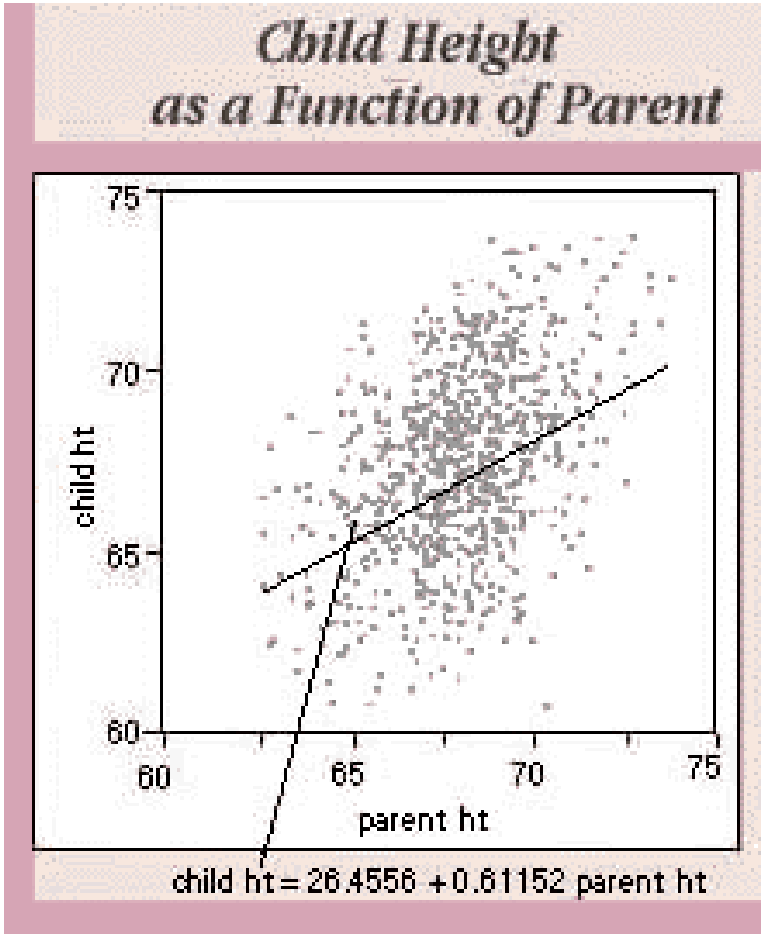
Regression and correlation

Timothy Mak
Postdoctoral Fellow
Centre for Genomic Sciences

Topics

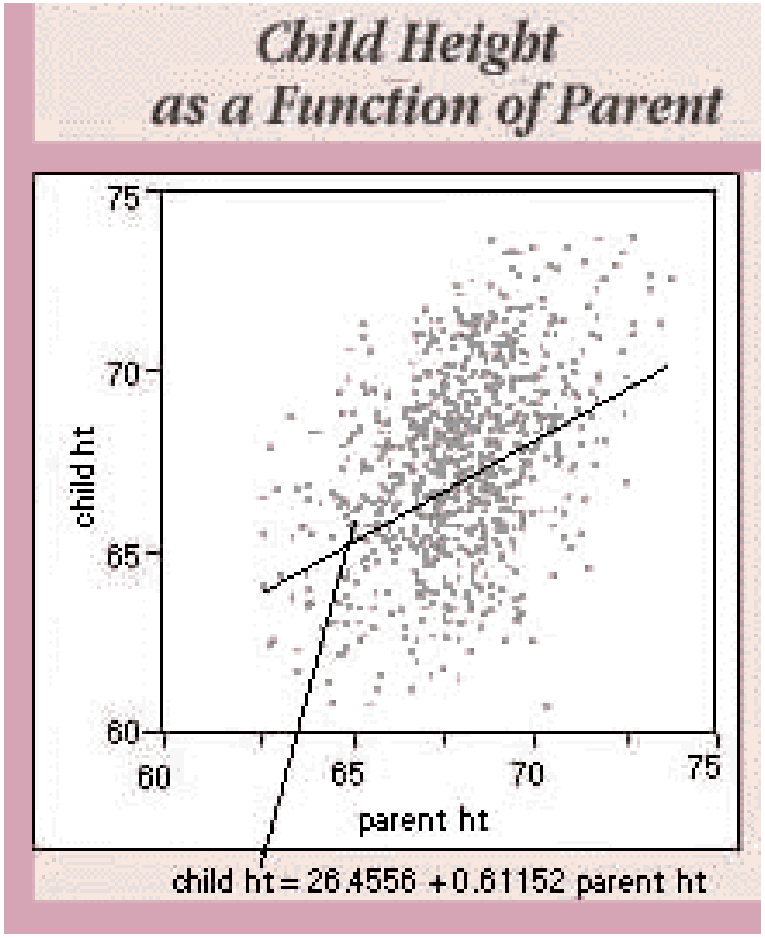
- Linear regression
- Correlation
- Multiple regression
- Categorical covariates
- Interactions
- Logistic regression/Generalized linear model
- Applications in genetics (GWAS)

Regression to the mean (history)



http://www.biostat.jhsph.edu/courses/bio653/misc/Regr_FigB2.gif

Linear regression



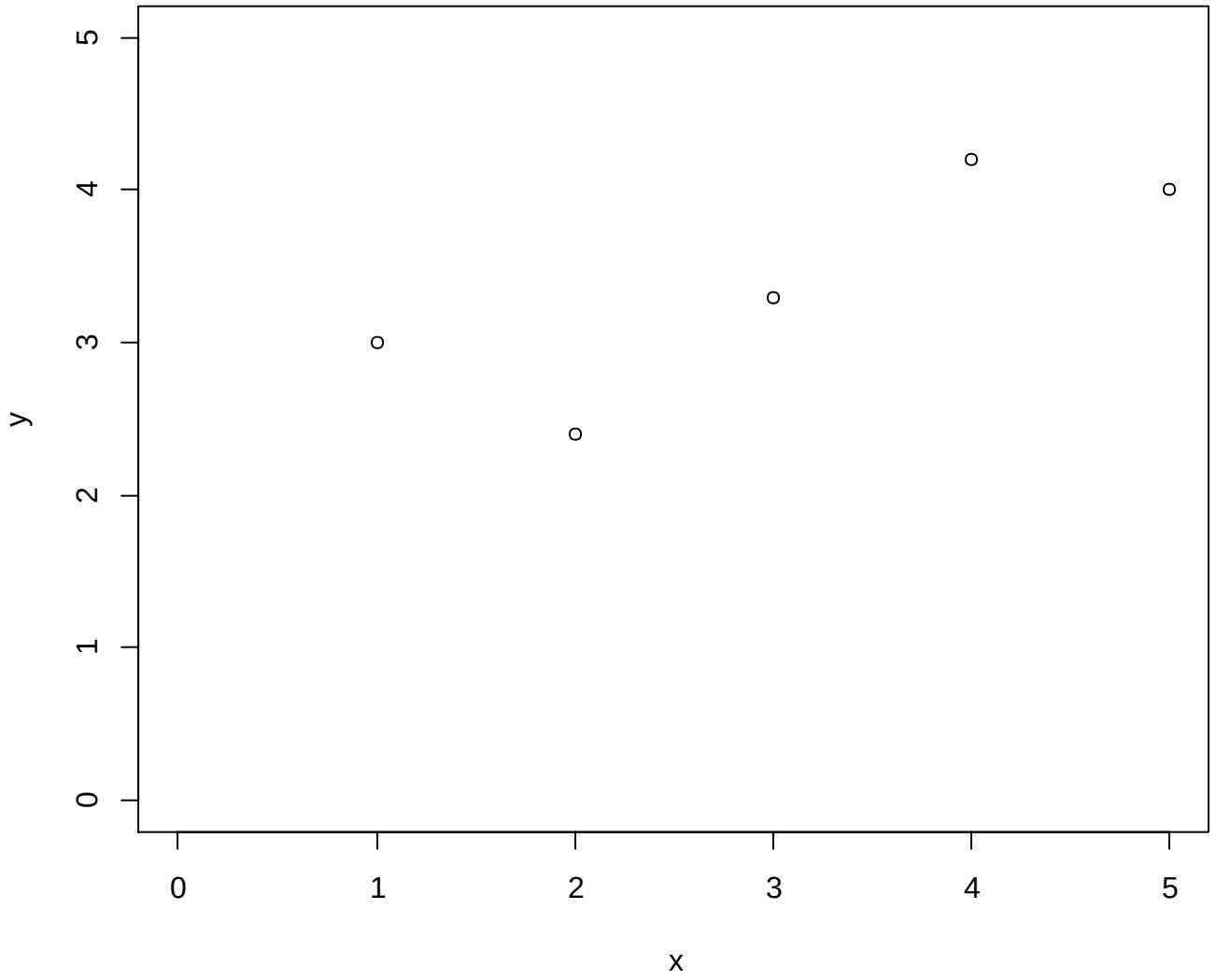
http://www.biostat.jhsph.edu/courses/bio653/misc/Regr_FigB2.gif

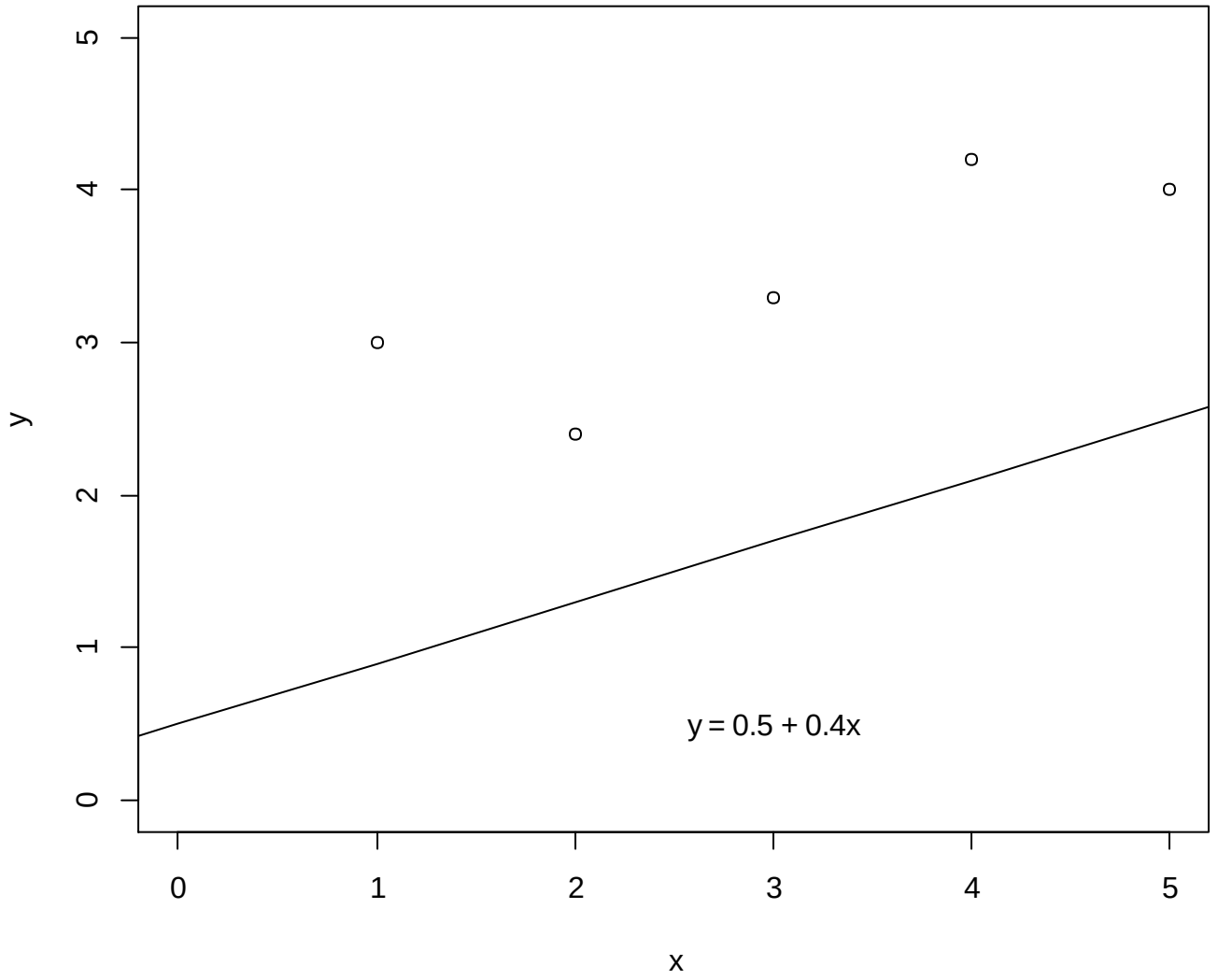
Linear regression

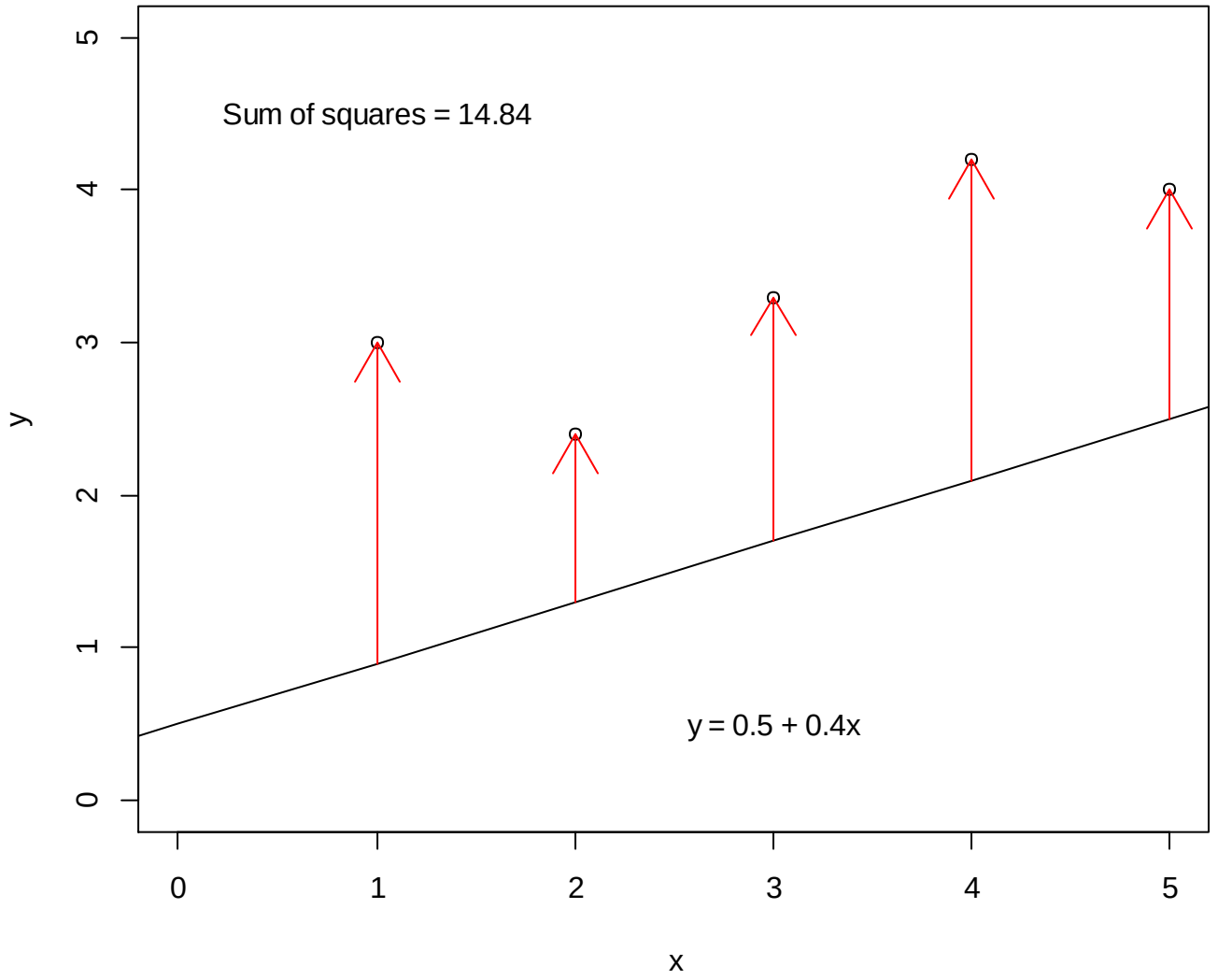
response predictor error

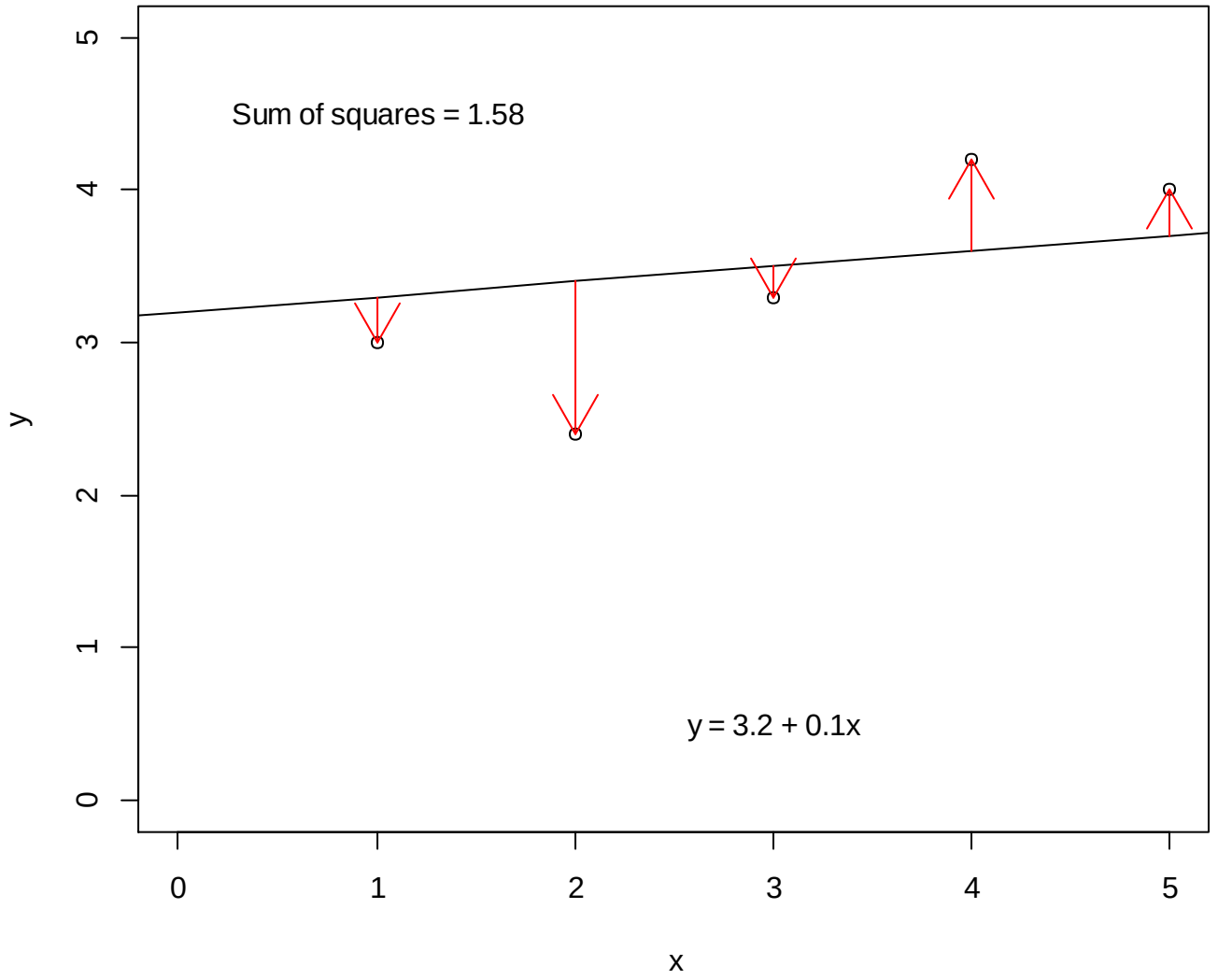
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

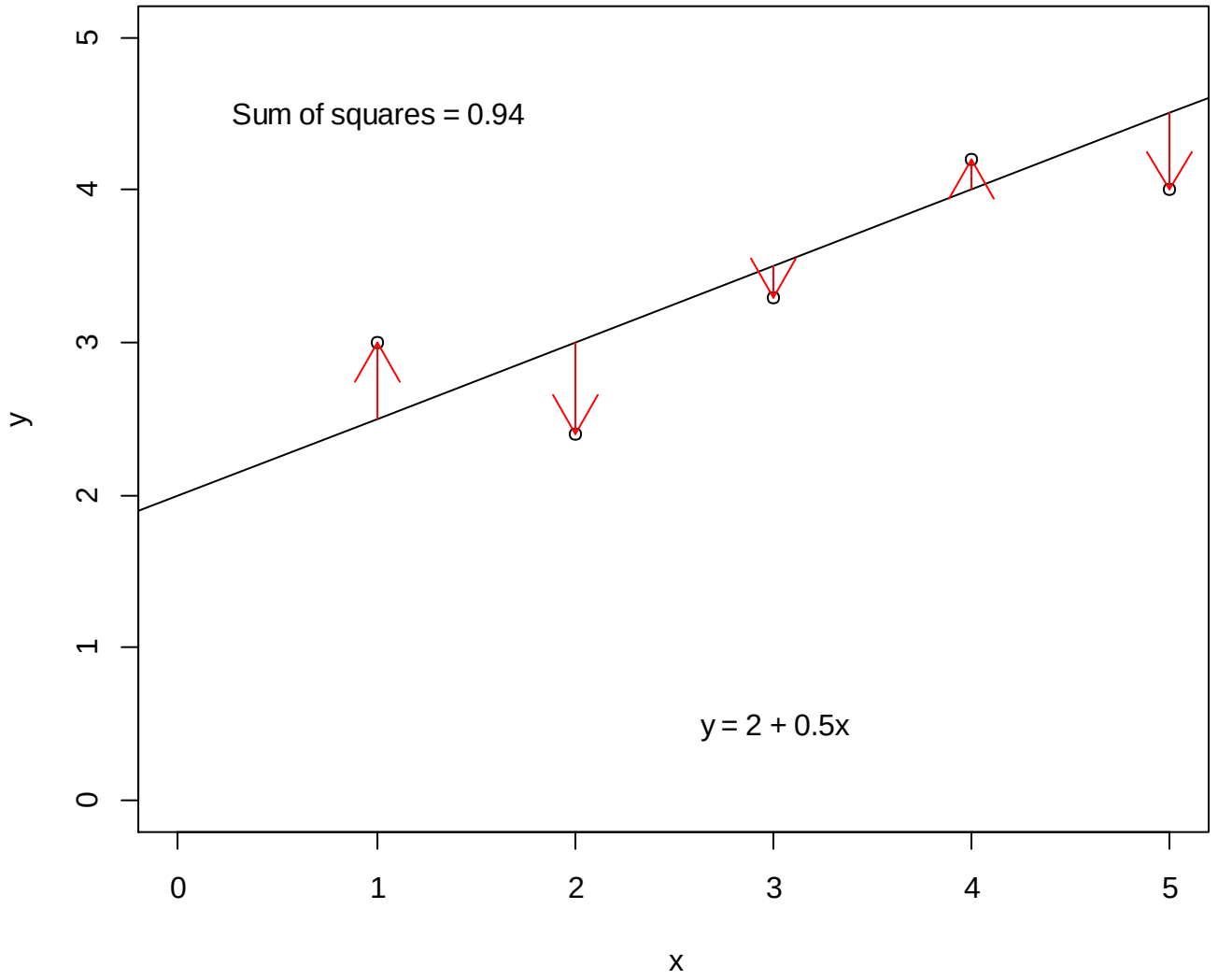
- Estimate α and β given data y_i and x_i.
- Find α and β that minimize $\sum (y_i - (\alpha + \beta x_i))^2$

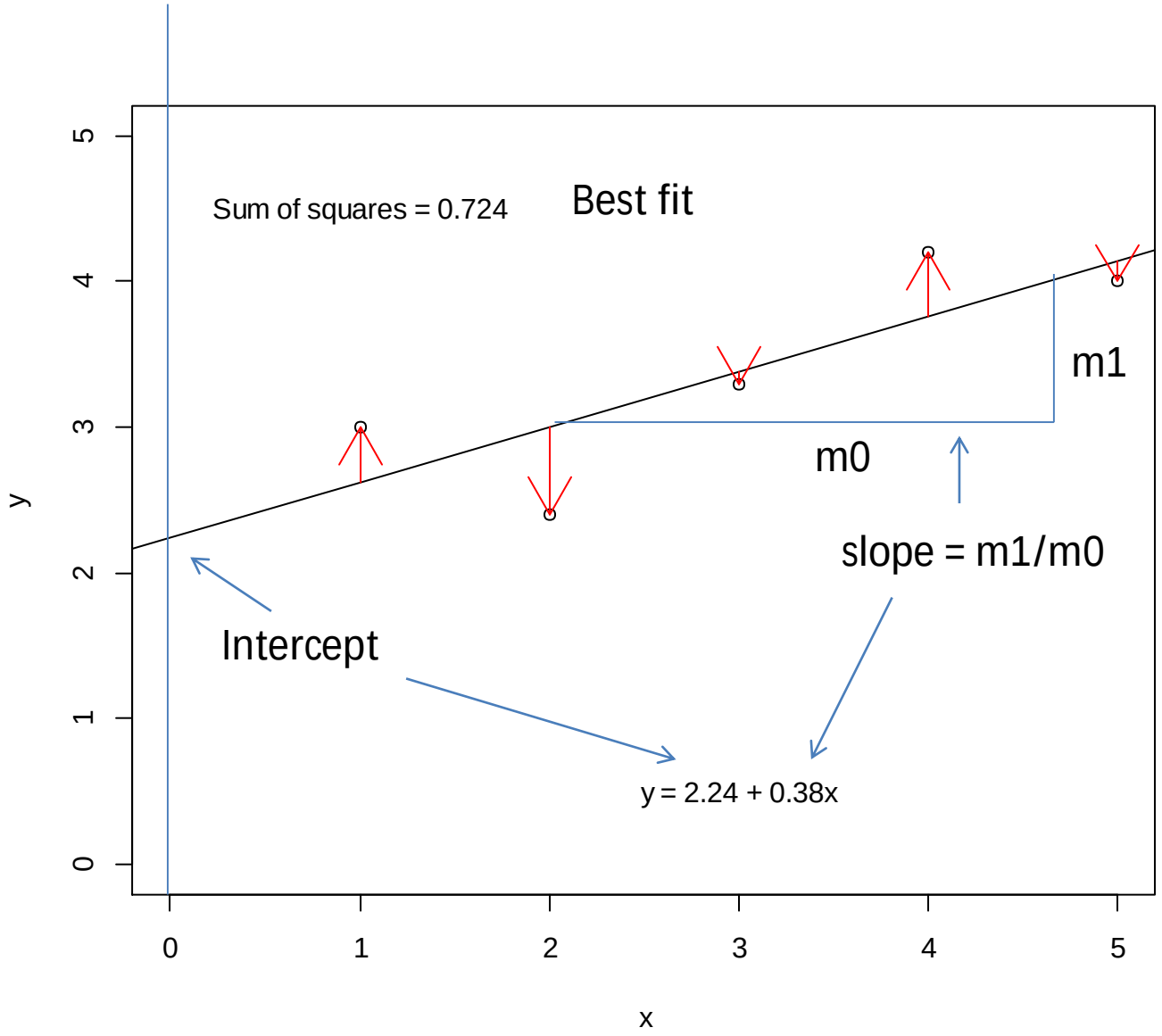

y
0
1
2
3
4
5
x
0
1
2
3
4
5


y
x
0
1
2
3
4
5
y = 0.5 + 0.4x


Sum of squares = 14.84
y = 0.5 + 0.4x
y
x
0
1
2
3
4
5


Sum of squares = 1.58
y = 3.2 + 0.1x
y
x
0
1
2
3
4
5


Sum of squares = 0.94
y = 2 + 0.5x
y
x
0
1
2
3
4
5


Sum of squares = 0.724
Best fit
m1
m0
slope = m1/m0
Intercept
y = 2.24 + 0.38x
y
x
0
1
2
3
4
5

Interpretation

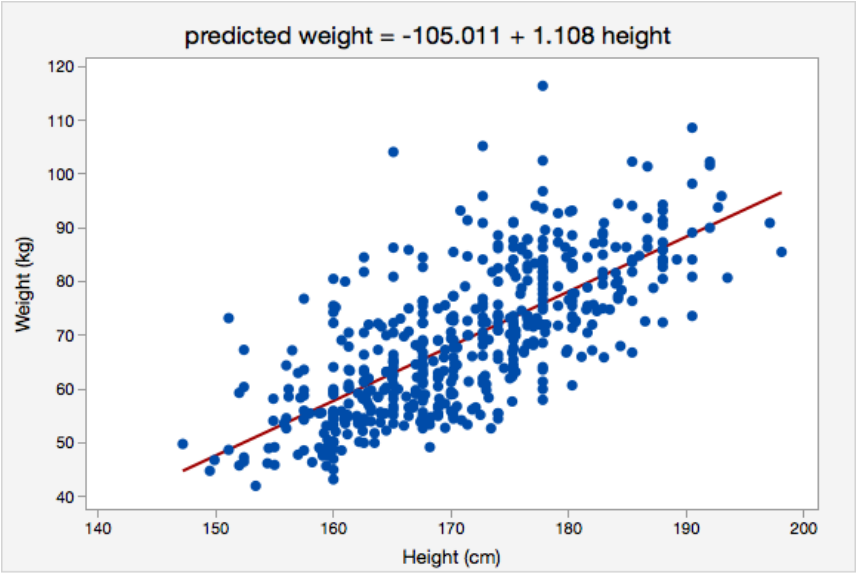
Intercept: (Not always meaningful)

Slope: Expected change in y per (unit) increase in x.

Example

Slope = 1.108

For every increase in height (cm), you expect the weight to increase by 1.1 kg.



https://onlinecourses.science.psu.edu/stat200/sites/onlinecourses.science.psu.edu.stat200/files/lesson12/weight_height_regression.png

Quiz

How would the slope change if:

- weight (y) was measured in lbs instead of kg? (1 kg = 2.2 lbs)
- Height (x) was measured in feet rather than cm? (1 foot = 30cm)

Regression and correlation

Regression slopes are scale dependent

It may not be appropriate for comparing the effects of different predictors (x) on an outcome (y).

One way to make the comparison more meaningful is if we standardize x.

$$\text{Standardized}(x) = \frac{x - \text{mean}(x)}{\text{SD}(x)}$$

Regression and correlation

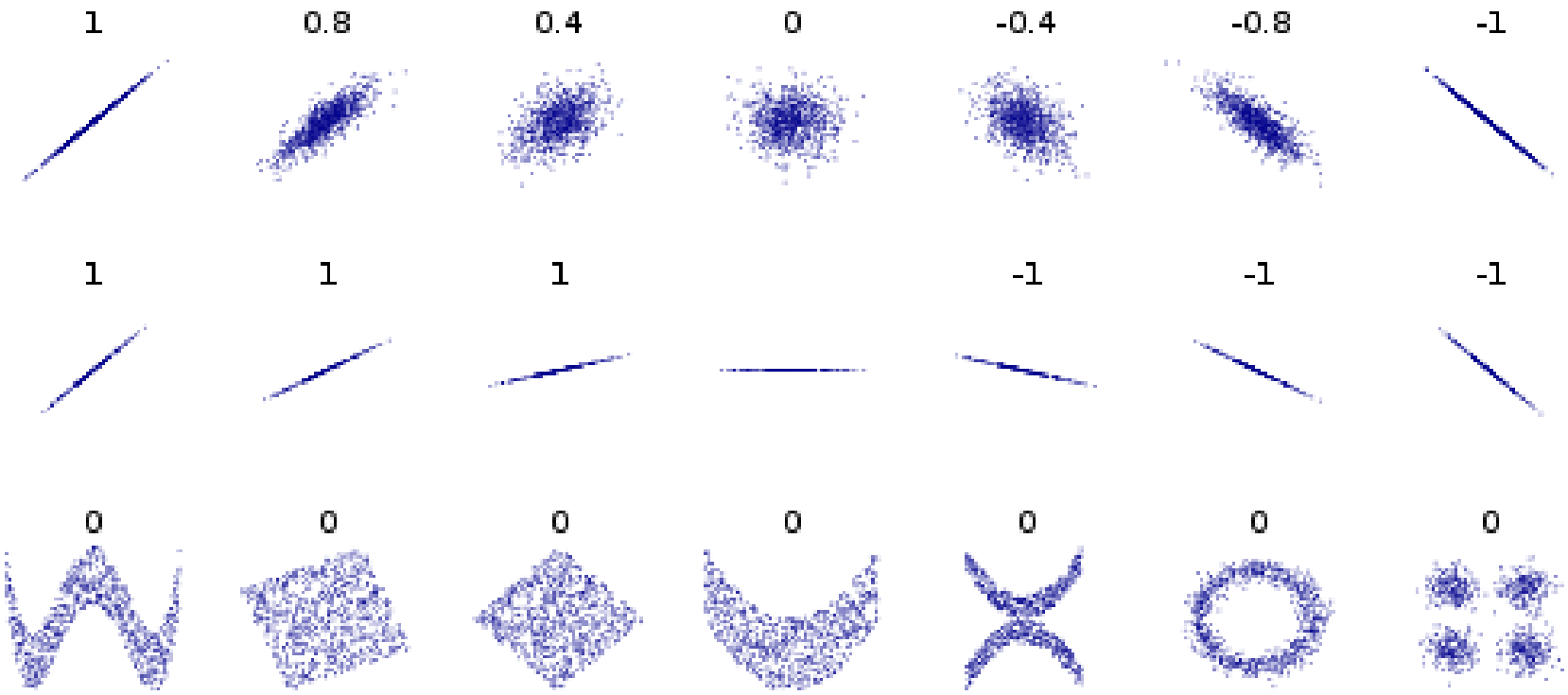
If we standardize our x and y variables before doing regression, we get correlation coefficients.

The usual definition of correlation

	Population / distribution	Sample
Mean	$E(X)$	$\bar{x} = \sum x / n$
Variance = SD^2 $Var(x)$	$E[(X - E(X))(X - E(X))]$	$\sum (x - \bar{x})(x - \bar{x}) / n$
Covariance $Cov(x,y)$	$E[(Y - E(Y))(X - E(X))]$	$\sum (y - \bar{y})(x - \bar{x}) / n$

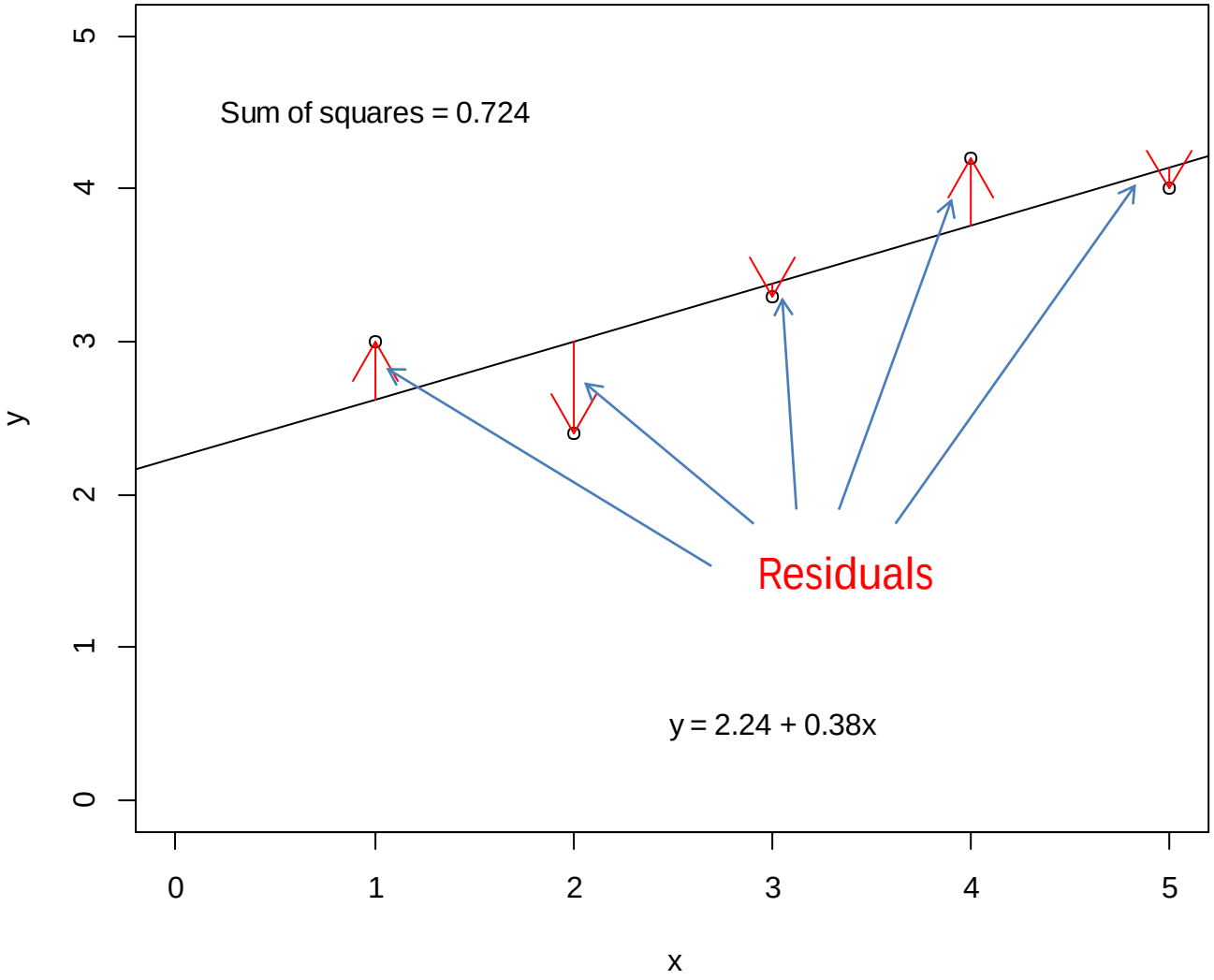
$$Corr(x,y) = \frac{Cov(x,y)}{SD(x)SD(y)}$$

Correlation as a measure of relatedness between two variables



https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Correlation_examples2.svg

Residuals



Facts:

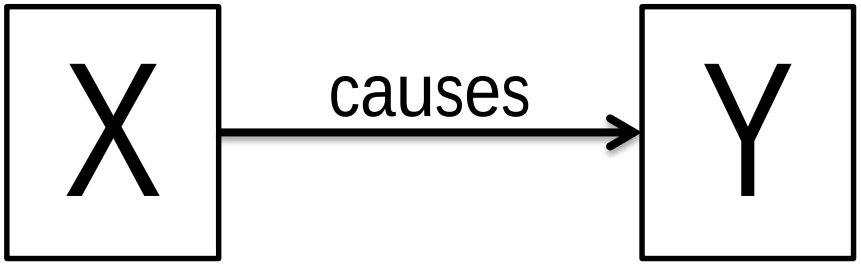
1. Sum of residuals = 0
2. Corr(x, res) = 0

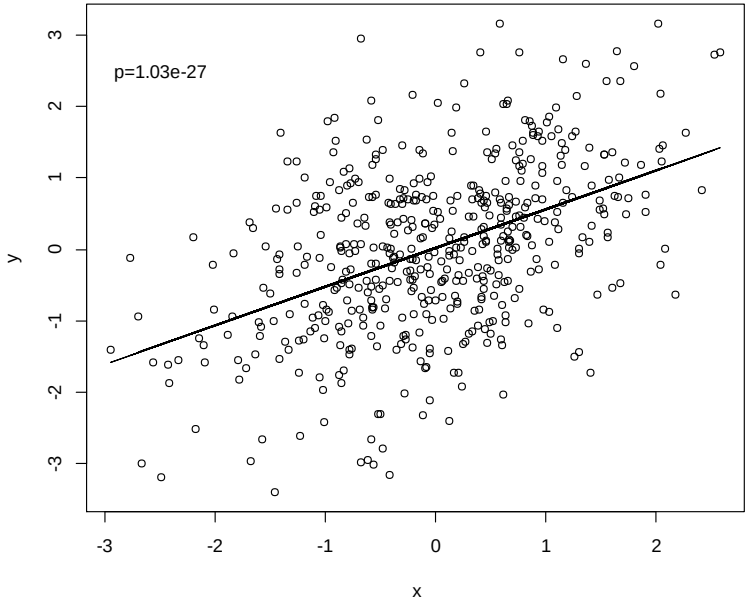
Significance testing

Testing if slope is significantly different from 0:

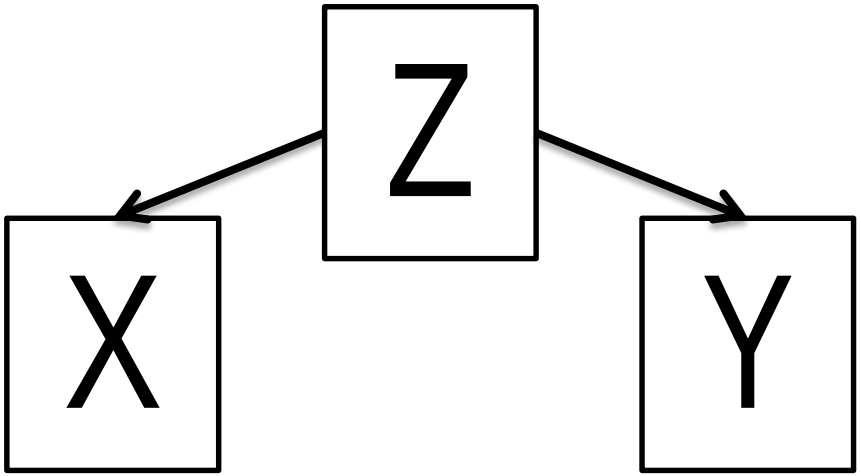
- Wald test: t = estimate / s.e.
- Same test for correlation and regression
- Assumptions: Residuals are Normally distributed.

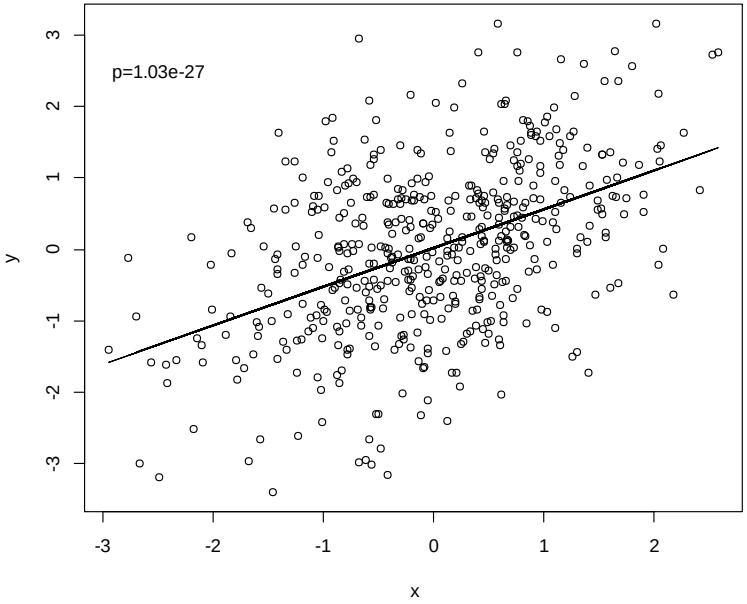
Causality





Confounding





Confounding

How do we “remove” the effect of Z when looking at the effect of X on Y?

1. Carry out regression of Y on Z
2. Carry out regression of Residuals of Y on X

Confounding

What about?

1. Carry out regression of Y on Z
2. Carry out regression of X on Z
3. Carry out regression of Residuals of Y on residuals of X

Better interpretation of slope: Change in Y for each change in X holding Z constant

Confounding

It turns out this approach is the same as doing a regression with two predictor variables.

A regression with two predictor variables

$$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i$$

Find α, β, γ which minimizes

$$\sum (y_i - (\alpha + \beta x_i + \gamma z_i))^2$$

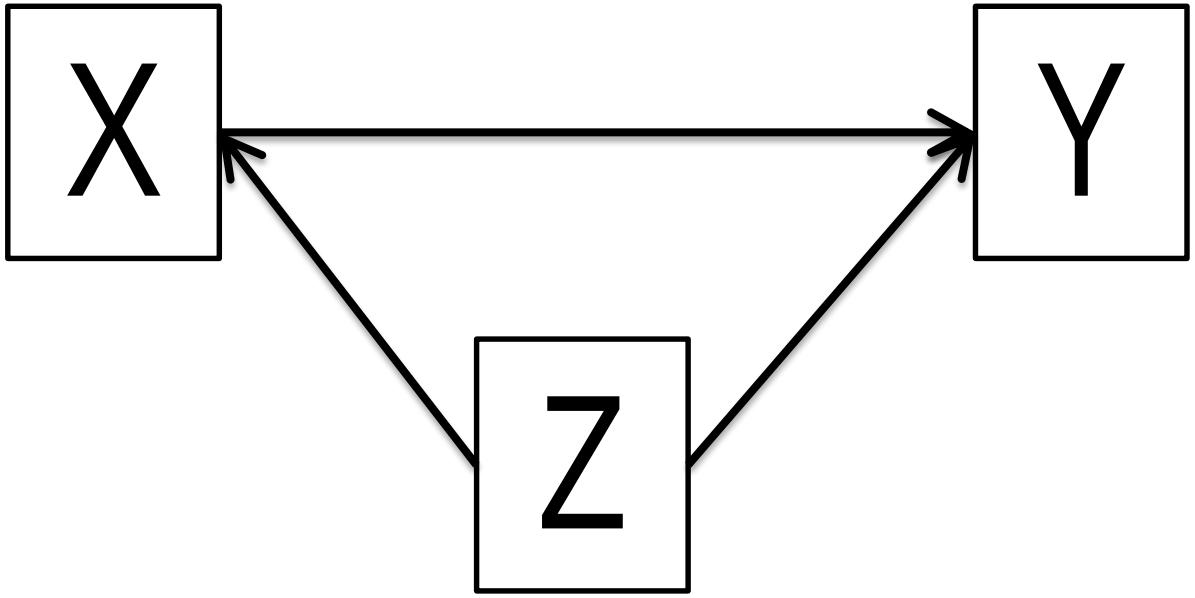
Multiple regression

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \varepsilon_i$$

Interpretation:

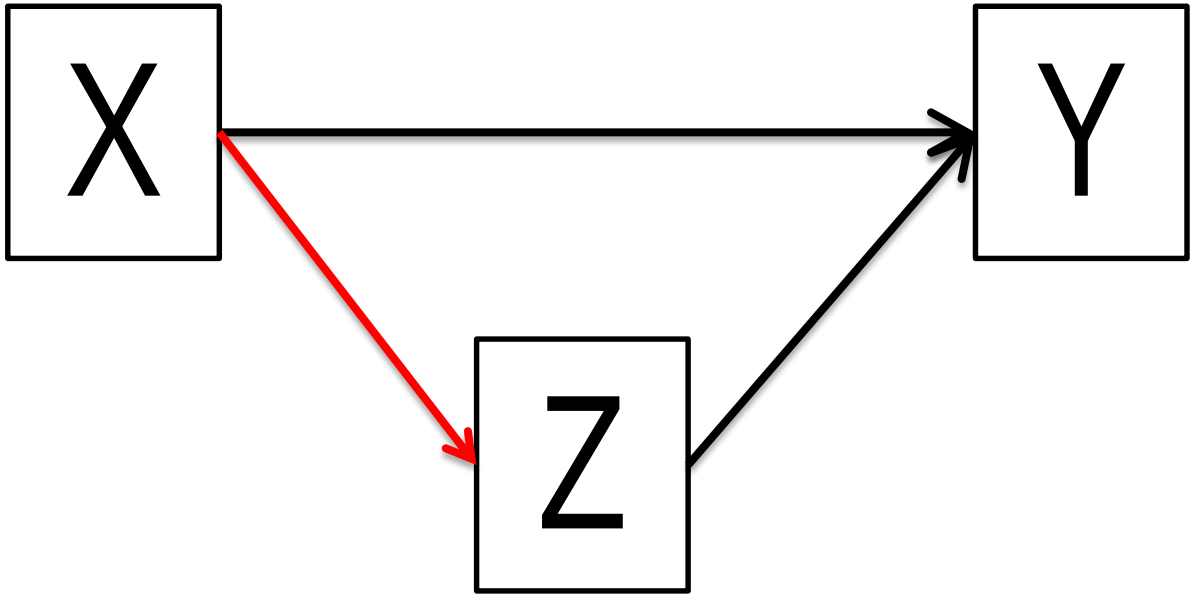
- β_1: Average change in y per unit change in x_1 assuming x_2, x_3, ..., stays constant.
- β_2: Also average change in y per unit change in x_2 assuming x_1, x_3, ..., stays constant, etc.

Uses of multiple regression



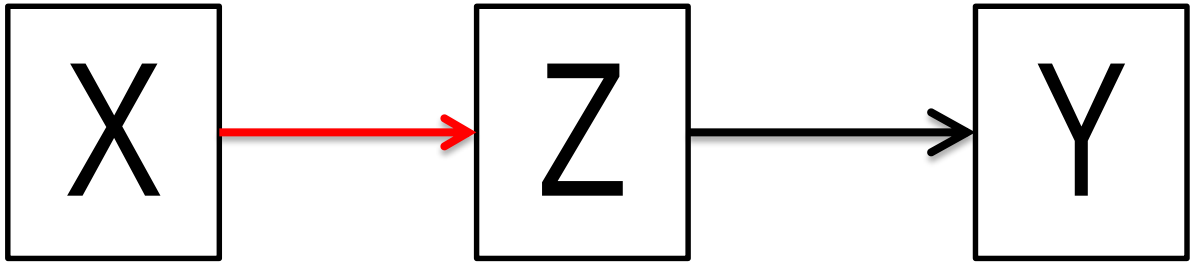
- Effect of X on Y independent of association with Z.

Uses of multiple regression



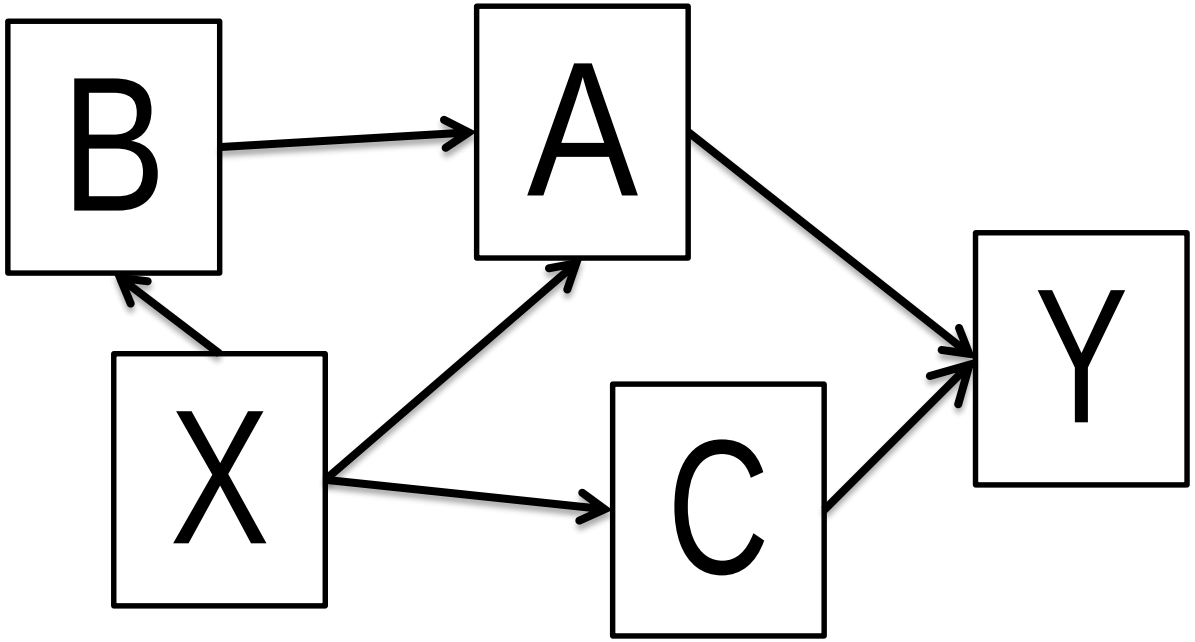
- Direct effect of X on Y independent of indirect effect through Z.

Uses of multiple regression



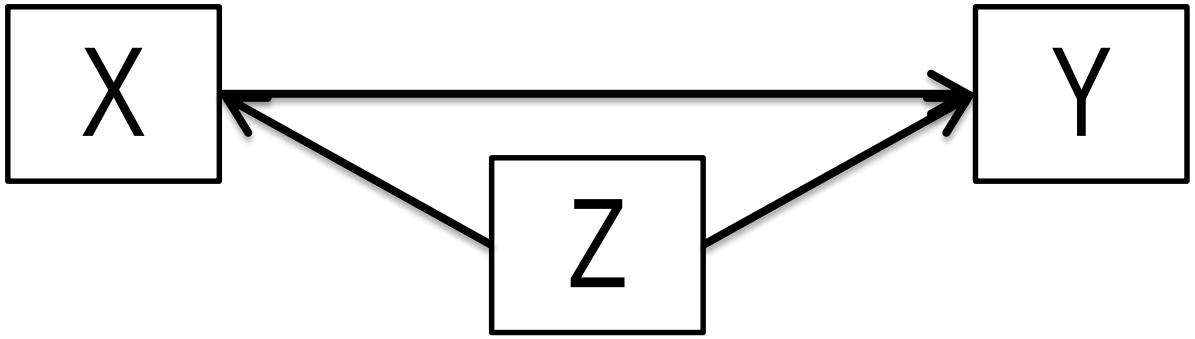
- If you expect all or most of the effect of X on Y is through Z, then you shouldn't use multiple regression to examine the effect of X on Y.

Uses of multiple regression



- Don’t just put all your variables into the same multiple regression model.
- Think about which variable should be included for each of the effects you want to estimate.

Some terminology



- Multiple regression / Multivariate regression
- y: Dependent variable / Outcome /response variable
- x,z: Independent variable(s) / Explanatory variable(s), covariate(s) / predictors
- z: Confounder(s), covariate(s)
- Controlling/adjustment for covariates/confounders

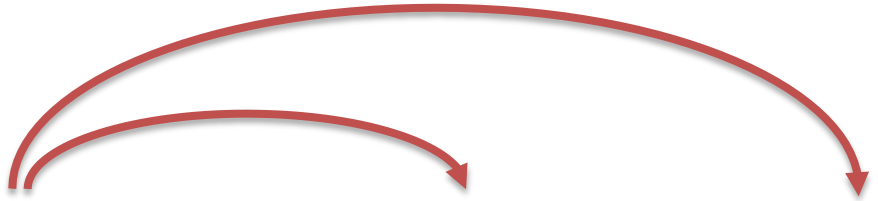
Assumptions in multiple regression to infer causal effects

- Covariates are measured without error
  - e.g. Controlling for Social Economic Status
- Linearity
- Lack of interactions (effect of one variable does not depend on another)
- All relevant confounders are included
- Errors are Normal and independent (and identically distributed)

The more variables there are, the less likely it is that all of the above are satisfied.

Categorical covariates

Introducing dummy variables:



Subject	Ethnicity (X)	Korean (X1)	Japanese (X2)
1	Chinese	0	0
2	Chinese	0	0
3	Korean	1	0
4	Chinese	0	0
5	Japanese	0	1
6	Japanese	0	1
7	Korean	1	0
8	Chinese	0	0

Categorical covariates

"Factors" in R are automatically turned into dummy variables in R when running regression. In the 'mtcars' dataset, the variable cyl (number of cylinders) has 3 levels: 4, 6, and 8.

```
> summary(lm(mpg ~ cyl, data=mtcars))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.8846     2.0738   18.27  < 2e-16 ***
cyl          -2.8758     0.3224   -8.92 6.11e-10 ***

> summary(lm(mpg ~ as.factor(cyl), data=mtcars))
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      26.6636     0.9718  27.437  < 2e-16 ***
as.factor(cyl)6  -6.9208     1.5583  -4.441 0.000119 ***
as.factor(cyl)8 -11.5636     1.2986  -8.905 8.57e-10 ***
```

Categorical covariates

Use the ‘relevel()’ command to switch reference category:

```
> summary(lm(mpg ~ relevel(as.factor(cyl), ref="6"), data=mtcars))
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         19.743      1.218  16.206 4.49e-16 ***
relevel(as.factor(cyl), ref = "6")4    6.921      1.558   4.441 0.000119 ***
relevel(as.factor(cyl), ref = "6")8   -4.643      1.492  -3.112 0.004152 **
```

For more flexible contrast testing:

http://www.ats.ucla.edu/stat/r/faq/testing_contrasts.htm

http://rstudio-pubs-static.s3.amazonaws.com/65059_586f394d8eb84f84b1baaf56ffb6b47f.html

Categorical covariates

Overall tests:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i$$

Say x_{1i} and x_{2i} are dummy variables for a 3-level categorical variable x. How do I test for the overall significance of x?

Null hypothesis: $\beta_1 = \beta_2 = 0$

```
> anova(lm(mpg ~ as.factor(cyl), data=mtcars))
               Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(cyl)  2 824.78  412.39  39.697 4.979e-09 ***
Residuals      29 301.26   10.39
```

Interactions

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

- Assume both x_1 and x_2 cause y
- Effect of x_1 on y is the same for all values of x_2. Vice versa.
- What if the effect of x_1 on y depends on x_2?

Interactions

Assume x_2 is a binary dummy variable (coded 0 or 1).

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

If $x_{2i} = 0$,

$$y_i = \alpha + \beta_1 x_{1i} + \varepsilon_i$$

If $x_{2i} = 1$,

$$y_i = (\alpha + \beta_2) + (\beta_1 + \beta_3) x_{1i} + \varepsilon_i$$

- Effectively fitting 2 different linear regressions
- Only difference: ε_i assumed to have the same variance

Interactions

R outputs

```
> summary(lm(mpg ~ as.factor(cyl)*wt, data=mtcars))
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          39.571      3.194  12.389 2.06e-12 ***
as.factor(cyl)6     -11.162      9.355  -1.193 0.243584
as.factor(cyl)8     -15.703      4.839  -3.245 0.003223 **
wt                   -5.647      1.359  -4.154 0.000313 ***
as.factor(cyl)6:wt    2.867      3.117   0.920 0.366199
as.factor(cyl)8:wt    3.455      1.627   2.123 0.043440 *

> anova(lm(mpg ~ as.factor(cyl)*wt, data=mtcars))
Analysis of Variance Table
                  Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(cyl)     2 824.78  412.39 68.7811 4.138e-11 ***
wt                 1 118.20  118.20 19.7147 0.0001473 ***
as.factor(cyl):wt  2  27.17   13.58  2.2658 0.1238570
Residuals         26 155.89    6.00
```

- Note that we the non-interaction effects and test may not be meaningful anymore.

Overall fit of model

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

- Residuals are uncorrelated with x_1 and x_2.

$$\mathrm{Var}(y_i) = \mathrm{Var}(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i}) + \mathrm{Var}(\varepsilon_i)$$

Total variance — Variance explained by model — Residual variance

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

Overall fit of model

- When there is only 1 variable, R^2 = correlation^2
- also called the coefficient of determination
- A high R^2 doesn't mean that your model is good.
- Estimates of R^2 always increase with increasing number of variables.
- Measures like AIC and BIC penalize for the number of parameters in the model.

(Binary) Logistic regression

Another way to write down a multiple regression model:

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots$$

Binary logistic regression

$$y_i \sim Bern(p_i)$$

$$\text{logit}(p_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots$$

(Binary) Logistic regression

Why?

- Dependent variable Binary – violation of assumptions in (linear) multiple regression for significance testing
- Prediction of outcome will not exceed 0 or 1

But:

- Assumes effects are linear on log odds (logit) scale (Effect size interpretable as log odds ratio)
- No decomposition of variance

Generalized linear model

Exponential family

$$y_i \sim f(\theta_i)$$

$$g(\theta_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots$$

- e.g. Poisson regression, Binomial regression, Negative Binomial regression, Cox regression
- in R, use glm() instead of lm(), e.g.
  - glm(y~x, family=binomial) for logistic regression
  - glm(y~x, family=poisson) for Poisson regression

Generalized linear model

What to do with violations of distributional assumptions? e.g.

- Count data but not Poisson or Neg Binomial
- Continuous variable but not Normal

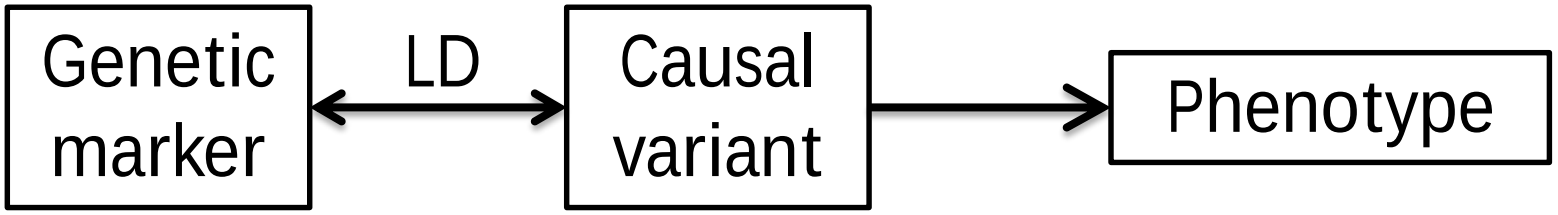
Estimates of β_1, β_2, etc. still valid, although standard error/tests not accurate.

The sandwich/Huber/White variance estimator

- http://thestatsgeek.com/2014/02/14/the-robust-sandwich-variance-estimator-for-linear-regression-using-r/
- Just add the ‘robust’ option in Stata

APPLICATIONS IN GENETICS

Linkage disequilibrium



- The correlation coefficient (R) or coefficient of determination (R^2) is a common measure of LD

GWAS

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i$$

However,

- Too many variants (SNPs) for number of participants
- But we cannot reliably estimate β_1, β_2 , ...

GWAS

SNP-wise regression:

For each SNP j,

$$y_i = a_j + b_j x_{ji} + \varepsilon_{ji}$$

- $b_j \neq \beta_j$

Controls for confounders in GWAS

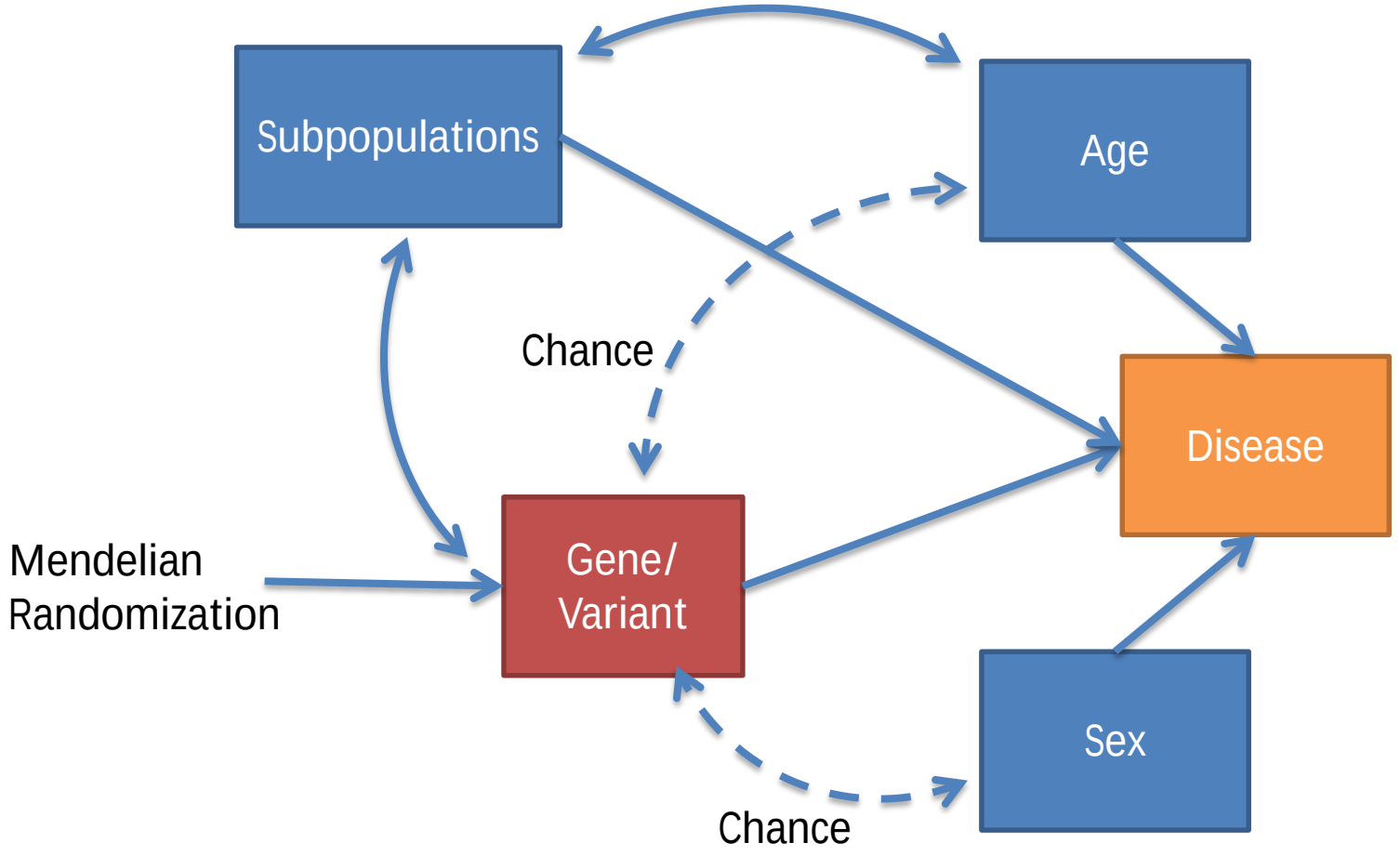
- A confounder must be associated with both the outcome and the genetic variant
- But beware of controlling for intermediate variables
- Usually, just sex, age, and Principal components (for population stratification)

GWAS

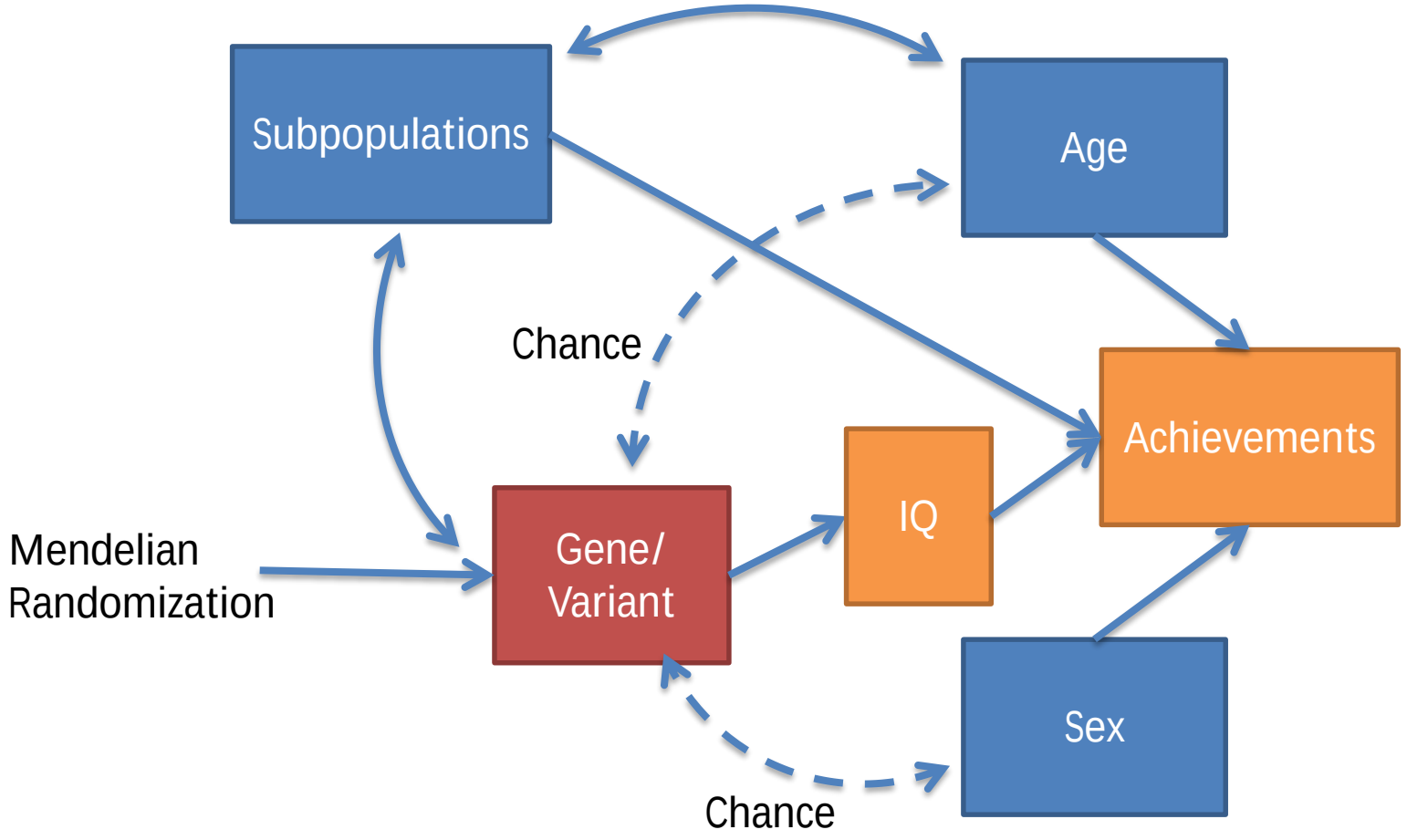
Controls for confounders in GWAS

- When populations are not randomly mating, there will be population stratification (population subgroups with different genotypic makeup)
- Control of population stratification by including principal components as covariates.

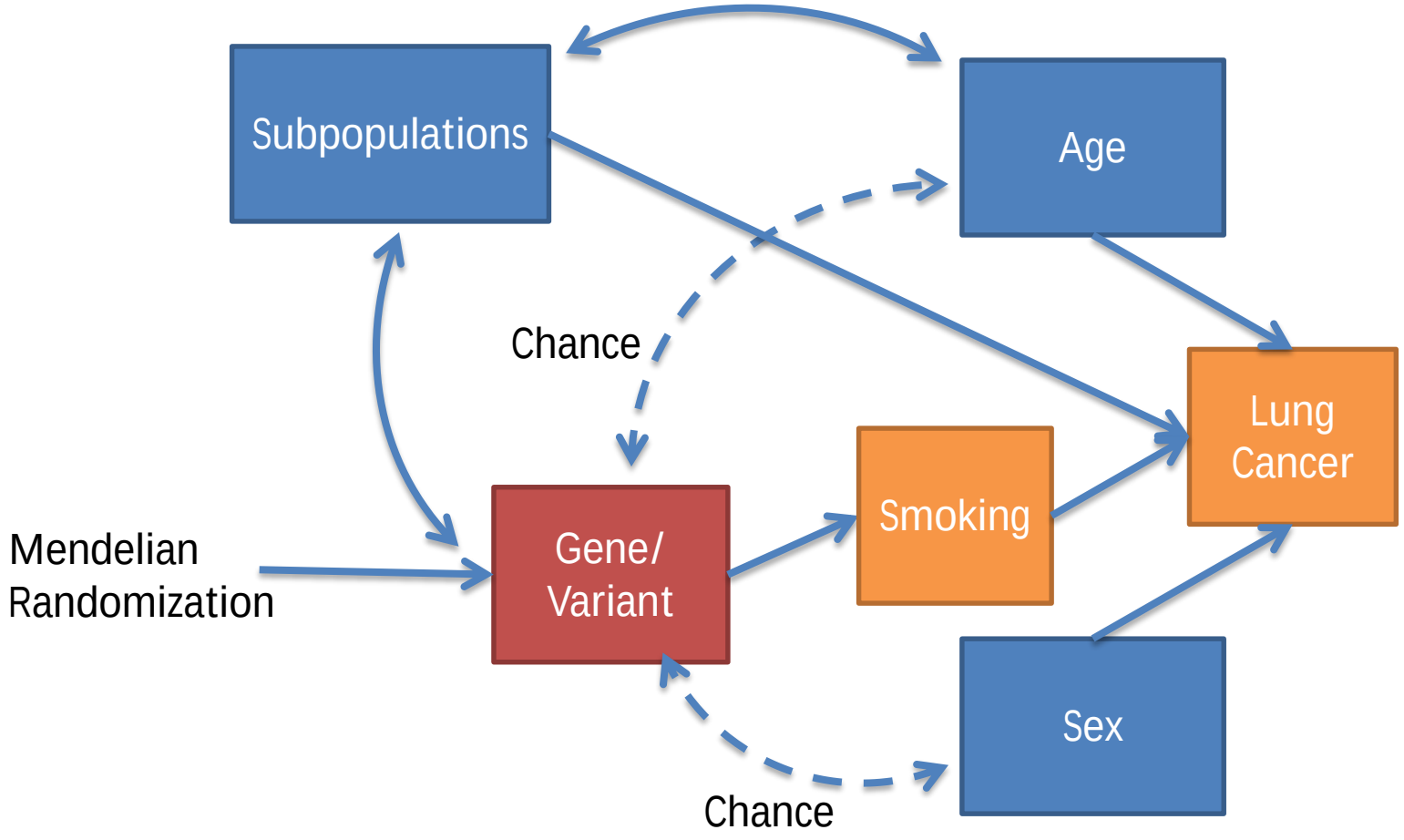
Some scenarios



Some scenarios



Some scenarios



Some scenarios