Applied Data Analysis (CS401)

ada

Lecture 5
Regression for disentangling data
18 Oct 2023

EPFL

Robert West

dlab

Announcements

- Homework H1 due end of next week
  - Due Fri 27 Oct 23:59
- Project milestone P1 feedback to be released next week
- Final exam has been scheduled: Tue 16 Jan 2024, 15:15–18:15
- Friday’s lab session:
  - Quiz 4
  - Exercise on regression analysis (in BCH 2201)
  - Homework office hours (on Zoom, in parallel to exercise)
- Indicative course feedback is being collected (until Sun 22 Oct)

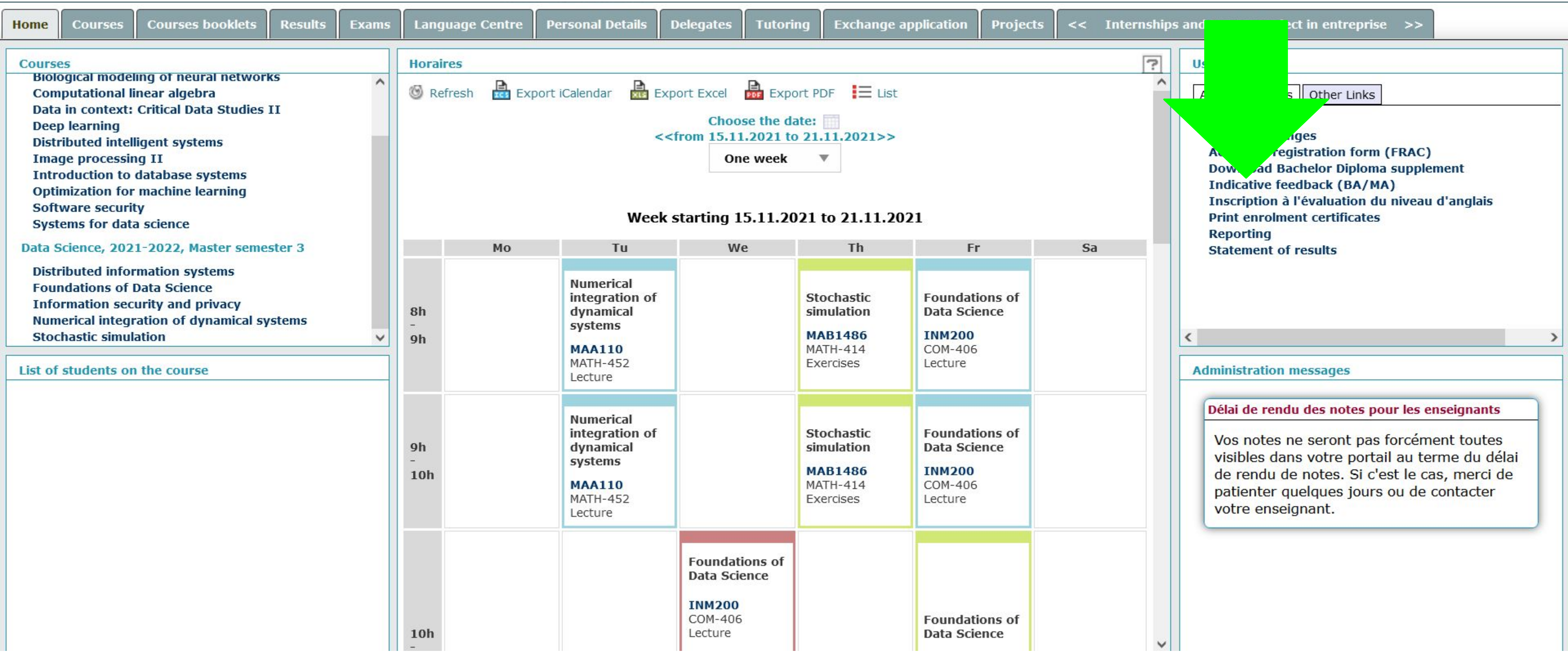
Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec5-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Course eval ("indicative feedback") open until 22 Oct

Go to https://isa.epfl.ch now!



Linear regression

Credits

- Much of the material in this lecture is based on Andrew Gelman and Jennifer Hill’s great book “Data Analysis Using Regression and Multilevel/Hierarchical Models”, available for free here

- For a neat and gentle written intro to linear regression, especially check out chapters 3 and 4

What you should already know about linear regression

POLLING TIME

- **"How familiar are you with linear regression?"**
- Scan QR code or go to https://web.speakup.info/room/join/66626

Linear regression as you know it

- **Given:** n data points (X_i, y_i), where X_i is k-dimensional vector of predictors (a.k.a. features) of i-th data point, and y_i is scalar outcome
- **Goal:** find the optimal coefficient vector $\beta = (\beta_1, \ldots, \beta_k)$ for approximating the y_i's as a linear function of the X_i's:

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n \end{aligned}$$

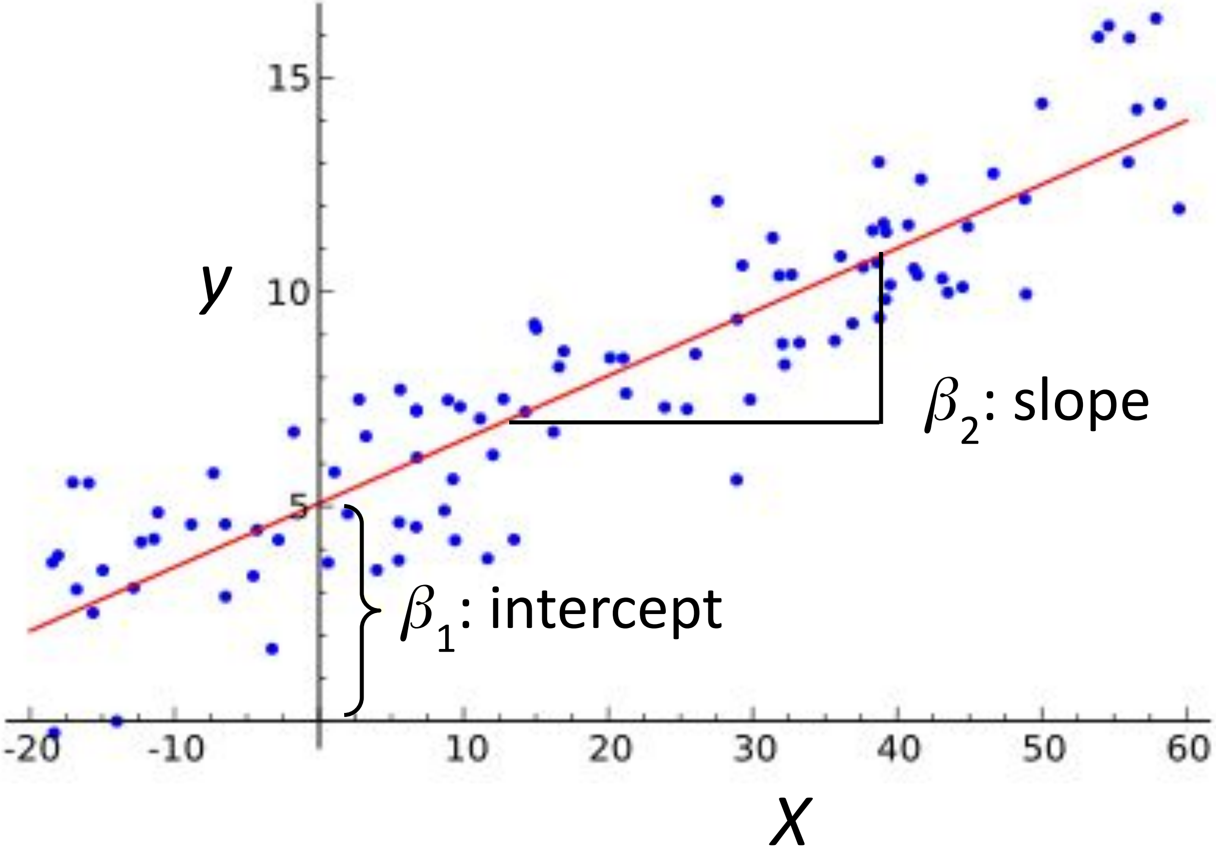
Scalar product (a.k.a. dot product) of 2 vectors

where ϵ_i are error terms that should be as small as possible

- X_{i1} usually the constant 1 (by def) ⇒ β_1 a constant intercept

Example with one predictor

$$y \approx \beta_1 + \beta_2 X$$



Linear regression as you know it

- **Given:** n data points (X_i, y_i), where X_i is k-dimensional vector of predictors (a.k.a. features), and y_i is scalar outcome, of i-th data point
- **Goal:** find the optimal coefficient vector $\beta = (\beta_1, \ldots, \beta_k)$ for approximating the y's as a linear function of the X's:

?

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n \end{aligned}$$

where ϵ_i are error terms that should be as small as possible

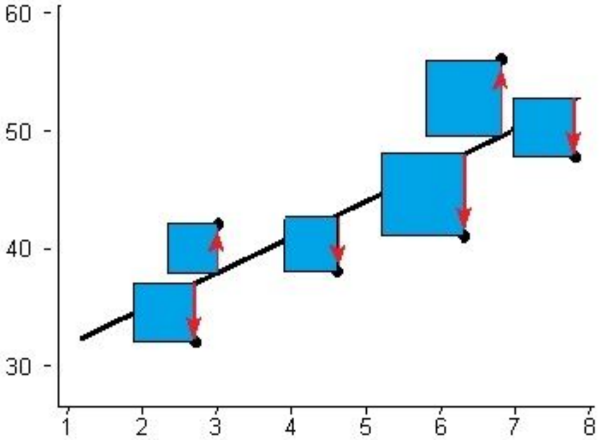
- X_{i1} usually the constant 1 $\longrightarrow$ β_1 a constant intercept

Optimality criterion: least squares

$$y_i \quad = \quad X_i\beta + \epsilon_i \qquad \text{for } i = 1, \ldots, n$$

- Intuitively, want errors ϵ_i to be as small as possible
- Technically, want sum of squared errors as small as possible
  ⇔ find $\hat{\beta}$ such that we minimize

$$\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^2$$



Use cases of regression

- **Prediction:** use fitted model to estimate outcome *y* for a new *X* not seen during model fitting (if you've seen regression before, then probably in the context of prediction)
- **Descriptive data analysis:** compare average outcomes across subgroups of data (today!)
- **Causal modeling:** understand how outcome *y* changes when you manipulate predictors *X* (next lecture is about causality, although not primarily using regression)

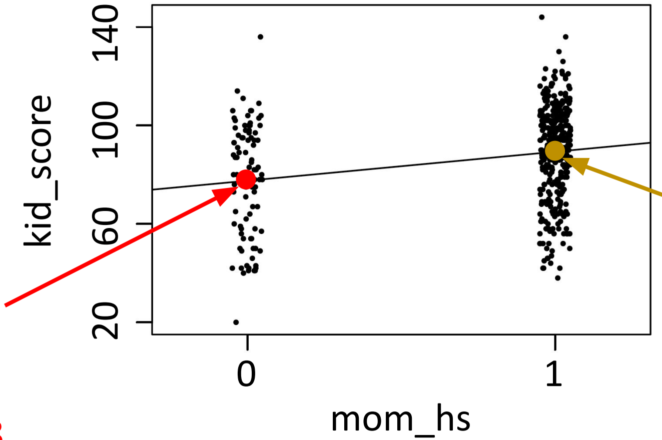
Regression as comparison of average outcomes

Example with one binary predictor X_i

- X_i = mom_hs = "Did mother finish high school?" $\in$ {0, 1} (No = 0, Yes = 1)
- y_i = kid_score = child's score on cognitive test $\in$ [0, 140]

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

kid_score = 78 + 12 · mom_hs + error



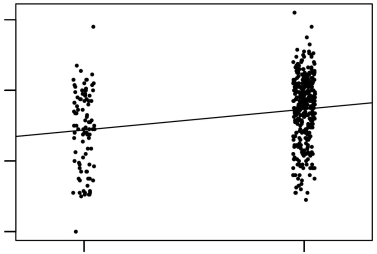
mean kid_score for moms who **didn't finish** high school: 78

mean kid_score for moms who **finished** high school: 78 + 12 = 90

One binary predictor X_i: Interpretation of fitted parameters β

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

- **Intercept** β_1: mean outcome for data points i with $X_i = 0$
- **Slope** β_2: difference in mean outcomes between data points with $X_i = 1$ and data points with $X_i = 0$
- Reason: means minimize least-squares criterion: $\sum^n_{i=1} (y_i - m)^2$ is minimized w.r.t. m when $-2 \sum^n_{i=1} (y_i - m) = 0$, i.e., when $m = (1/n) \sum^n_{i=1} y_i$



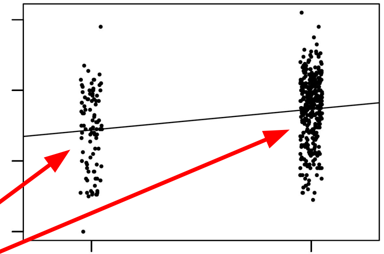
One binary predictor X_i: Interpretation of fitted parameters β

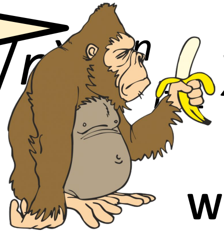
$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

- **Intercept** β_1: mean outcome for data points *i* with $X_i = 0$
- **Slope** β_2: difference in mean outcomes between data points with $X_i = 1$ and data points with $X_i = 0$
- Reason: means minimize least-squares criterion:

  $\sum^n_{i=1} (y_i$ … $)^2$ … n when

  $-2 \sum^n_{i=1}$ … $(1/n)$ … y_i





What a mean monkey!

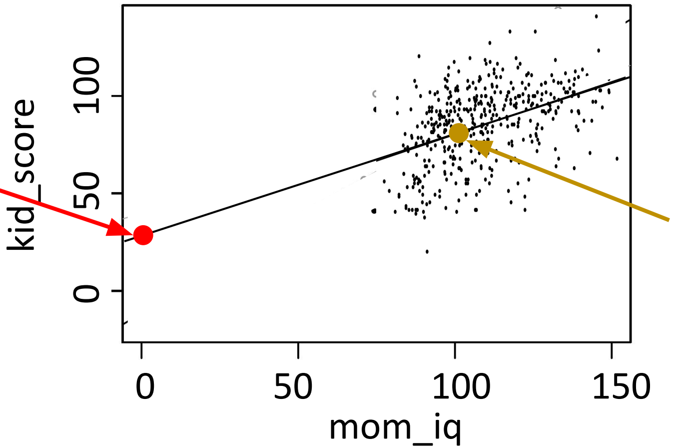
Example with one continuous predictor X_i

- X_i = mom_iq = mother's IQ score $\in$ [70, 140]
- y_i = kid_score = child's score on cognitive test $\in$ [0, 140]

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

kid_score = 26 + 0.6 · mom_iq + error

estimated (hypothetical) mean kid_score for moms with IQ = 0: 26

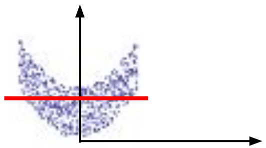


estimated mean kid_score for moms with IQ = 100: 26 + 0.6 · 100 = 86

One continuous predictor X_i: Interpretation of fitted parameters β

$$y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

- **Intercept** β_1: estimated mean outcome for data points *i* with $X_i = 0$
- **Slope** β_2: difference in estimated mean outcomes between data points whose X_i's differ by 1
- Why "estimated"? $\longrightarrow$ e.g.,



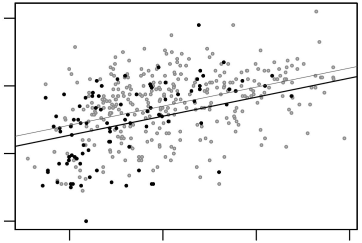
- NB: for binary predictor, we got "exact" instead of "estimated"

Example with multiple predictors

- (X_{i1} = 1 = constant)
- X_{i2} = mom_hs = "Did mother finish high school?" $\in$ {0, 1} (No Yes)
- X_{i3} = mom_iq = mother's IQ score $\in$ [70, 140]
- y_i = kid_score = child's score on cognitive test $\in$ [0, 140]

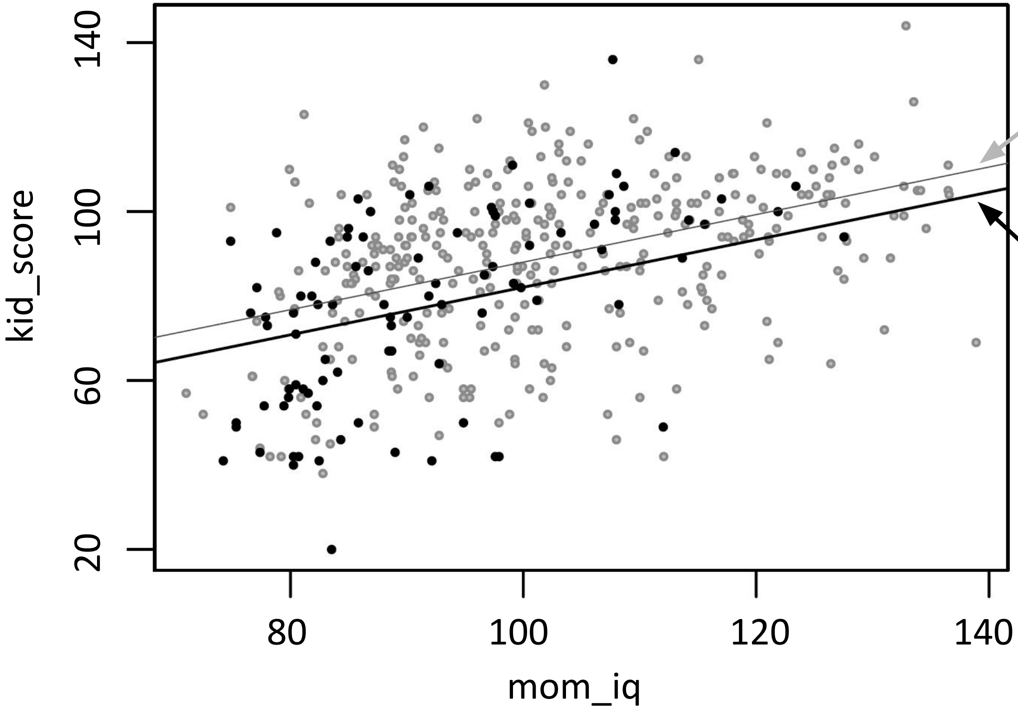
$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

kid_score = 26 + 6 · mom_hs + 0.6 · mom_iq + error



Example with multiple predictors

kid_score = 26 + 6 · mom_hs + 0.6 · mom_iq + error



kids of moms who **finished** high school:
intercept = 26 + 6 = 32
slope = 0.6

kids of moms who **didn't finish** high school:
intercept = 26
slope = 0.6

Example with interaction of predictors

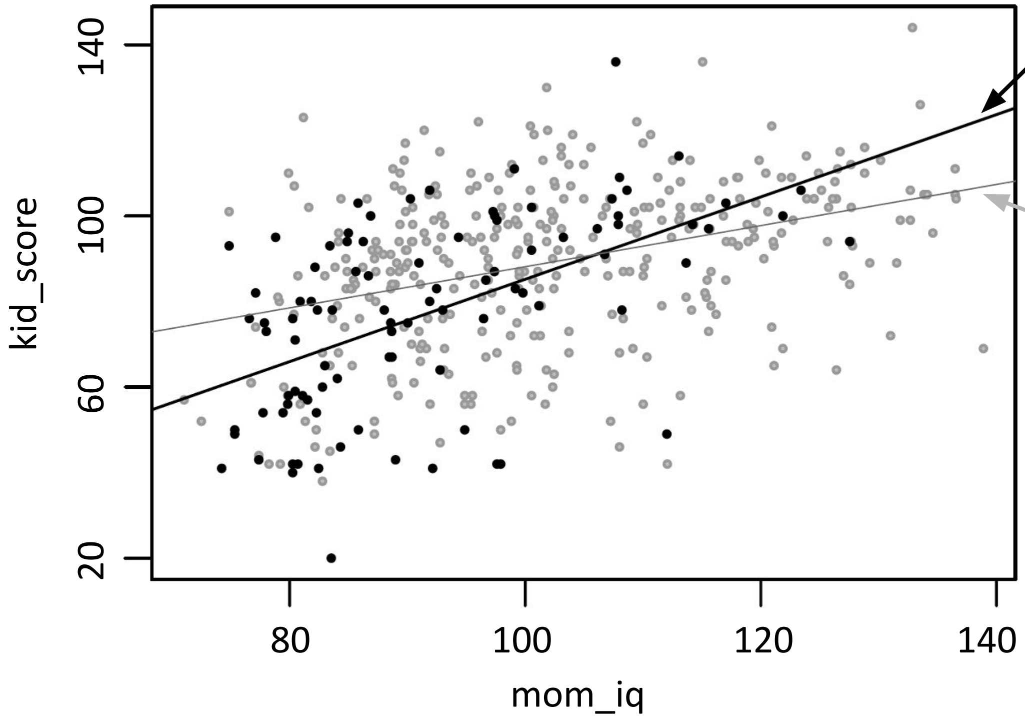
- X_{i2} = mom_hs = "Did mother finish high school?" $\in \{0, 1\}$ (No = 0, Yes = 1)
- X_{i3} = mom_iq = mother's IQ score $\in [70, 140]$
- y_i = kid_score = child's score on cognitive test $\in [0, 140]$

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i2} X_{i3} + \epsilon_i$$

kid_score = −11 + 51 · mom_hs + 1.1 · mom_iq − 0.5 · mom_hs · mom_iq + error

Example with interaction of predictors

$$\text{kid\_score} = -11 + 51 \cdot \text{mom\_hs} + 1.1 \cdot \text{mom\_iq} - 0.5 \cdot \text{mom\_hs} \cdot \text{mom\_iq} + \text{error}$$



kids of moms who **didn't finish** high school:

intercept = −11

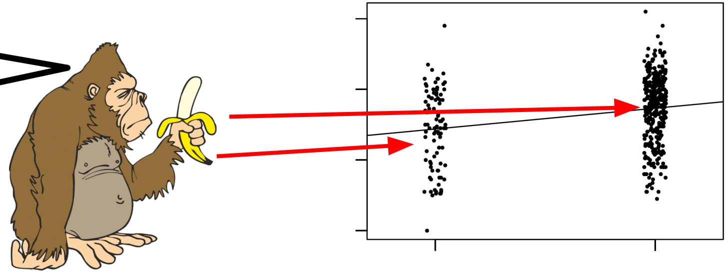
slope = 1.1

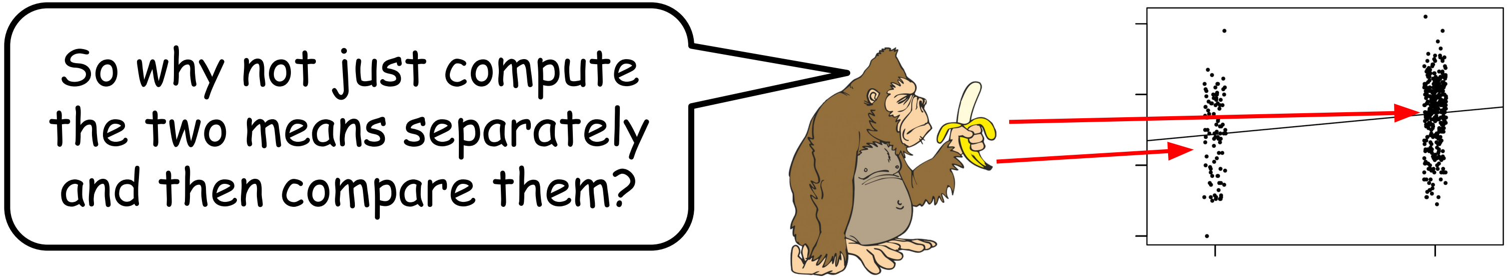
kids of moms who **finished** high school:

intercept = −11 + 51 = 40

slope = 1.1 − 0.5 = 0.6

So why not just compute the two means separately and then compare them?





	Mom drives Mercedes	Mom doesn't drive Mercedes
Mom finished high school	avg kid_score 90	avg kid_score 90
Mom didn't finish high school	avg kid_score 78	avg kid_score 78

	Mom drives Mercedes	Mom doesn't drive Mercedes
Mom finished high school	990 women	10 women
Mom didn't finish high school	10 women	990 women

	Mom drives Mercedes	Mom doesn't drive Mercedes
Mom finished high school	avg kid_score 90	avg kid_score 90
Mom didn't finish high school	avg kid_score 78	avg kid_score 78

	Mom drives Mercedes	Mom doesn't drive Mercedes
Mom finished high school	990 women	10 women
Mom didn't finish high school	10 women	990 women

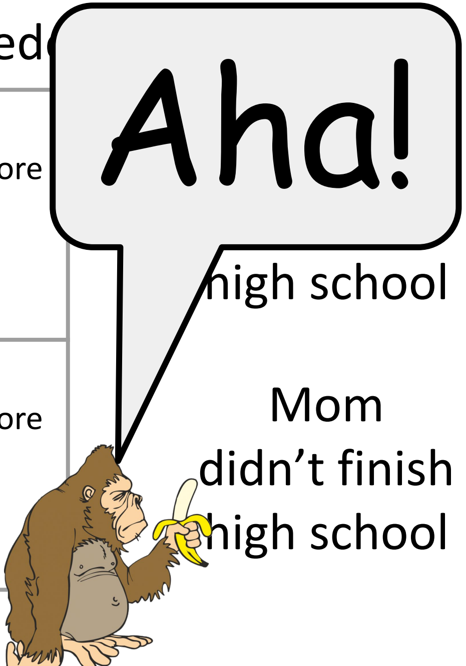
THINK FOR A MINUTE:

What is the mean outcome for Mercedes-driving moms vs. for non-Mercedes-driving moms? Compare the two means! What does the comparison tell you about the link between Mercedes-driving and kid_score?

(Feel free to discuss with your neighbor.)

- Mean kid_score for Mercedes drivers: **0.99 · 90 + 0.01 · 78 ≈ 90**
- Mean kid_score for non-Mercedes drivers: **0.01 · 90 + 0.99 · 78 ≈ 78**
- But really driving Mercedes makes no difference (for fixed high-school predictor)!
- Root of evil: **correlation** between finishing high school and driving Mercedes
- **Regression** to the rescue: **kid_score = 78 + 12 · mom_hs + 0 · mercedes + error**

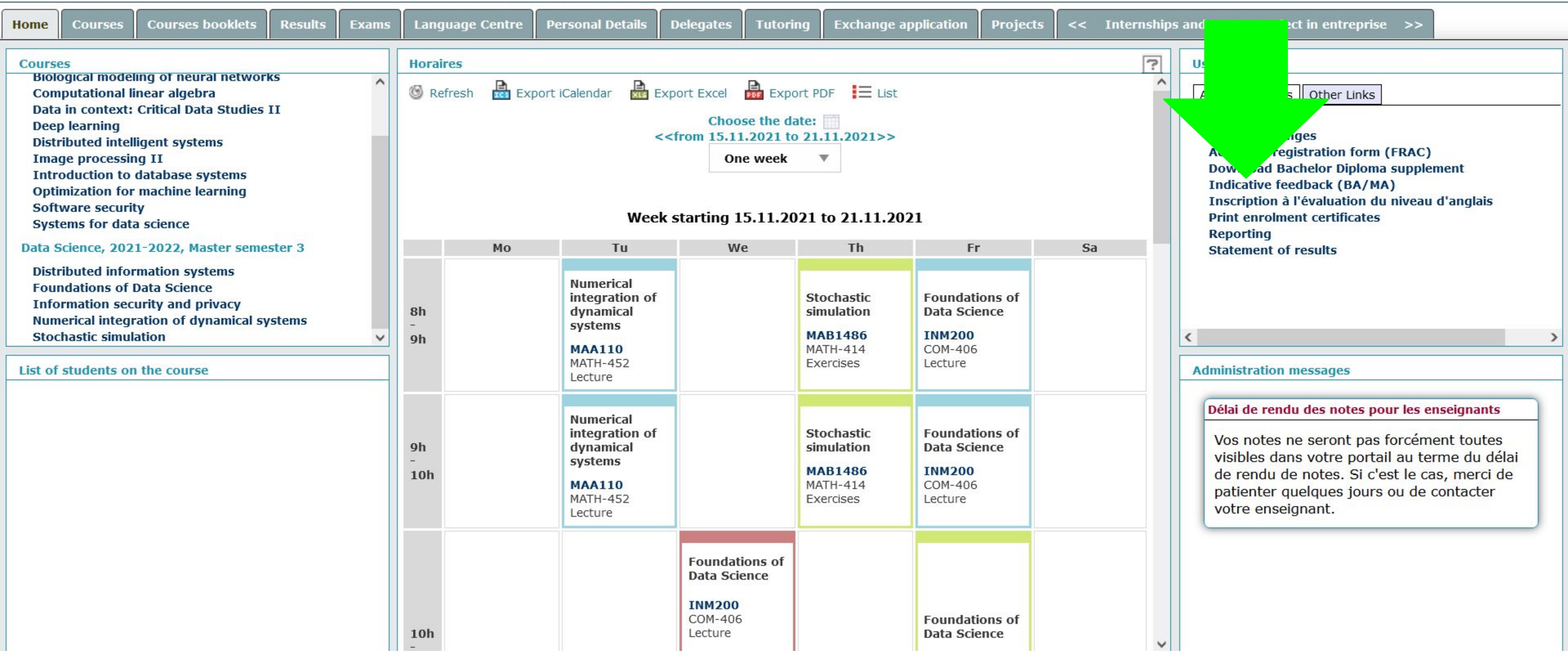
	Mercedes	No Merced
Mom finished high school	mean kid_score 90	mean kid_score 90
Mom didn't finish high school	mean kid_score 78	mean kid_score 78



	Mercedes	No Mercedes
high school	990 women	10 women
Mom didn't finish high school	10 women	990 women

Course eval ("indicative feedback") open until 22 Oct

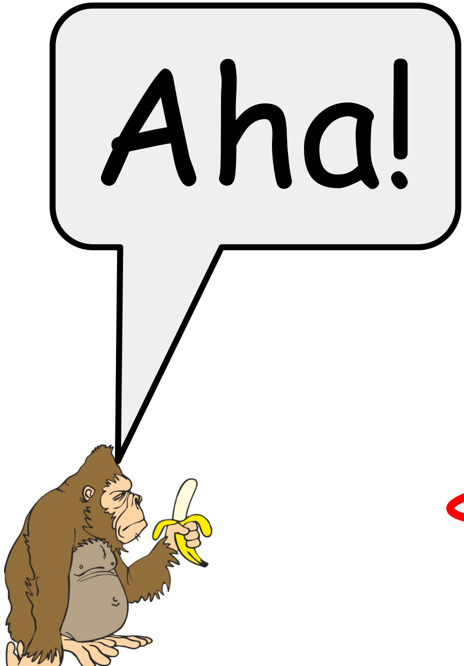
Go to https://isa.epfl.ch now!



Quantifying uncertainty

Quantifying uncertainty

- Statistical software gives you more than just coefficients β:



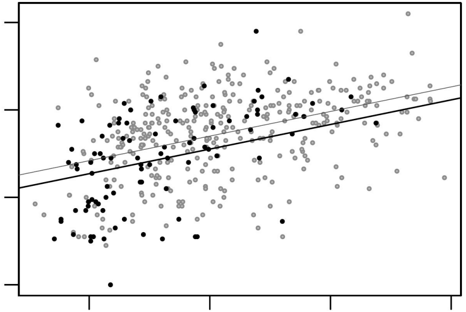

```
Residuals:
    Min      1Q  Median      3Q     Max
-52.873 -12.663   2.404  11.356  49.545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
mom.hs       5.95012    2.21181   2.690  0.00742 **
mom.iq       0.56391    0.06057   9.309  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.14 on 431 degrees of freedom
Multiple R-Squared: 0.2141,     Adjusted R-squared: 0.2105
F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```

p-value: probability of estimating such an extreme coefficient if the true coefficient were zero (= null hypothesis)

Residuals and R^2

- **Residual** for data point *i*: estimation error on data point *i*:

$$r_i = y_i - X_i\hat{\beta}$$



- Mean of residuals = 0
  (total overestimation = total underestimation)
- Variance of residuals
  = avg squared distance of predicted value from observed value
  = "unexplained variance"
- Fraction of variance explained by the model:

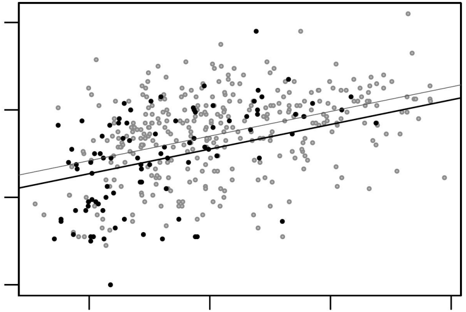
$$R^2 = 1 - \hat{\sigma}^2 / s_y^2$$

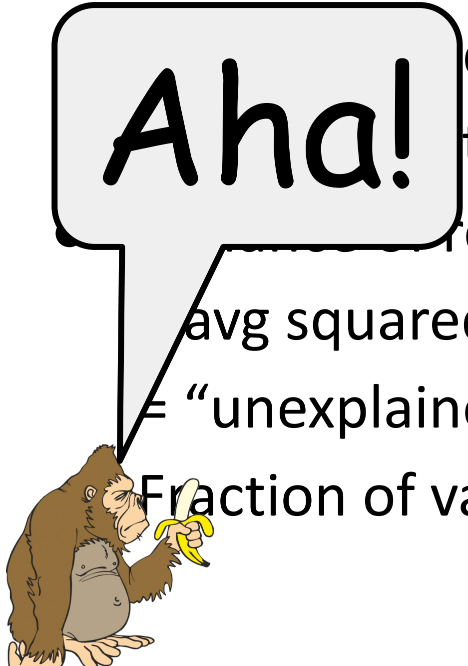
Variance of outcomes *y*

Residuals and R^2

- **Residual** for data point *i*: estimation error on data point *i*:

$$r_i = y_i - X_i\hat{\beta}$$





duals = 0

timation = total underestimation)

residuals

avg squared distance of predicted value from observed value

= “unexplained variance”

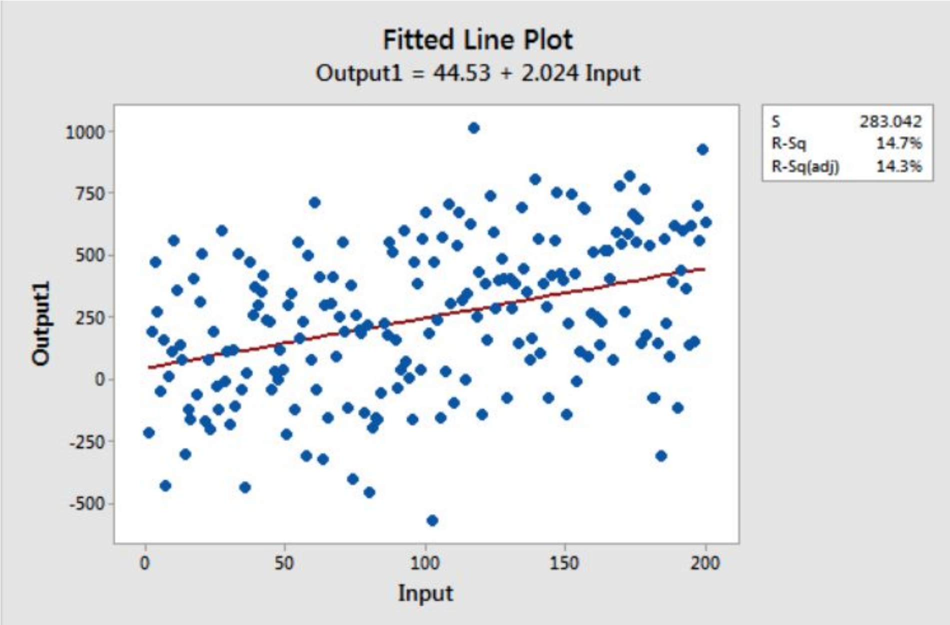
Fraction of variance explained by the model:

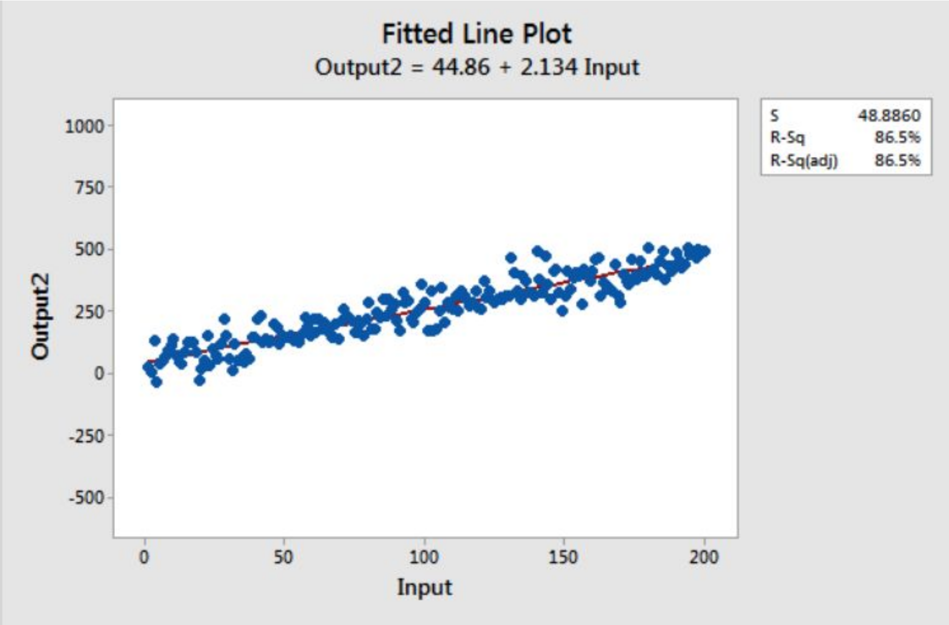
$$R^2 = 1 - \hat{\sigma}^2 / s_y^2$$

Variance of outcomes *y*

Coefficient of determination: R^2

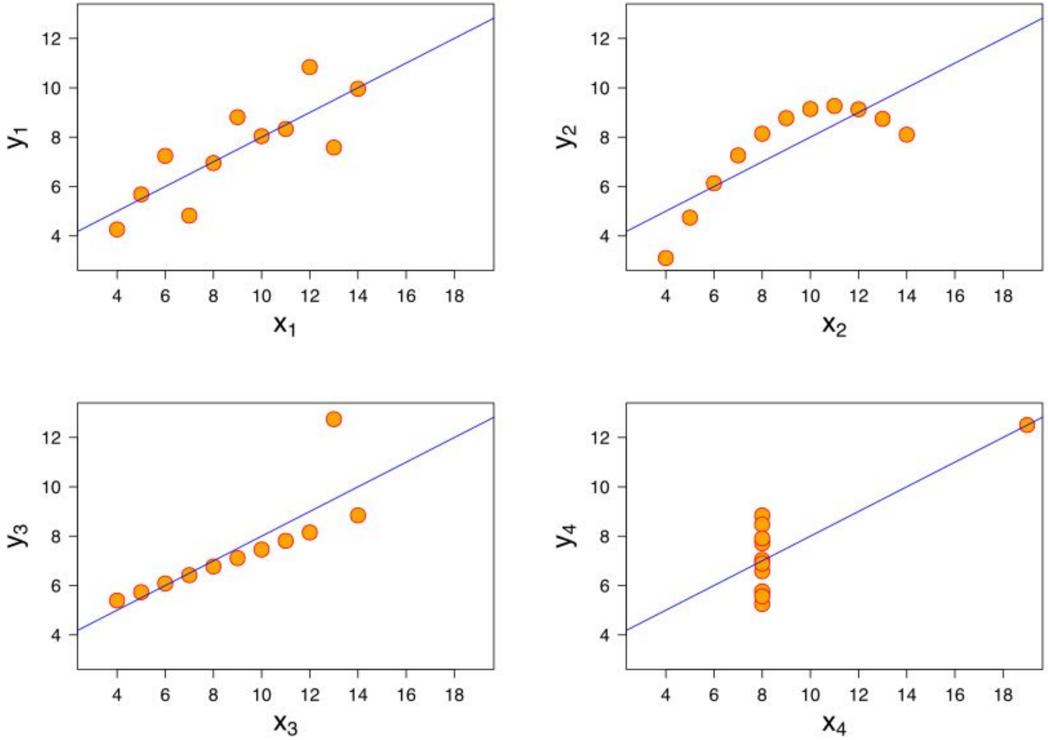
$$R^2 = 1 - \hat{\sigma}^2 / s_y^2$$



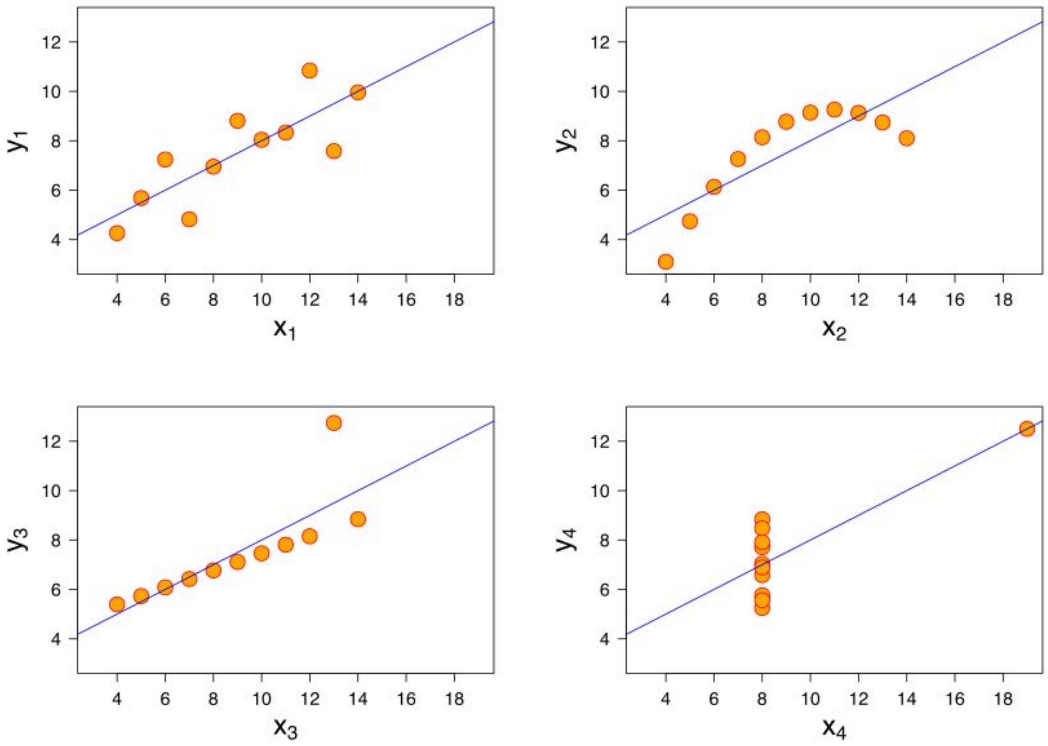


$R^2 = 0.147$ $R^2 = 0.865$

Coefficient of determination: R^2



Coefficient of determination: R^2



R^2 = 0.67 everywhere!

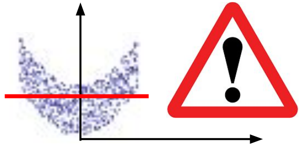
Assumptions made in regression modeling

Assumptions for regression modeling

1. Validity:
    a. Outcome measure should accurately reflect the phenomenon of interest
    b. Model should include all relevant predictors
    c. Model should generalize to cases to which it will be applied

Assumptions for regression modeling (2)

2. Linearity:



$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n \end{aligned}$$

But very flexible: we require linearity in *predictors* (not necessarily in raw inputs); predictors can be arbitrary functions of raw inputs, e.g.,

- logarithms, polynomials, reciprocals, …
- interactions (i.e., products) of multiple inputs
- discretization of raw inputs, coded as indicator variables

Assumptions for regression modeling (3)

3. Independence of errors: no interaction between data points
4. Equal variance of errors
5. Normality (Gaussianity) of errors

less important in practice

Transformations of predictors and outcomes

Transformations of predictors

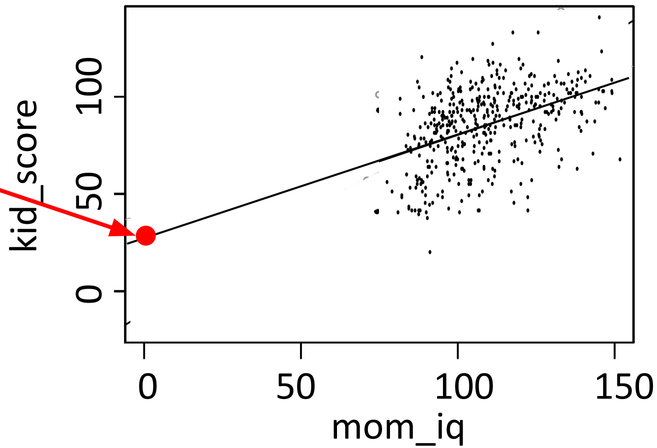
- When we apply linear transformations to predictors, the model remains "equally good":
  - The fitted coefficients may change, but predicted outcomes and model fit (R^2) won't change
- For instance,

$$
\begin{aligned}
\text{earnings} &= -61000 + 51 \cdot \text{height (in millimeters)} + \text{error} \\
\text{earnings} &= -61000 + 81000000 \cdot \text{height (in miles)} + \text{error.}
\end{aligned}
$$

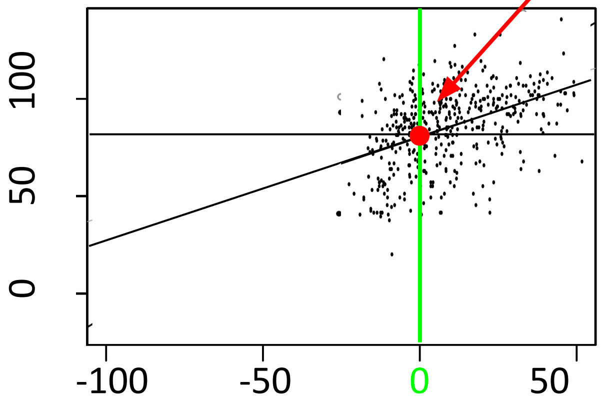
Mean-centering of predictors

- Compute the mean value of a predictor over all data points, and subtract it from each value of that predictor:
  $X_{ij} \leftarrow X_{ij} - \text{mean}(X_{1j}, \ldots, X_{nj})$
- $\Rightarrow$ the predictor X_{ij} now has mean 0

(hypothetical) mean kid_score for moms with IQ = 0: 26



mean kid_score for moms with mean IQ: 86



After mean-centering of predictors, …

… you have a convenient interpretation of coefficients β_j of main predictors (i.e., non-interaction predictors):

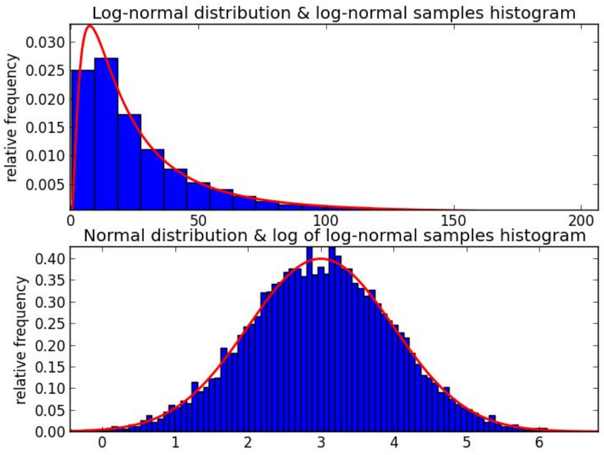
- j = 1 (i.e., intercept):
  - Estimated mean outcome when each predictor has its mean value
- j > 1:
  - Model w/o interactions: estimated mean increase in outcome y for each unit increase in X_{ij}
  - Model with interactions: estimated mean increase in outcome y for each unit increase in X_{ij} **when each other predictor has its mean value**

Standardization via z-scores

- First **mean-center** all predictors, then **divide them by their standard deviations:**
  $X_{ij} \leftarrow [X_{ij} - \text{mean}(X_{1j}, \ldots, X_{nj})] \,/\, \text{sd}(X_{1j}, \ldots, X_{nj})$
- Resulting values are called "**z-scores**"
- All predictors now have the same units:
  distance (in terms of standard deviations) from the mean
- This lets us compare coefficients for predictors with previously incomparable units of measurement, e.g., IQ score vs. earnings in Swiss francs vs. height in centimeters

Logarithmic outcomes

- **Practical:** makes sense if the outcome *y* follows a heavy-tailed distribution
- Only works for non-negative outcomes
- **Theoretical:** turns an additive model into a **multiplicative model:**



$$\log y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i$$

Exponentiating both sides yields

$$\begin{aligned} y_i &= e^{b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i} \\ &= B_0 \cdot B_1^{X_{i1}} \cdot B_2^{X_{i2}} \cdots E_i \end{aligned}$$

Logarithmic outcomes: Interpreting coefficients

$$\begin{aligned} y_i &= e^{b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i} \\ &= B_0 \cdot B_1^{X_{i1}} \cdot B_2^{X_{i2}} \cdots E_i \end{aligned}$$

- An **additive** increase of 1 in predictor $X_{\cdot 1}$ is associated with a **multiplicative** increase of B_1 := exp(b_1) in the outcome
- If $b_1 \approx 0$, we can immediately interpret b_1 (without needing to exponentiate it first to get B_1!) as the **relative increase** in outcomes, since exp(b_1) ≈ 1 + b_1
- E.g., $b_1 = 0.05 \Rightarrow B_1 = \exp(b_1) \approx 1.05$
  ⇒ "+1 in predictor $X_{\cdot 1}$" is associated with "+5% in outcome"

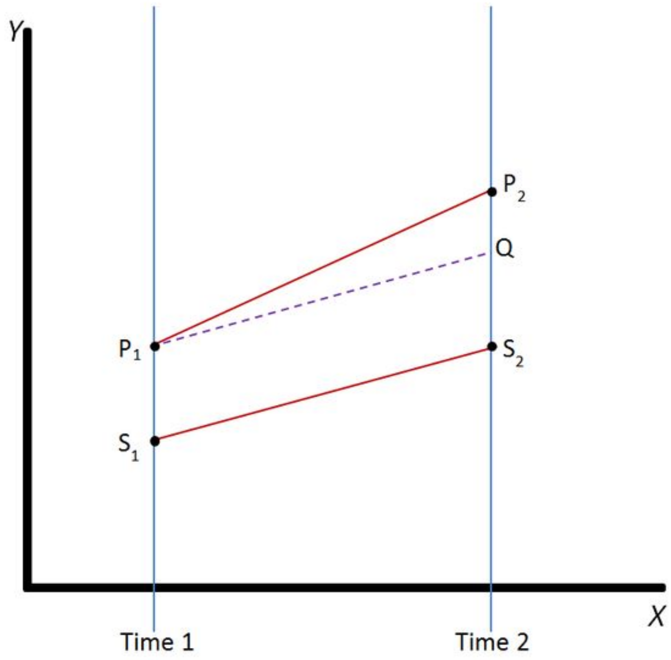
Going beyond linear regression for comparing means

Beyond linear regression: generalized linear models

- Logistic regression: binary outcomes
- Poisson regression: non-negative integer outcomes (e.g., counts)

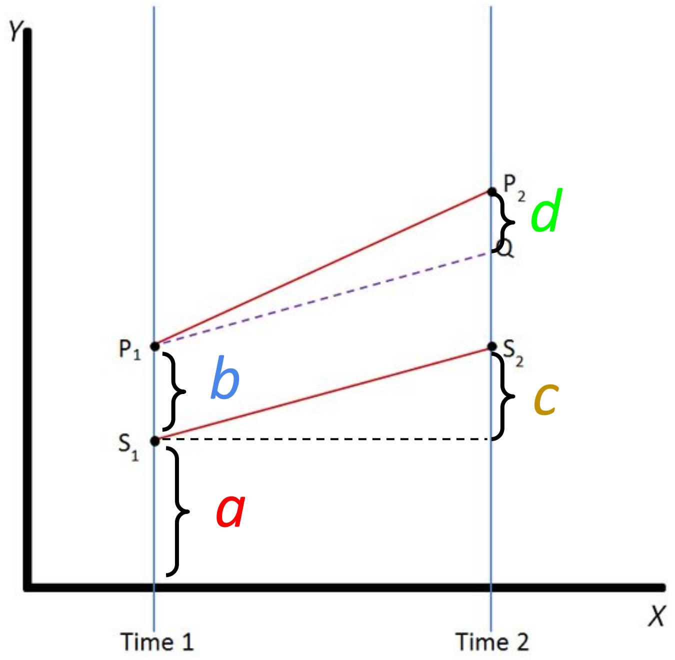
Beyond comparing means; or, A taste of causality: “Difference in differences”

- Two groups: *P*, *S*
- At time 2, group *P* receives a **treatment**, group *S* doesn’t
- Question: Did the treatment have an **effect?** If so, how large was it?
- *P* and *S* don’t start out the same at time 1
- There is a temporal “baseline effect” even w/o treatment

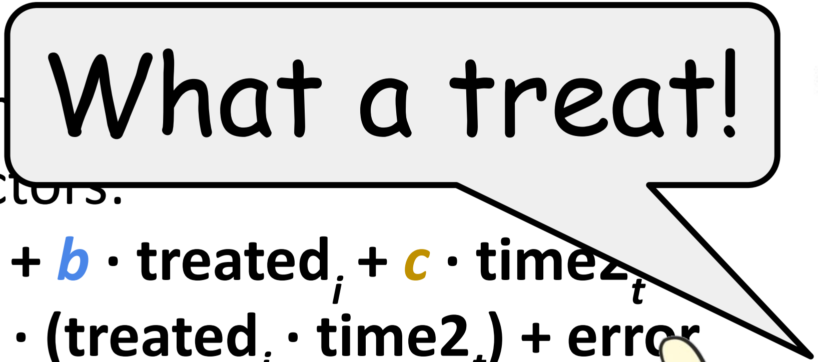


Beyond comparing means; or, A taste of causality: "Difference in differences" (2)

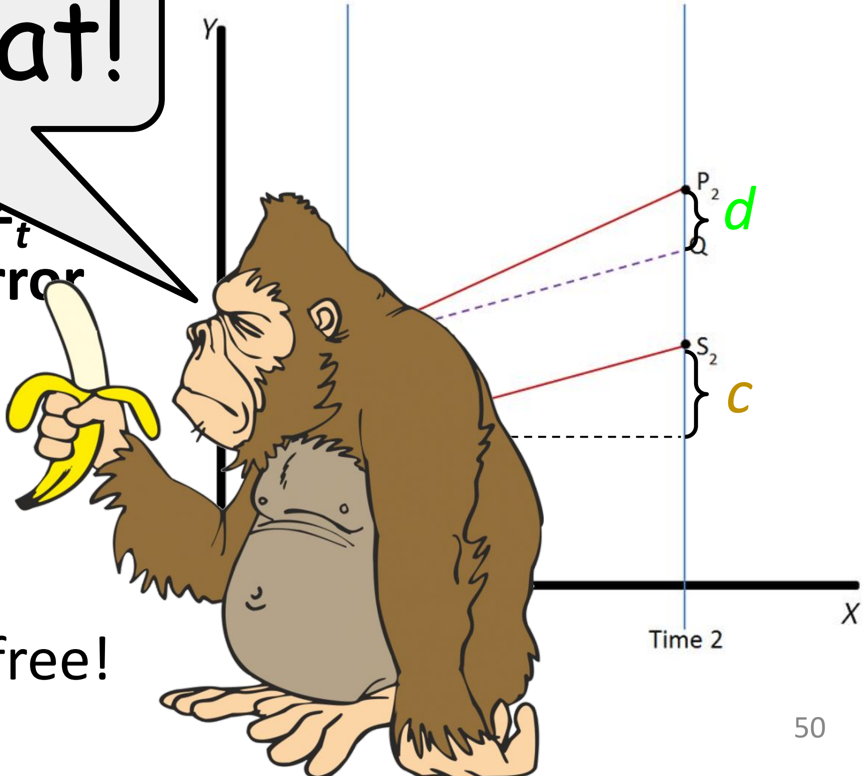
- Elegant linear model with binary predictors:
  $\boldsymbol{y_{it} = a + b \cdot \text{treated}_i + c \cdot \text{time2}_t}$
  $\boldsymbol{+ d \cdot (\text{treated}_i \cdot \text{time2}_t) + \text{error}_i}$
- d = treatment effect
- All of this with one single regression!
- You get quantification of uncertainty (significance) for free!



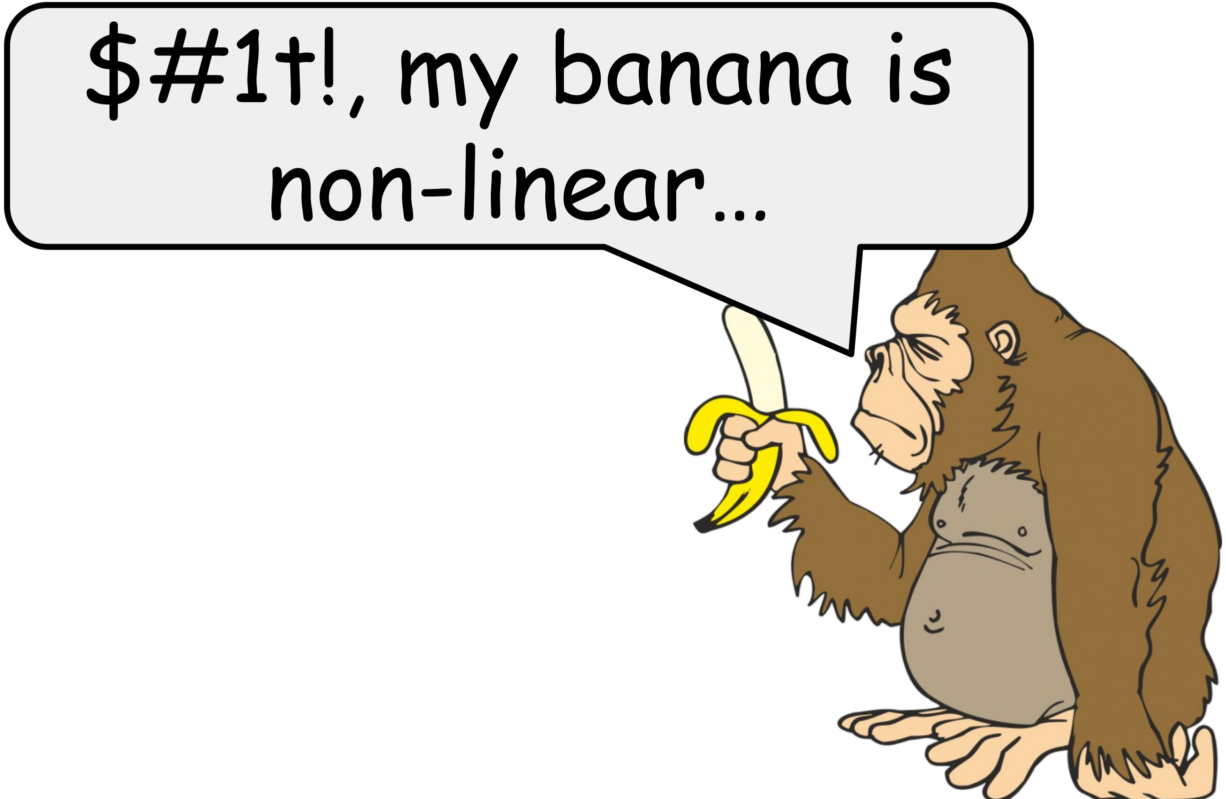
Beyond comparing means; or, A taste of causality: "Difference in differences" (2)



- Elegan predictors:
  $y_{it} = a + b \cdot \textbf{treated}_i + c \cdot \textbf{time2}_t$
  $+\ d \cdot (\textbf{treated}_i \cdot \textbf{time2}_t) + \textbf{error}$
- d = treatment effect
- All of this with one single regression!
- You get quantification of uncertainty (significance) for free!



A bonanza of causality:
Next lecture!

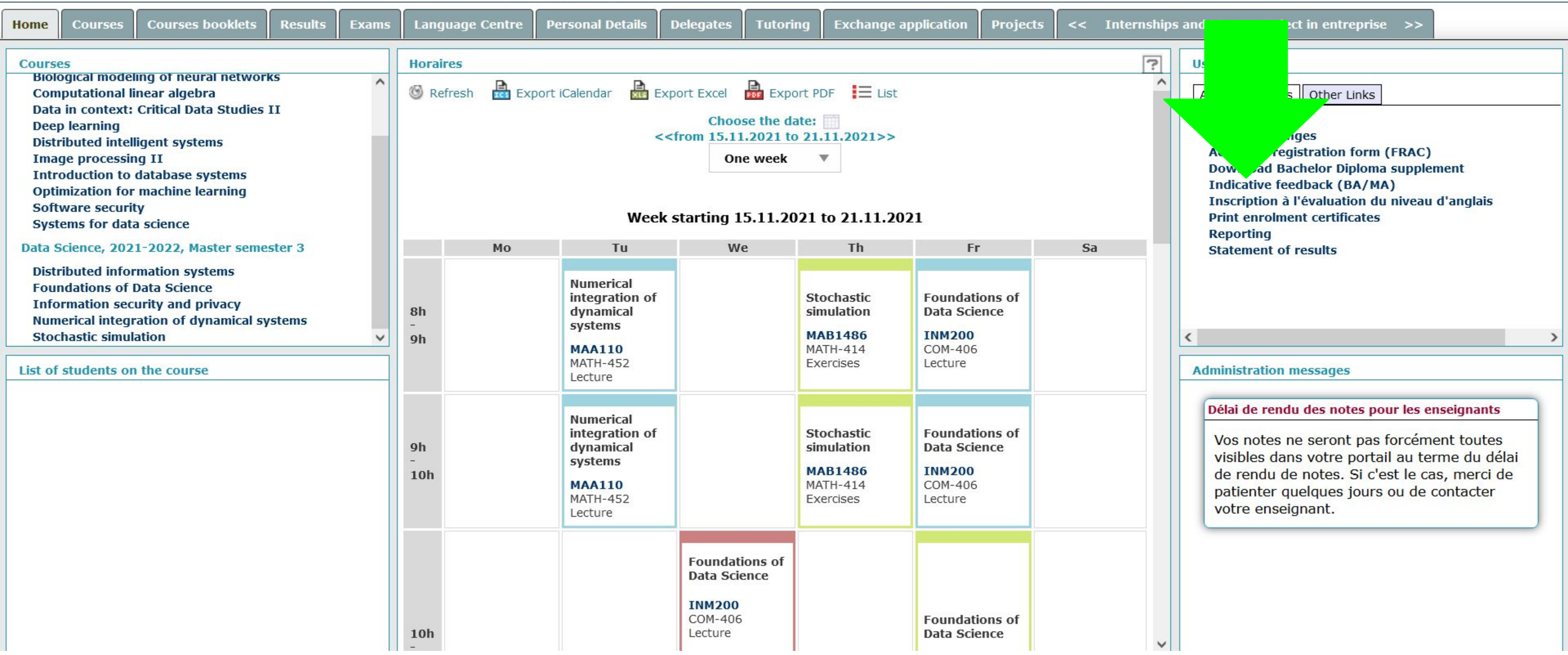


Summary

- Linear regression as a tool for comparing means across subgroups of data
- How? Read group means off from fitted coefficients
- Advantages over plain comparison of means “by hand”:
  - Accounting for correlations among predictors
  - Quantification of uncertainty (significance) “for free”
  - Additive vs. multiplicative model: all it takes is a log
- Caveat emptor:
  - Model must be appropriately specified, else nonsense results $\longrightarrow$ stay critical, run diagnostics (e.g., R^2, data viz)

Course eval ("indicative feedback") open until 22 Oct

Go to https://isa.epfl.ch now!



Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2023-lec5-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Credits

- Much of the material in this lecture is based on Andrew Gelman and Jennifer Hill's great book "Data Analysis Using Regression and Multilevel/Hierarchical Models", available for free here

- For a neat and gentle written intro to linear regression, especially check out chapters 3 and 4

Bonus: Logarithmic outcomes and predictors

Interpretation of coefficient of logarithmic predictor:

- **Multiplicative** increase by 1% in predictor $X_{\cdot 1}$ is associated with a **multiplicative** increase by b_1% in the outcome
- Why?
  - $\log(y) = a + b \log(X) \Rightarrow y = \exp(a) * X^b$
  - Multiplying X by a factor c multiplies y by a factor of c^b
  - $c^b \approx 1 + b*(c-1)$ for $c \approx 1$ (hint: Taylor approximation!)
  - Example when using $c = 1.01$ (i.e., increase by 1%):
    $b = 2 \Rightarrow$ increasing X by 1% increases y by 2%