Regression part 2

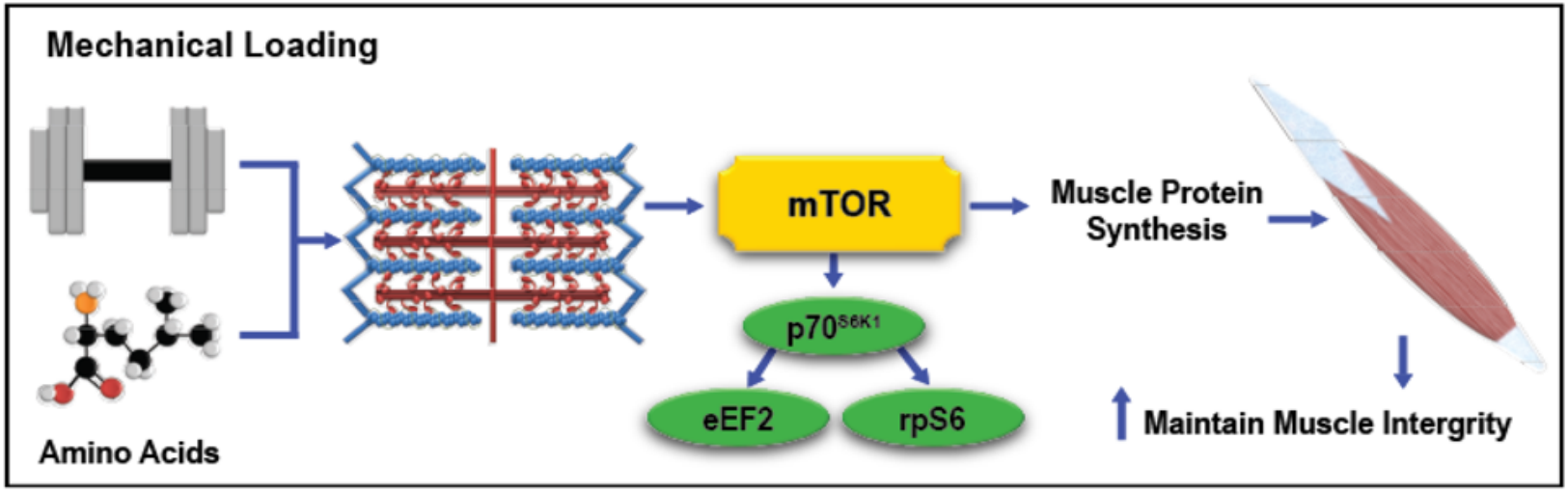
Daniel Hammarström

2019-10-28

Issues in studies of association

- Influential data points
- Correlation does not imply causation
- Regression towards the mean

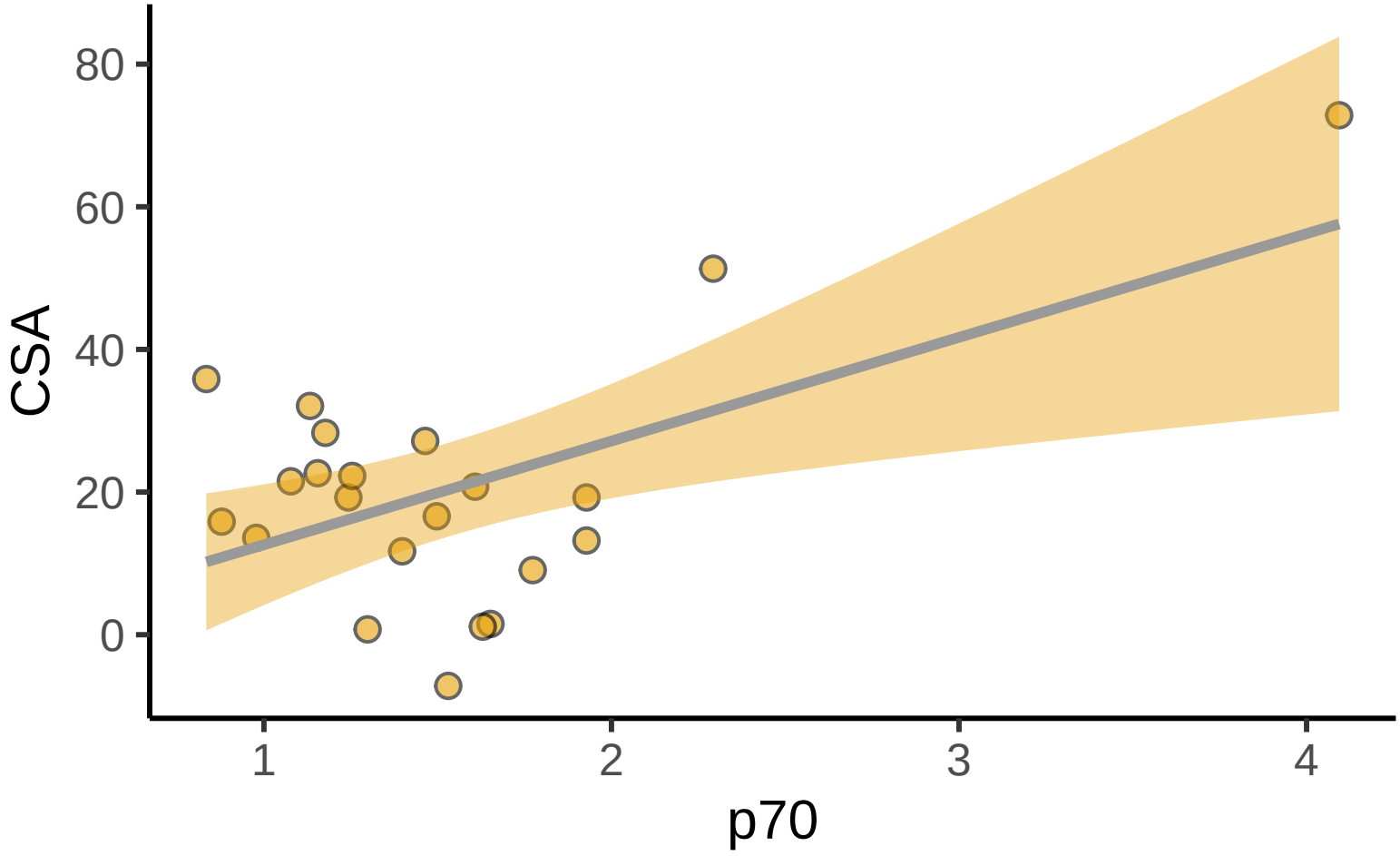
Influential data points – mTOR signaling and exercise induced muscle hypertrophy



Mechanical loading and amino acids maximize mTORC1 signaling and muscle protein synthesis, thus contributing to the maintenance of skeletal muscle mass. Abbreviations: mTOR, mammalian target of rapamycin; $p70^{S6K1}$, 70-kDa ribosomal protein S6 kinase 1; eEF2, eukaryotic elongation factor 2; rpS6, ribosomal protein S6.

Pasiakos, S. M. (2012). "Exercise and Amino Acid Anabolic Cell Signaling and the Regulation of Skeletal Muscle Mass." Nutrients 4(7).

Exercise induced P70 S6-kinase phosphorylation predicts muscle hypertrophy (Mitchell et al. 2013)



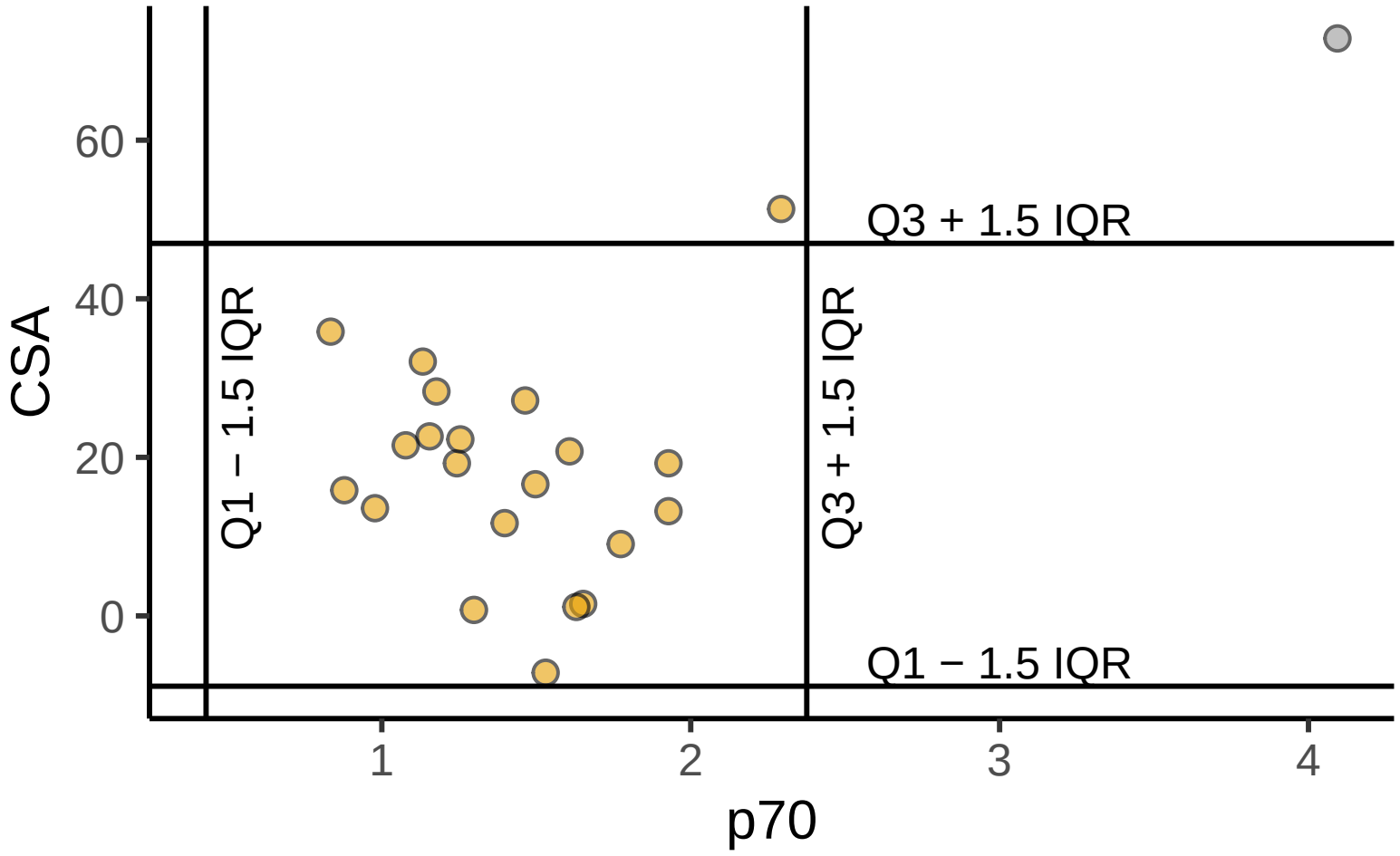
Mitchell, C. J., et al. (2013). "Muscular and Systemic Correlates of Resistance Training-Induced Muscle Hypertrophy." PLoS One 8(10): e78636.

Influential data points

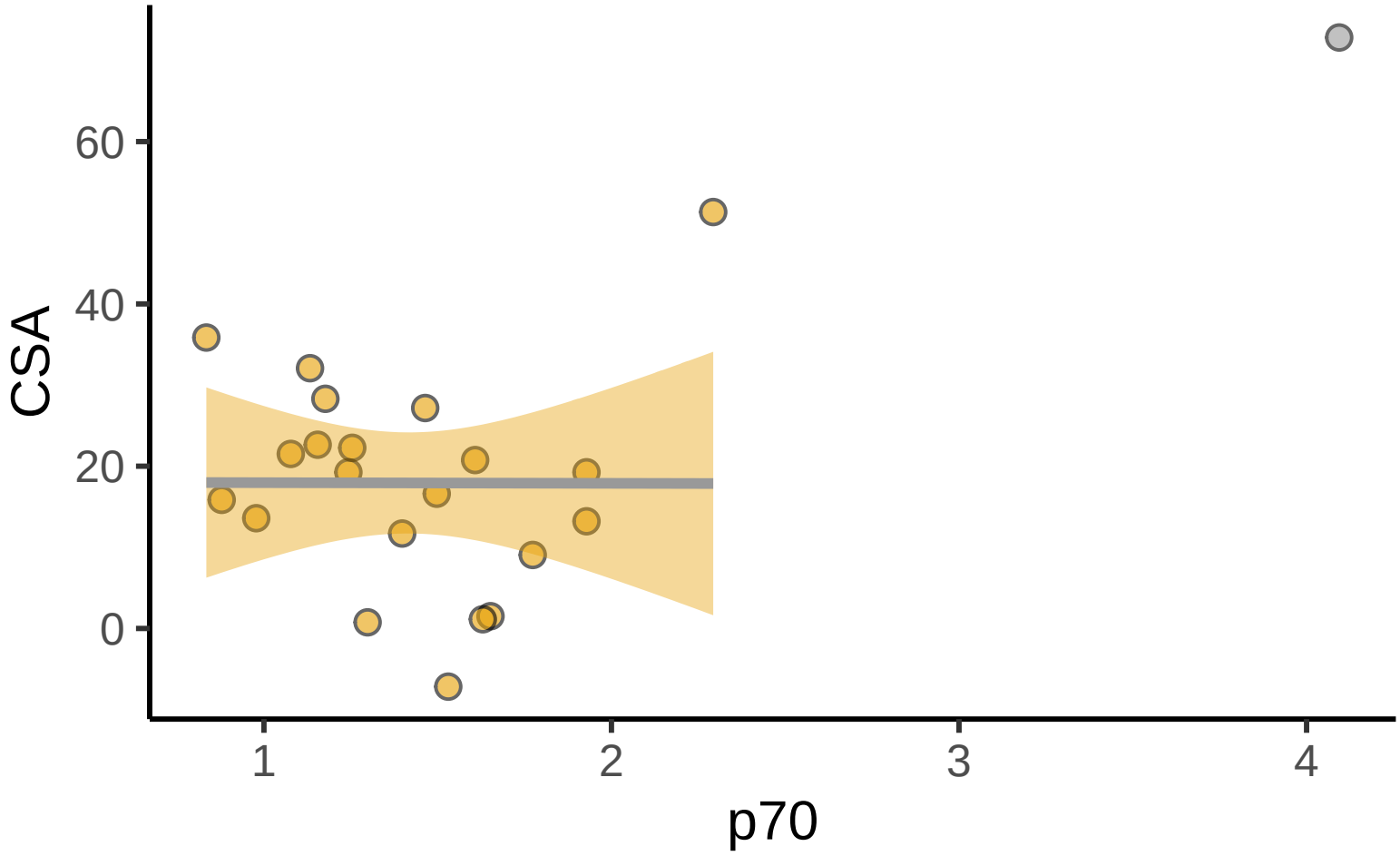
- Data points that substantially deviates from the rest of the data may affect the interpretation of regression models.
- "Leverage" is the effect each data point has on the model, unusual X-values produces larger leverage
- This can be assessed by looking at the graph, and numerically
- A tool in simple regression would be to assess outliers (in the X-axis) on model characteristics

Detect outliers

- An outlier is defined as $Q3/Q1 \pm 1.5 \times IQR$



Re-do analysis without outlier

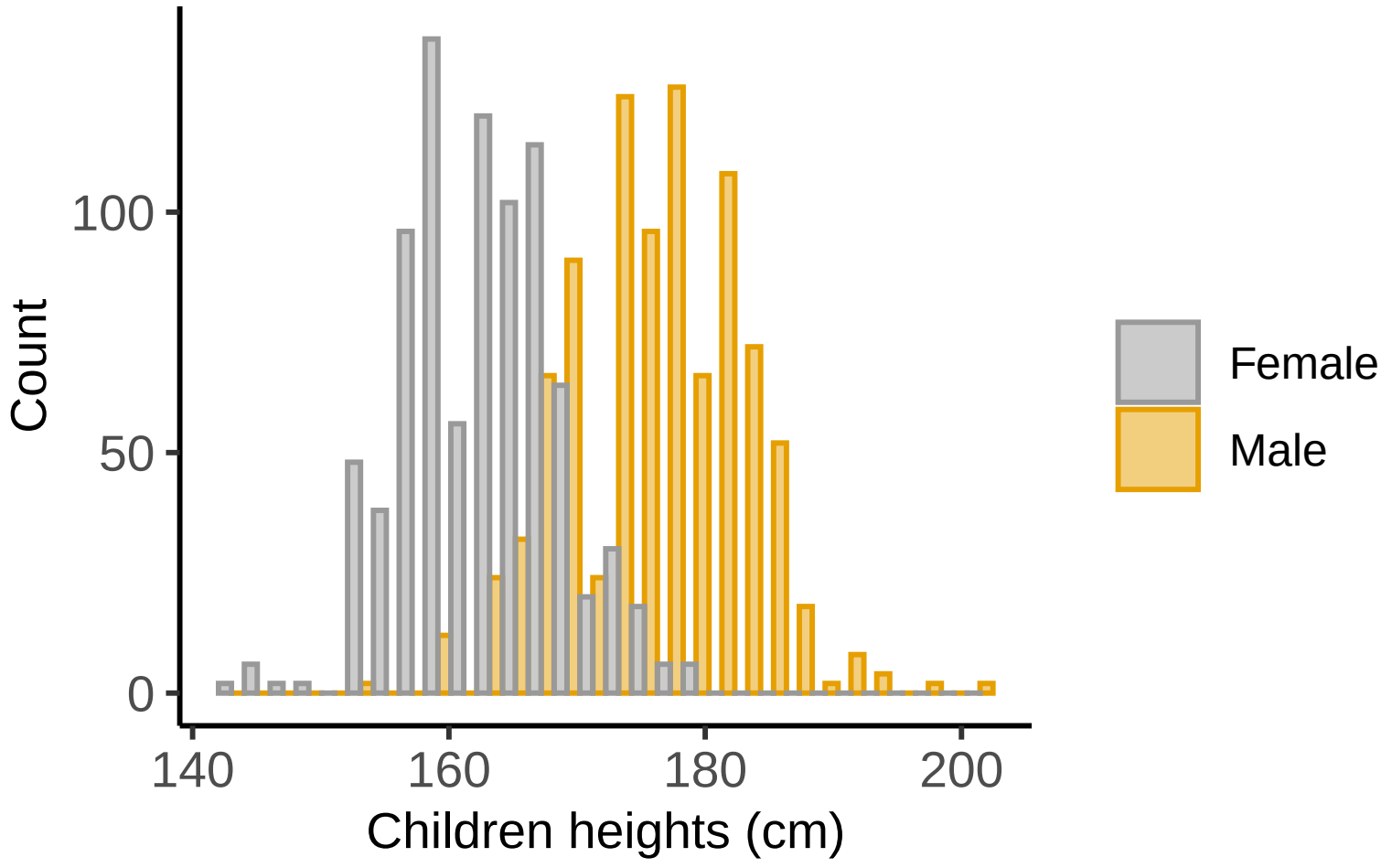


What can we conclude from the Mitchell data-set?

Regression towards the mean

- Francis Galton analyzed parents and children heights to study heritability (how much of a trait can be explained by genetics?)
- Does parents heights determine children heights?

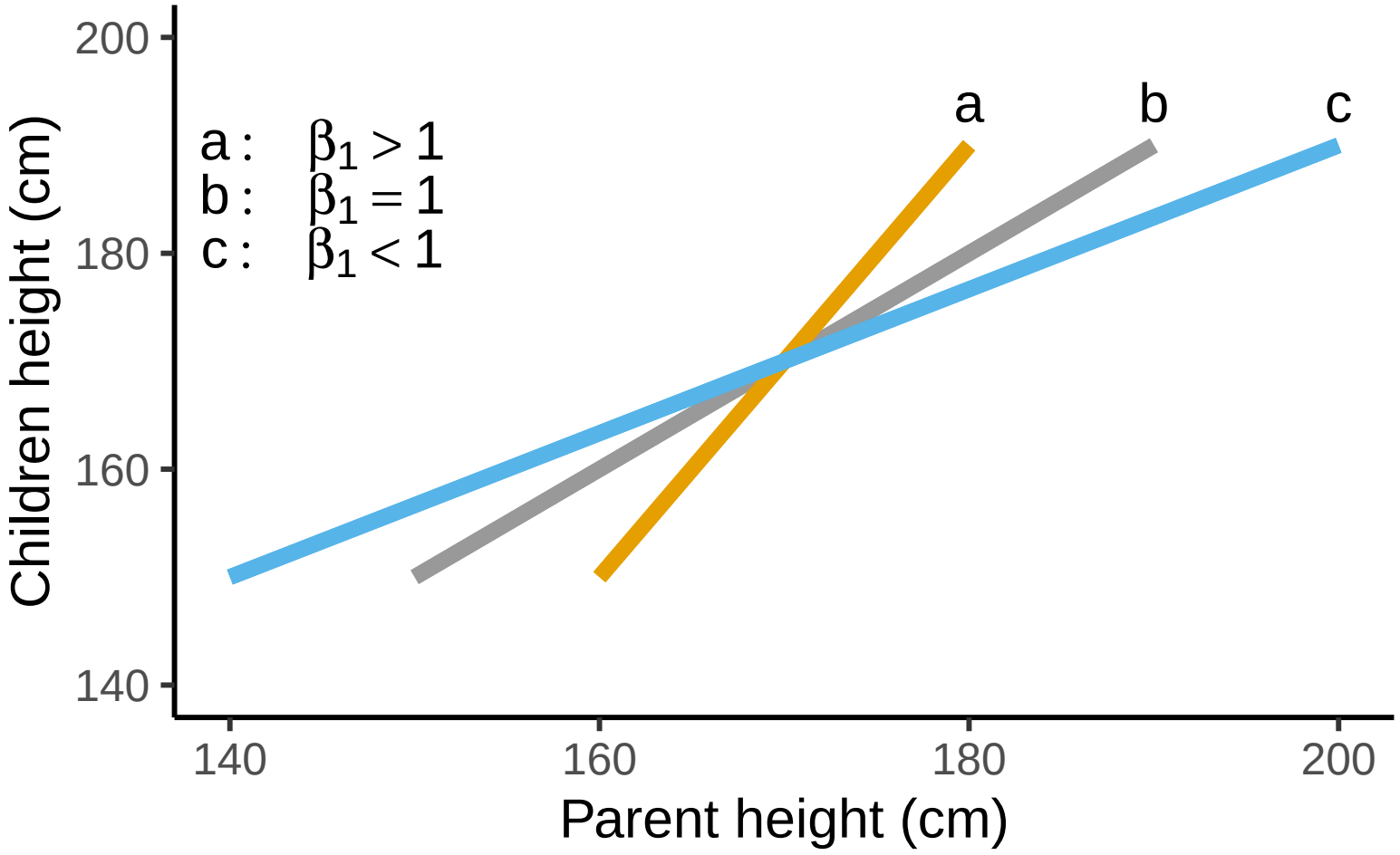
Regression towards the mean



- Do tall parents have tall children?

Regression towards the mean

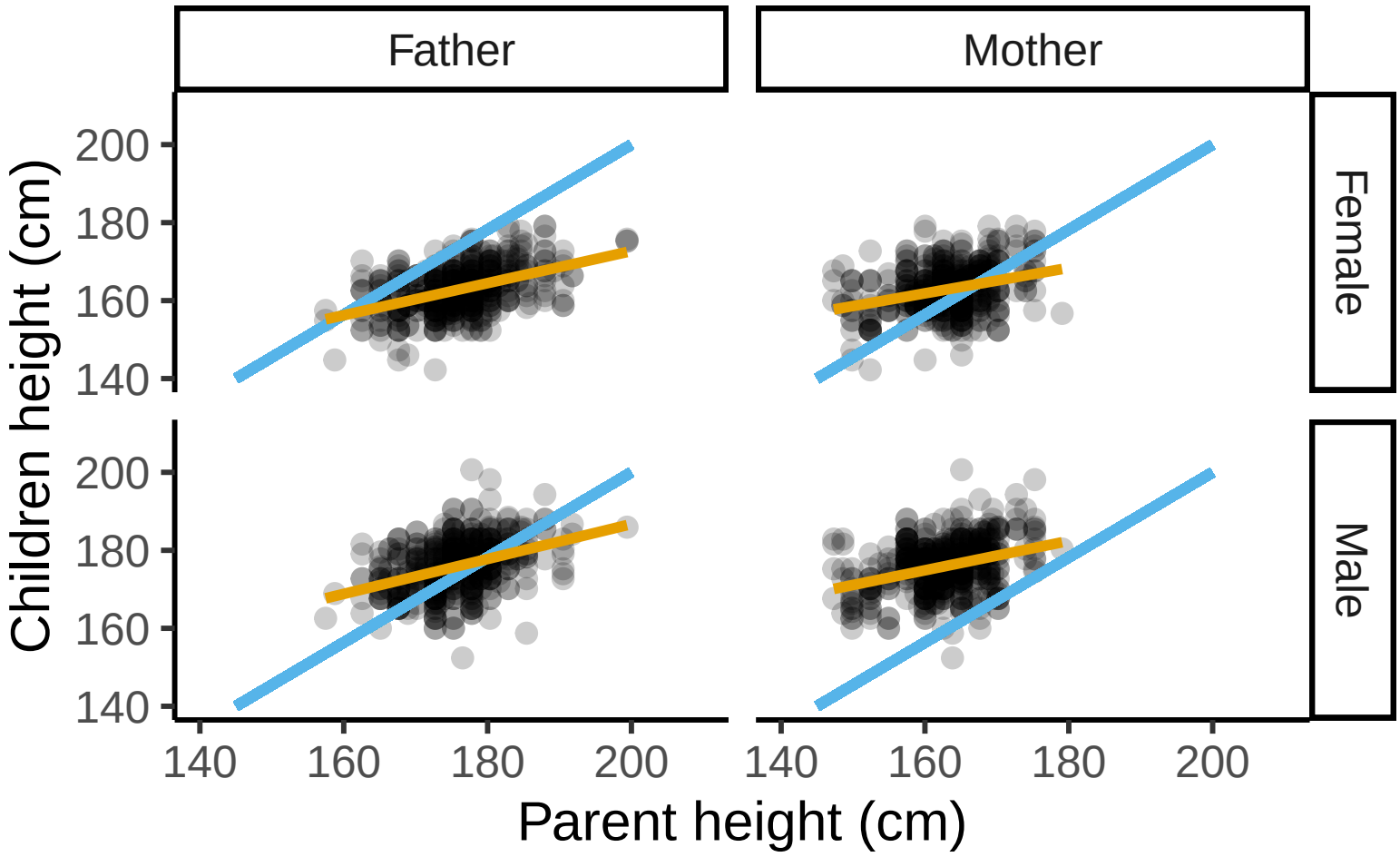
- If parents heights would predict children heights, what would the regression line look like?



Regression towards the mean

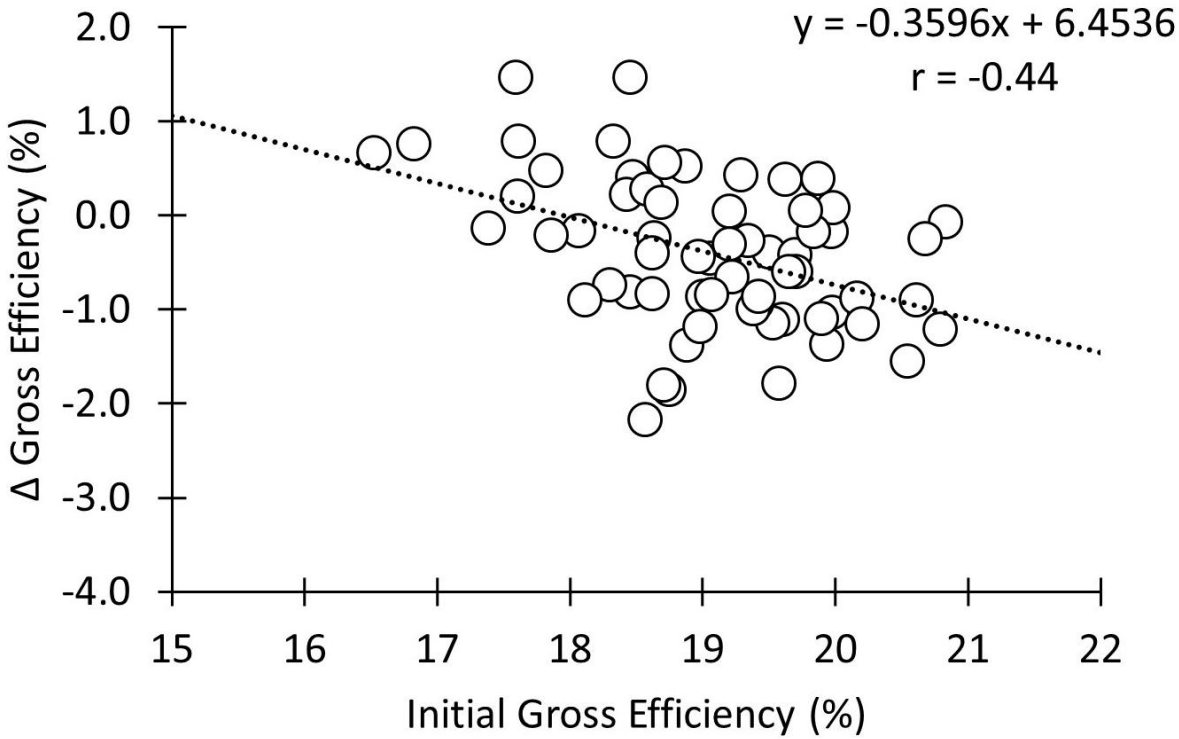
- Regression towards the mean predicts that upon repeated sampling from a normal distribution, extreme values will be less frequent than values close to the mean.
- An extreme value **within** a family will be "replaced" by a less extreme.
- How would the regression line look?

Regression towards the mean



Regression towards the mean

- This poses a problem when analyzing baseline characteristics and change due to training



When using a correction "*. . . to minimize the effect* [of regression to the mean], *the correlations in the present study were weakened.*"

Skovereng, K., et al. (2018). "Effects of Initial Performance, Gross Efficiency and O_{2peak} Characteristics on Subsequent Adaptations to Endurance Training in Competitive Cyclists." Front Physiol 9(713).

Multiple linear regression

- The simple linear regression model...

$$y = \beta_0 + \beta_1 X_1 + \epsilon$$

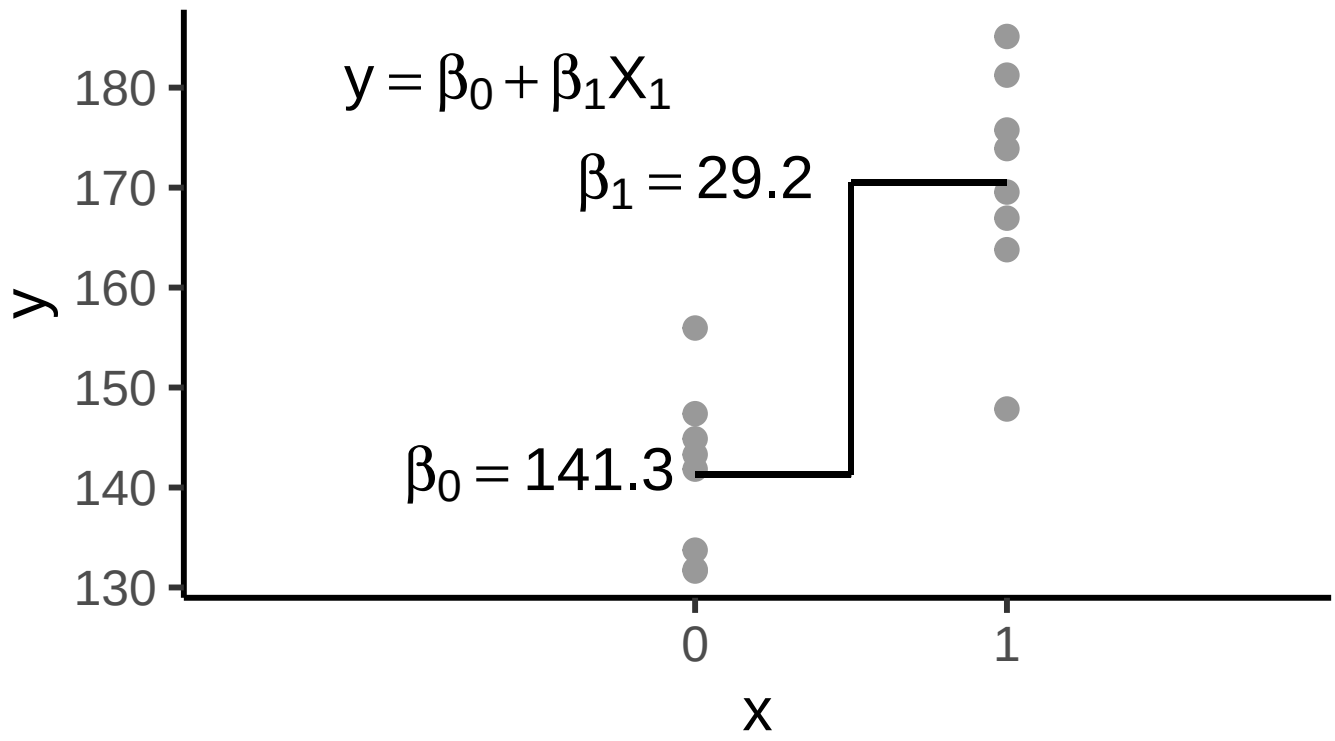
- ... can be extended to include multiple covariates (or independent variables)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... \beta_p X_p + \epsilon$$

- The model can contain continuous covariates and covariates that only take 0 and 1, these are called *dummy variables*

Dummy variables

- A dummy variable can be used in a regression model representing a qualitative variable (e.g. Male and Female) where the first **level** of the variable is **coded** 0 and the second level is **coded** 1
- In the regression model, the numerical coded variable is used, a simple uni-variate example:



Dummy variables

- In the case of Female and Male the dummy variable for sex is coded

if $sex = Female$ then $X = 0$

if $sex = Male$ then $X = 1$

Mean values for women:

$$y = \beta_0 + \beta_1 \times 0 = \beta_0$$

Mean values for men:

$$y = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

Dummy variables can used to code more levels than 2

- Using dummy variables, more **levels** can be coded into the model
- More parameters will have to be estimated, if three groups (A, B and C) are to be included in the model, 3-1 dummy variables are needed

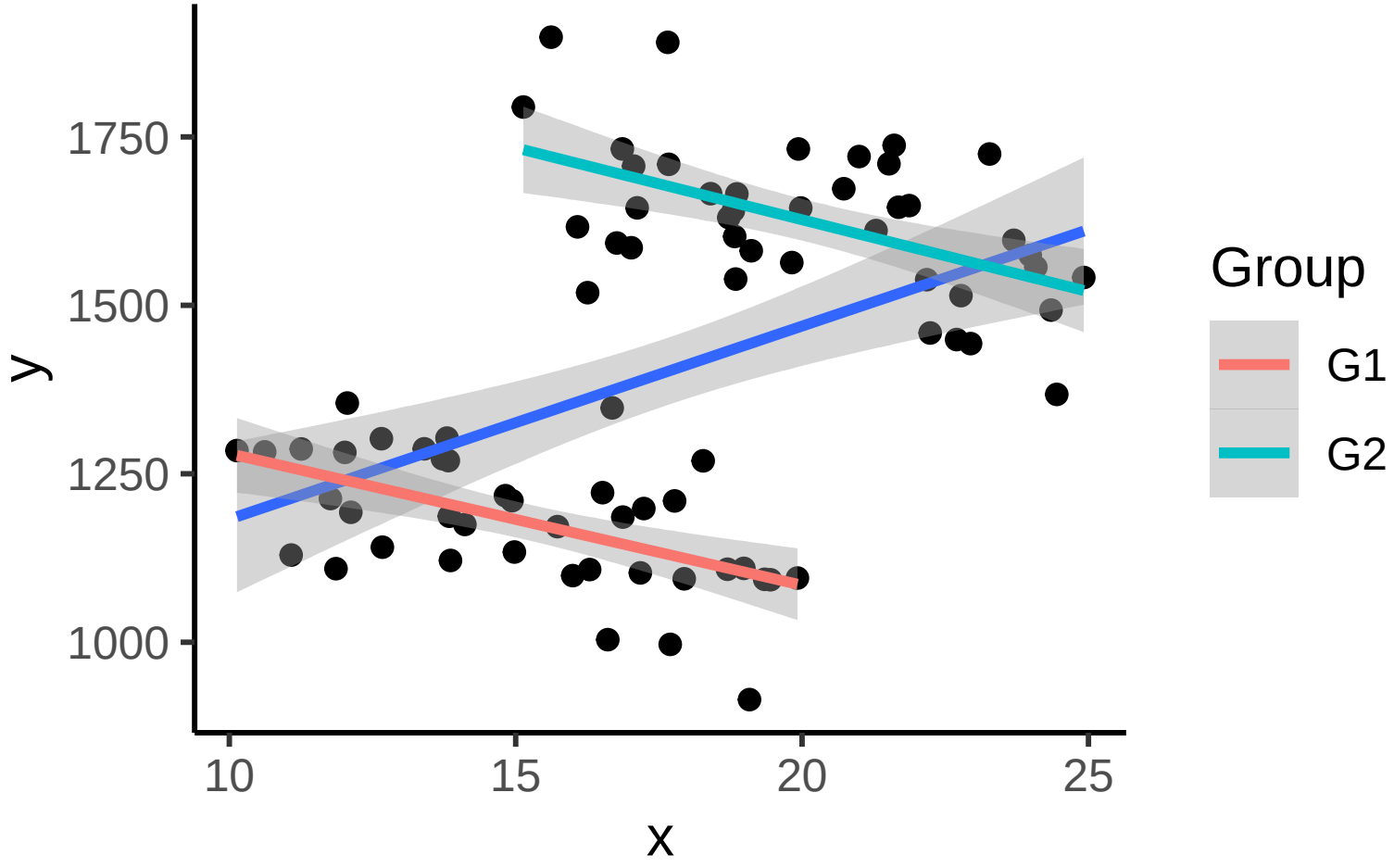
If $group = A$ then $X_1 = 0, X_2 = 0$

If $group = B$ then $X_1 = 1, X_2 = 0$

If $group = C$ then $X_1 = 0, X_2 = 1$

Dummy variables can be used to control for group effects

- Simpson's paradox is when **marginal** and **partial** relationships in the data set have different signs, i.e. a positive relationship in the whole data-set and negative relationships within subgroups



Dummy variables can be used to control for group effects

Table 1: Simple model

	Estimate	Std. Error	t value
(Intercept)	895.83	121.07	7.40
x	28.67	6.71	4.27

Table 2: Controlling for groups

	Estimate	Std. Error	t value
(Intercept)	1489.18	56.69	26.27
x	-20.45	3.62	-5.64
GroupG2	546.72	27.03	20.22