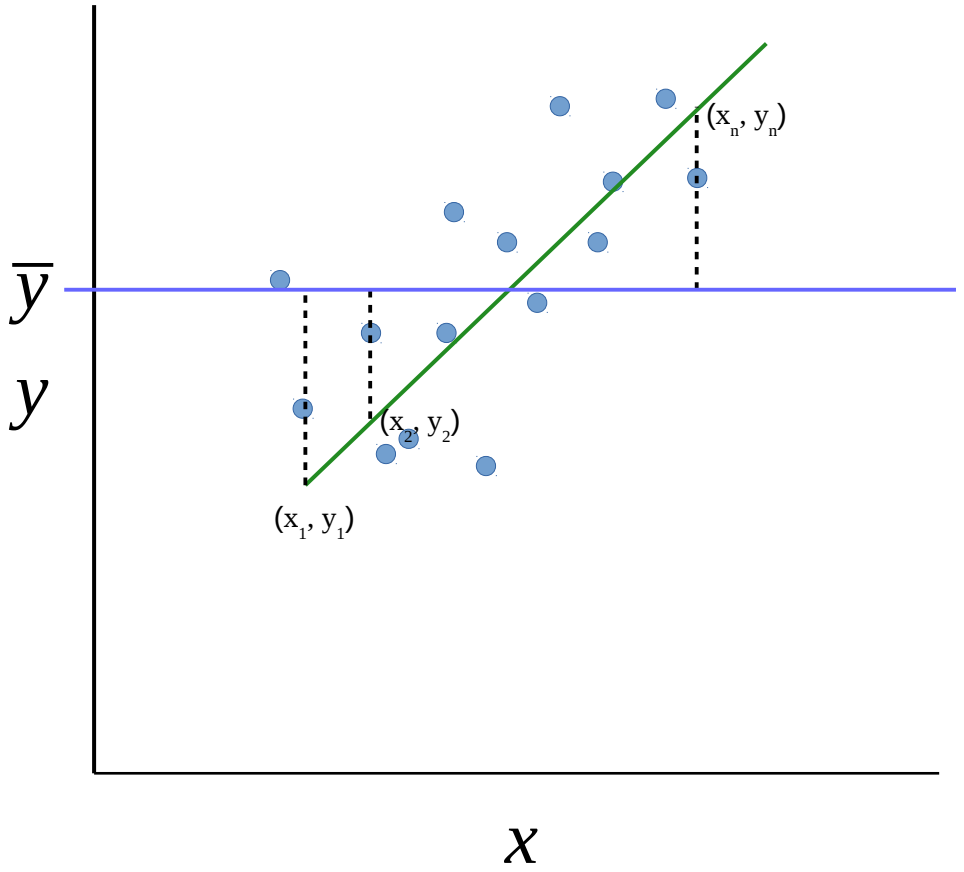
- R-squared (coefficient of determination)



$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

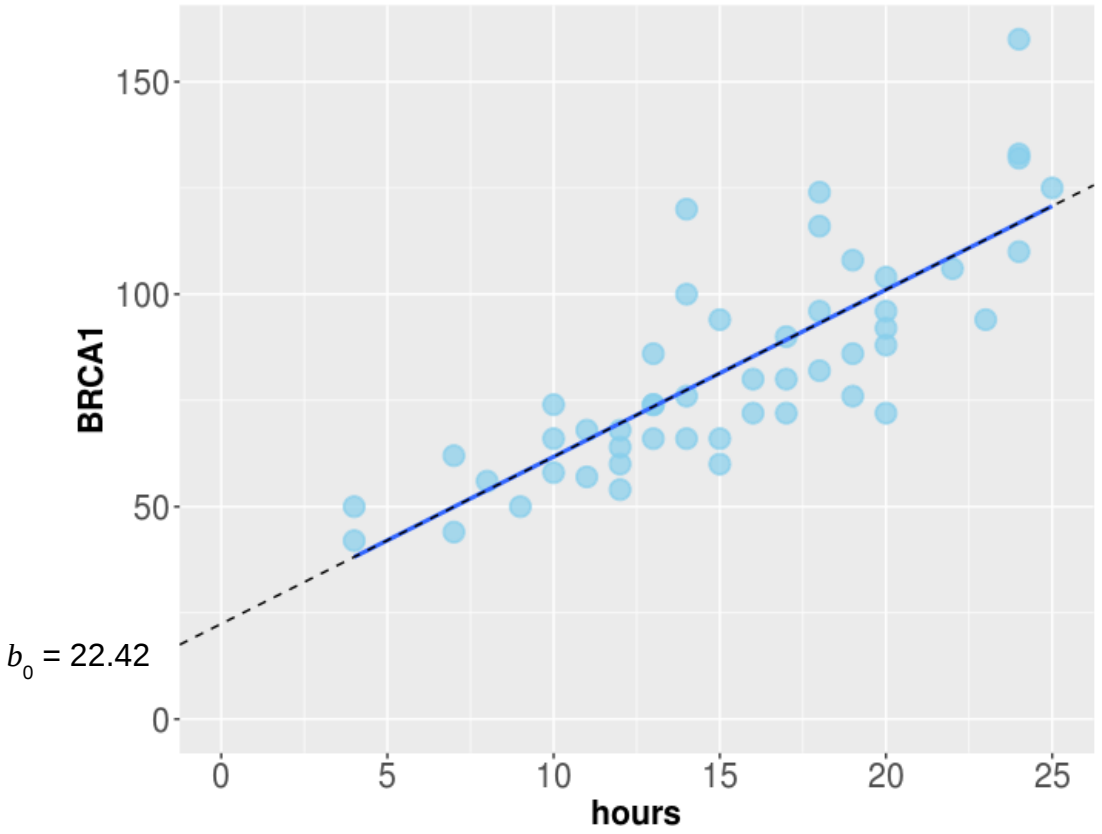
Total variation in $y = (y_1-\bar{y})^2 + (y_2-\bar{y})^2 + \ldots + (y_n-\bar{y})^2$

Simple linear regression



Variation explained by the model = $(y_1-\bar{y})^2 + (y_2-\bar{y})^2 + \ldots + (y_n+\bar{y})^2$

Simple linear regression



	BRCA1	PAX3	hours
cell-1	42	52	4
cell-2	50	43	4
cell-3	44	92	7
cell-4	62	36	7
cell-5	56	31	8
cell-6	50	21	9

$$y = b_0 + b_1 x$$

$$y = 22.42 + 3.93\, x$$

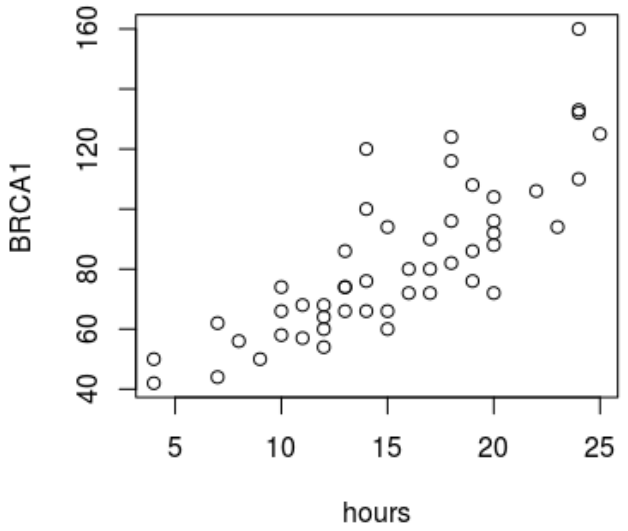
$$BRCA1 = 22.42 + 3.92\ hours$$

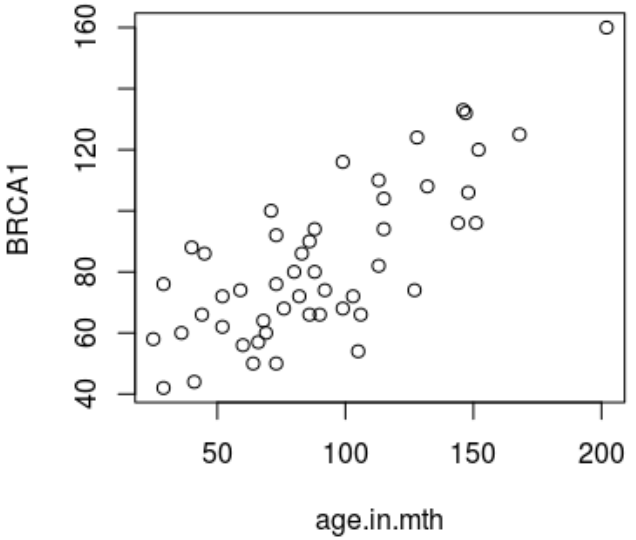
Multiple linear regression

- Extension of simple linear regression
- Association of two or more independent variables to one dependent variable.

$$y = b_0 + b_1 x_1 + b_2 x_2$$

	BRCA1	PAX3	hours	age.in.mth
cell-1	42	52	4	29
cell-2	50	43	4	73
cell-3	44	92	7	41
cell-4	62	36	7	52
cell-5	56	31	8	60
cell-6	50	21	9	64





Multiple linear regression

- Including only *hours* in a model

$$y = b_0 + b_1 x$$
$$y = 22.42 + 3.93 x$$
$$BRCA1 = 22.42 + 3.39\ hours$$

$$\frac{3.39 - 2.64}{2.64} = 0.284 = 28.4\%$$

Regression coeff. decreased by 28.4%

- Including both *hours* and *age.in.mth* in a model

$$y = b_0 + b_1 x_1 + b_2 x_2$$
$$y = 16.19 + 2.64 x_1 + 0.28 x_2$$
$$BRCA1 = 16.19 + 2.64\ hours + 0.23\ age.in.mth$$

- Effect of x_2 is regressed out by including it in the model

Regressing out in Seurat

- ScaleData(seuratObject,
  vars.to.regress = c("nCount_RNA",
  "nFeature_RNA"))
- ScaleData() calls RegressOutResid() which returns residuals of a regression model
- Returned residuals are later scaled
- *vars.to.regress* should be in *seuratObject@meta.data*

	orig.ident	nCount_RNA	nFeature_RNA
Prog_013	Prog	2563089	10211
Prog_019	Prog	3030620	9991
Prog_031	Prog	1293487	10192
Prog_037	Prog	1357987	9599

Other tools

- ScLVM (beta pre-release)
  - Designed for cell-cycle variation correction.
  - Also correction of other confounding variables.
- ccRemover (stable version from CRAN).
  - “ccRemover outperforms scLVM slightly.”
- scTransform
  - Part of Seurat

Thank you!