$x \rightarrow y$

Regression recap I

Revisiting aspects of regression models that we will need for CFA and SEM

Regression recap I

Today's goal:

Go over methods that we learned in M&E I that will be useful for understanding CFA and SEM.

Outline:

- Covariance and correlation
- Linear regression
- Non-linear regression
- Bootstrapping

Covariance and correlation

An explanation

Covariance matrix

	A	B	C	D	E	F
A	2.25	2.96	0.85	1.68	0.88	0.31
B	2.96	7.29	1.71	3.12	1.04	0.52
C	0.85	1.71	0.64	0.77	0.32	0.17
D	1.68	3.12	0.77	10.89	2.93	1.60
E	0.88	1.04	0.32	2.93	1.44	0.54
F	0.31	0.52	0.17	1.60	0.54	0.36

Correlation matrix

	A	B	C	D	E	F
A	1.00	0.73	0.71	0.34	0.49	0.34
B	0.73	1.00	0.79	0.35	0.32	0.32
C	0.71	0.79	1.00	0.29	0.33	0.35
D	0.34	0.35	0.29	1.00	0.74	0.81
E	0.49	0.32	0.33	0.74	1.00	0.75
F	0.34	0.32	0.35	0.81	0.75	1.00

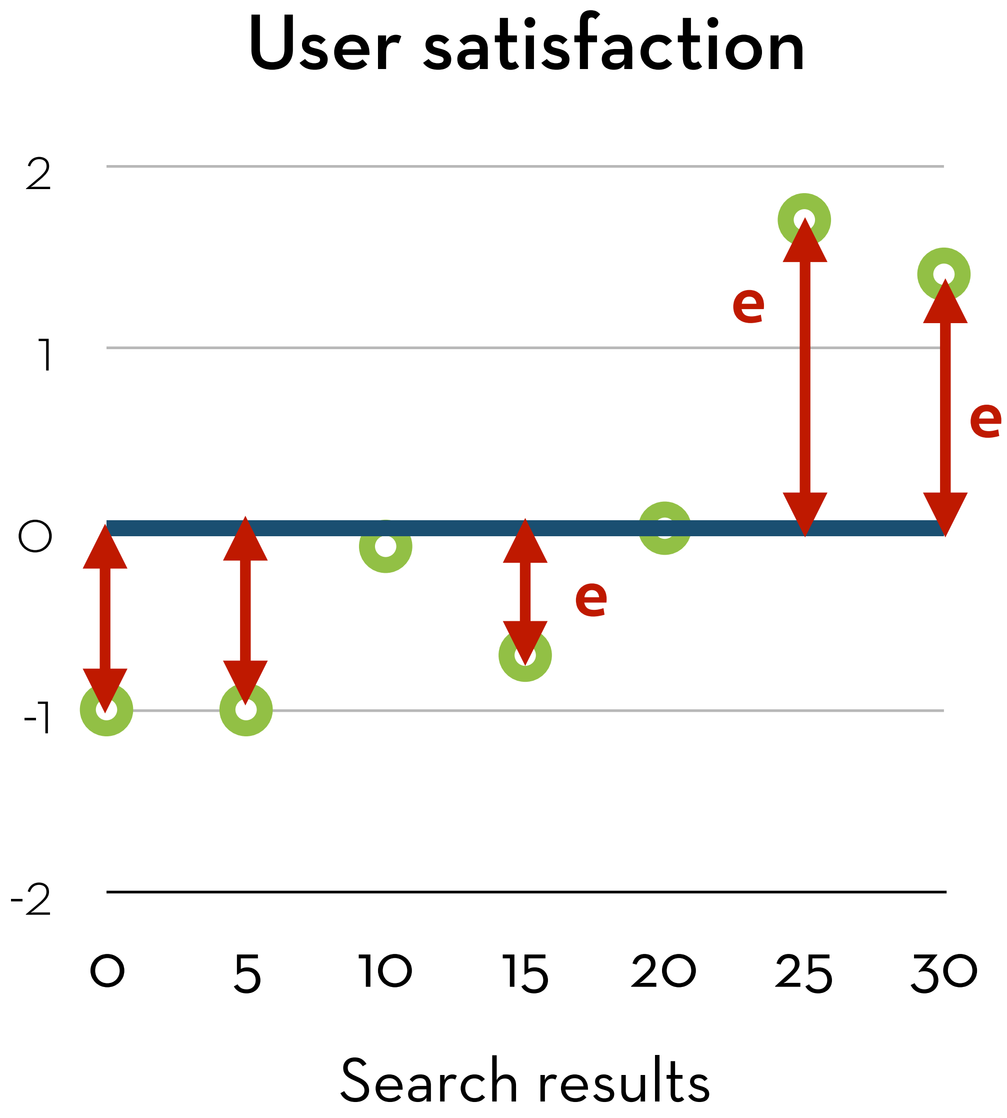
Modeling data

A model is a way to explain or summarize the data

The mean is a model

The quality of the model depends on how well it fits the data

We can measure the deviance between the model and the data



Modeling data

$\text{error}_i = x_i - \text{mean}$

$SS = \sum \text{error}_i^2$

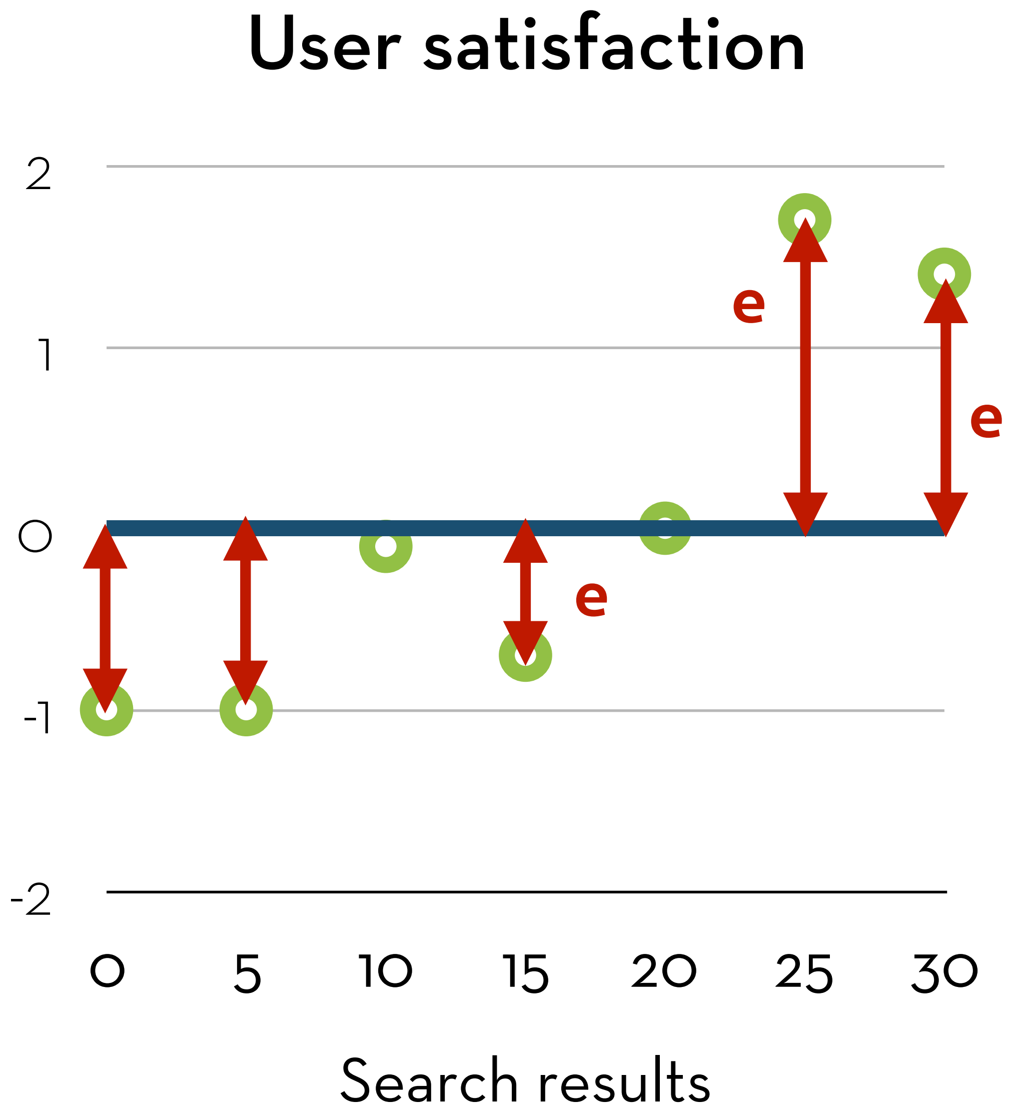
SS = sum of squared errors

$s^2 = SS/(N-1)$

s^2 = variance

s = standard deviation

N-1 = degrees of freedom



Why N-1?

Let's say you have 4 data points:

1, 3, 4, 8

Mean: 4

If you know the mean, how many data points are "free"?

Answer: Only three!

Once you know the first three, you will know the fourth one as well, because the mean needs to be 4!

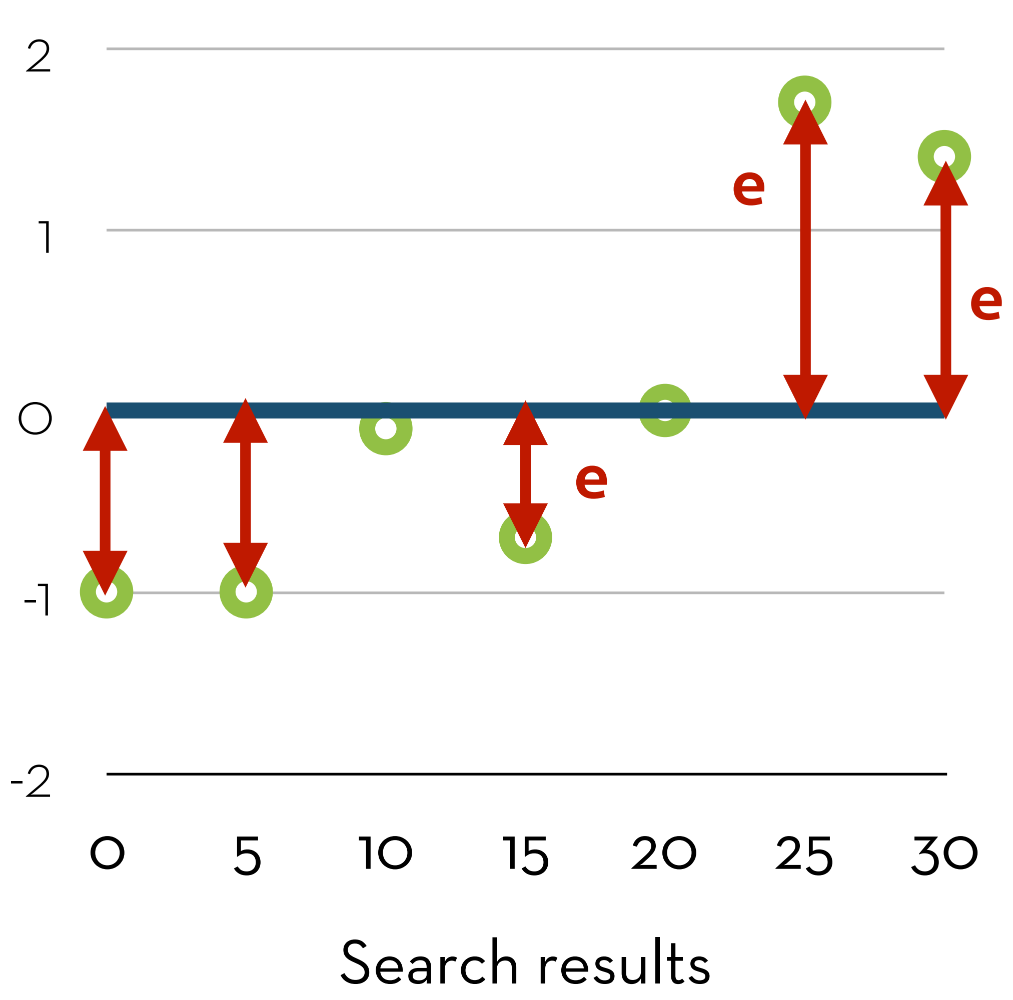
(1+3+4+x)/4 = 4 —> x has to be 8!

Variance

Variance is the variation of the data around a model (e.g. the mean)

$s^2 = \sum(x_i - \text{mean}_x)^2/(N\text{-}1)$

It is the sum of the **error in x times the error in x**, divided by the degrees of freedom



User satisfaction

Covariance

Covariance measures the relationship between the variations of two variables, x and y

$$\mathrm{cov}(x,y) = \sum(x_i - \mathrm{mean}_x)(y_i - \mathrm{mean}_y)/(N-1)$$

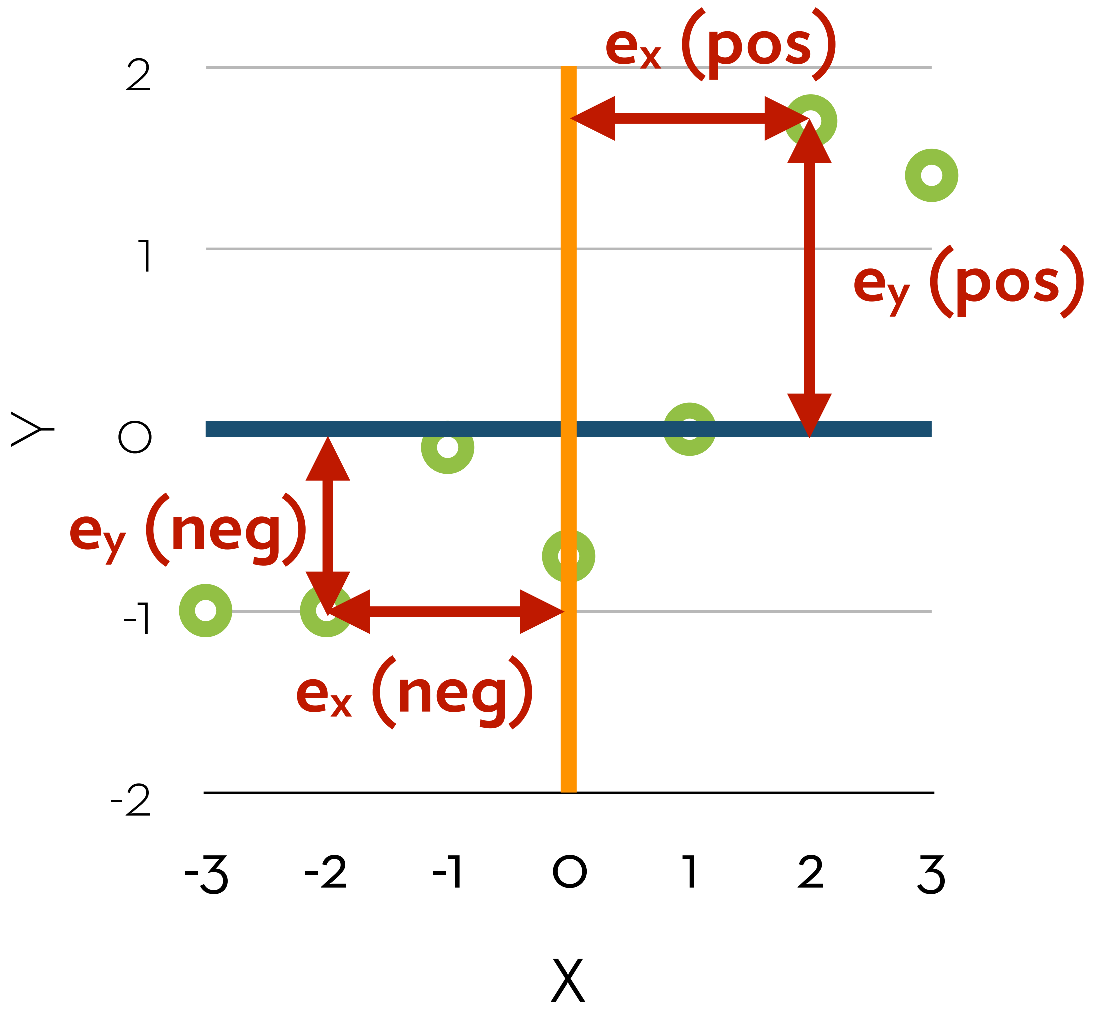
It is the sum of the **error in x times the error in y**, divided by the degrees of freedom

Covariance

Covariance measures the relationship between the variations of two variables, x and y

cov(x,y) = ∑(x$_i$ – mean$_x$) * (y$_i$ – mean$_y$)/(N-1)

It is the sum of the **error in x times the error in y**, divided by the degrees of freedom



Correlation

Standardization:

We can standardize any deviation by dividing it by the **standard deviation** of the measure (√variance)

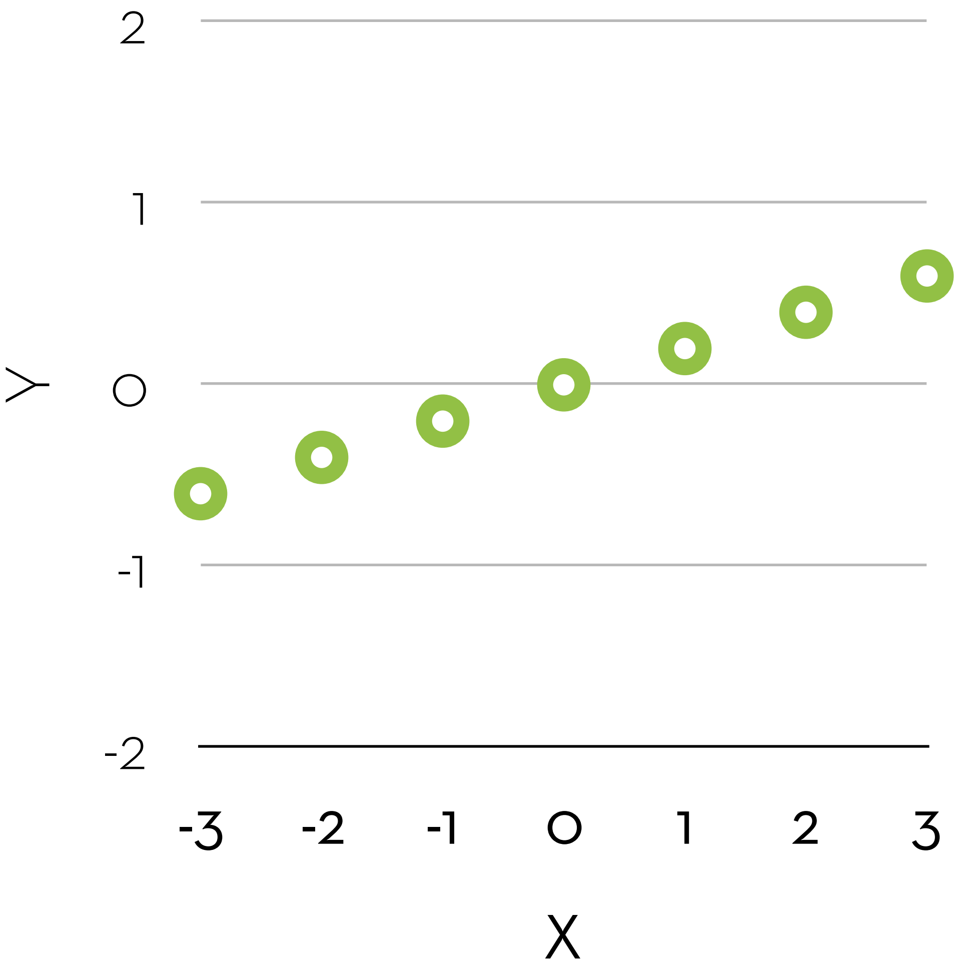
If we want to standardize the covariance, we divide by **both** the standard deviation of x and the standard deviation of y.

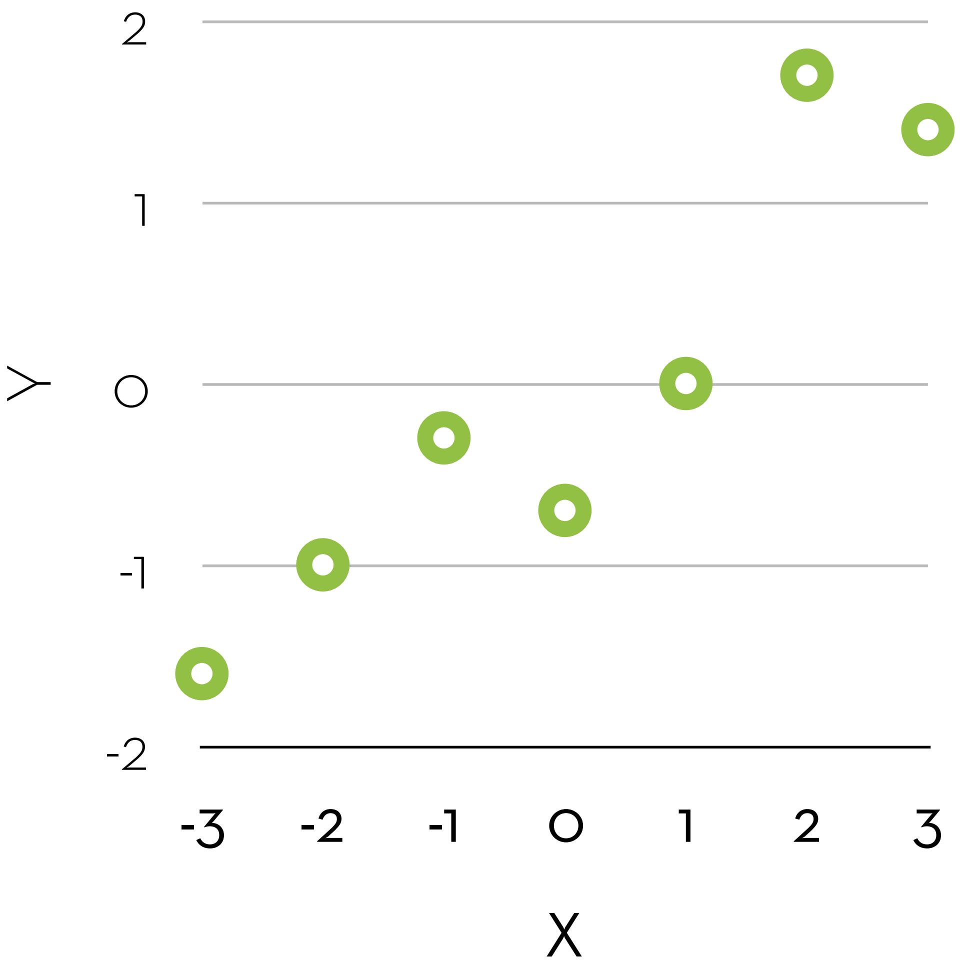
The resulting metric is the **correlation coefficient**:

$$r = \mathrm{cov}(x,y)/s_x s_y = \sum(x_i - \mathrm{mean}_x)(y_i - \mathrm{mean}_y)/(N-1)s_x s_y$$

Correlation

Which of these two graphs shows the strongest correlation?





Correlation types

Pearson: continuous x continuous

Biserial and Point-biserial: dichotomous x continuous

- Point-biserial: dichotomy is strict (e.g. dummy variable)
- Biserial: dichotomy represents an underlying continuous trait (e.g. yes/no decision)

Phi and tetrachoric: dichotomous x dichotomous

- The latter assumes underlying an underlying continuous trait

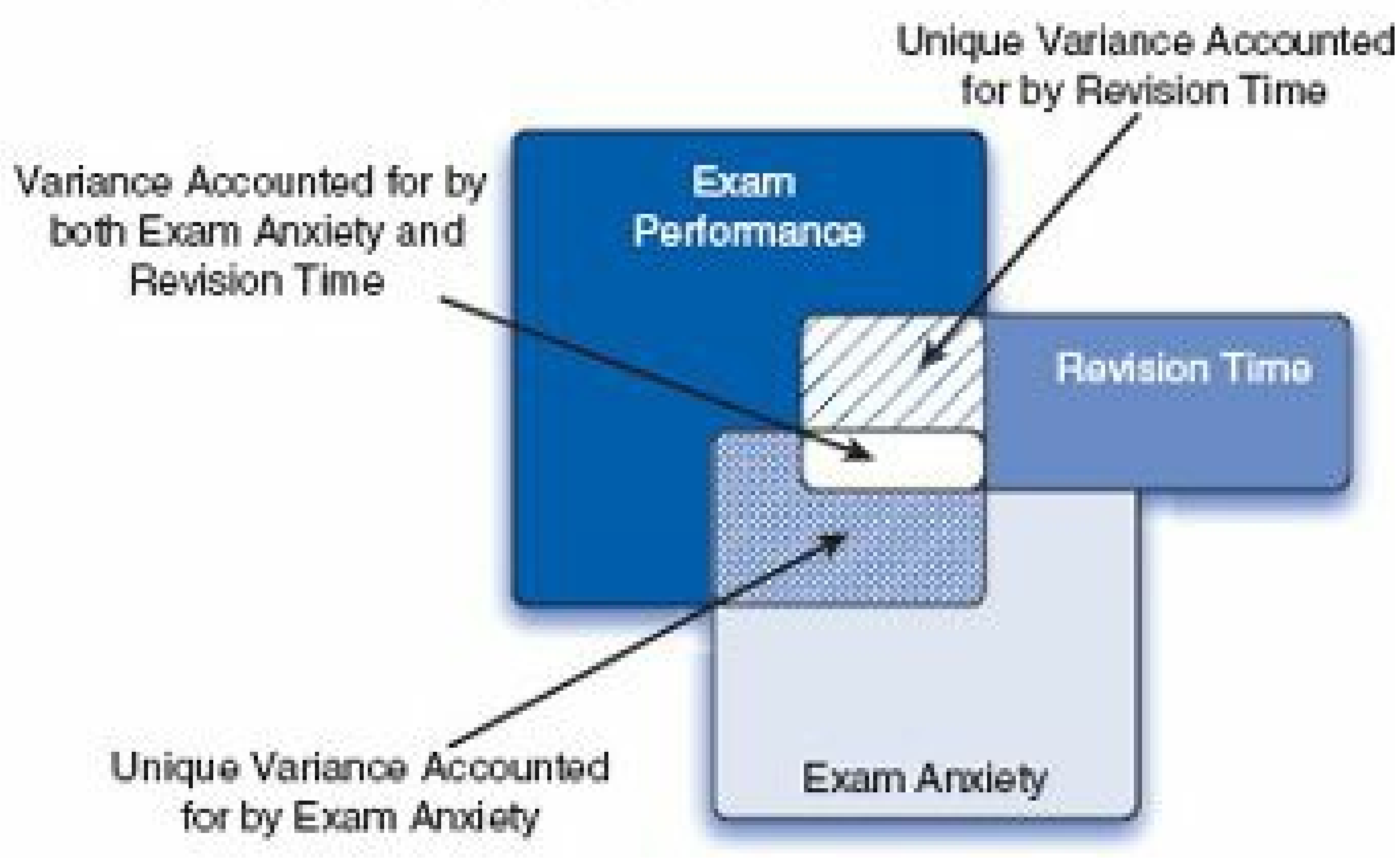
Correlation types

Polyserial: categorial x continuous

Polychoric: categorical x categorical

Spearman Rho: ordered x ordered

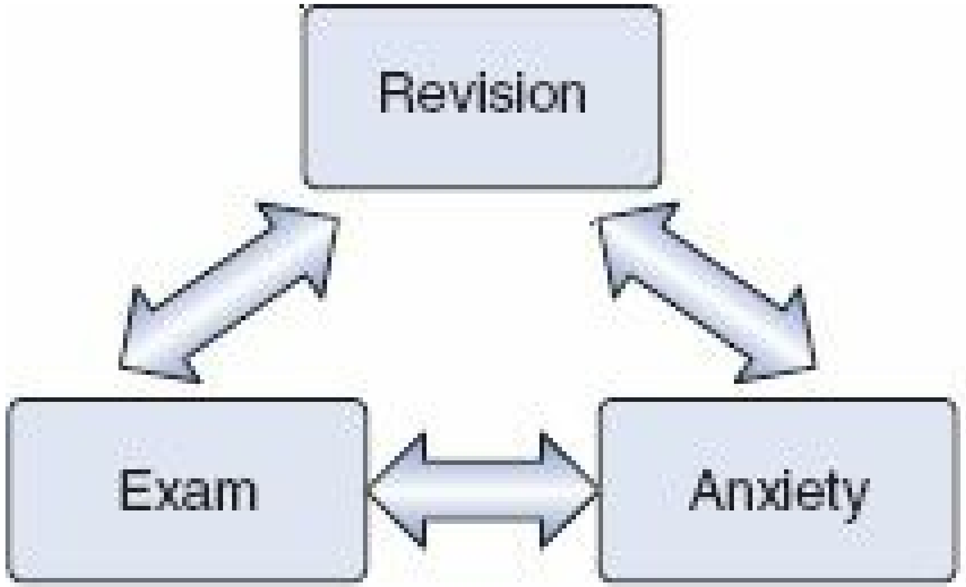
Partial correlation


Unique Variance Accounted for by Revision Time
Variance Accounted for by both Exam Anxiety and Revision Time
Exam Performance
Revision Time
Unique Variance Accounted for by Exam Anxiety
Exam Anxiety

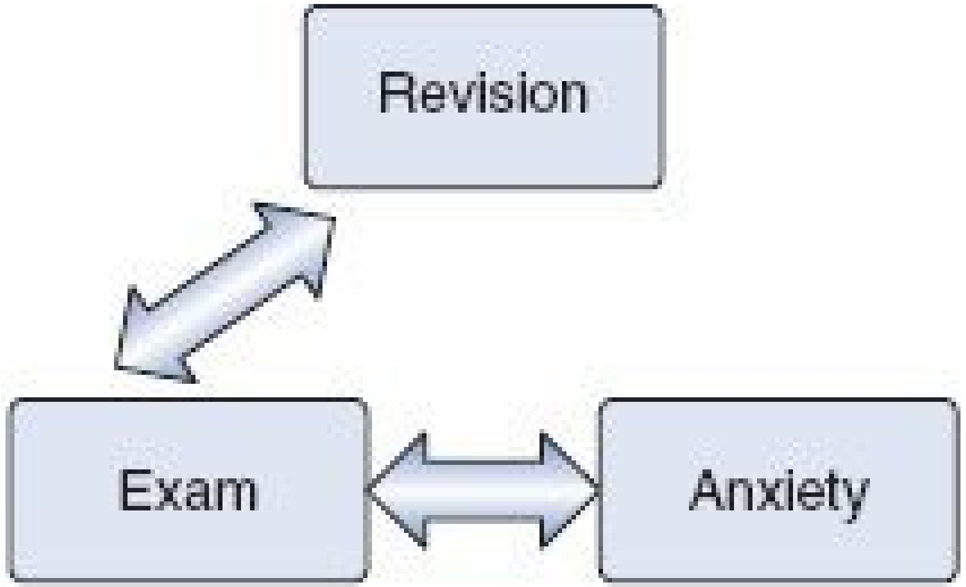
Partial correlation

The correlation between Exam score and Anxiety, without the part that could be explained by (i.e. controlling for) Revision time

The correlation between Exam score and Revision time, controlling for Anxiety



Partial Correlation



Semi-Partial Correlation

Partial correlation

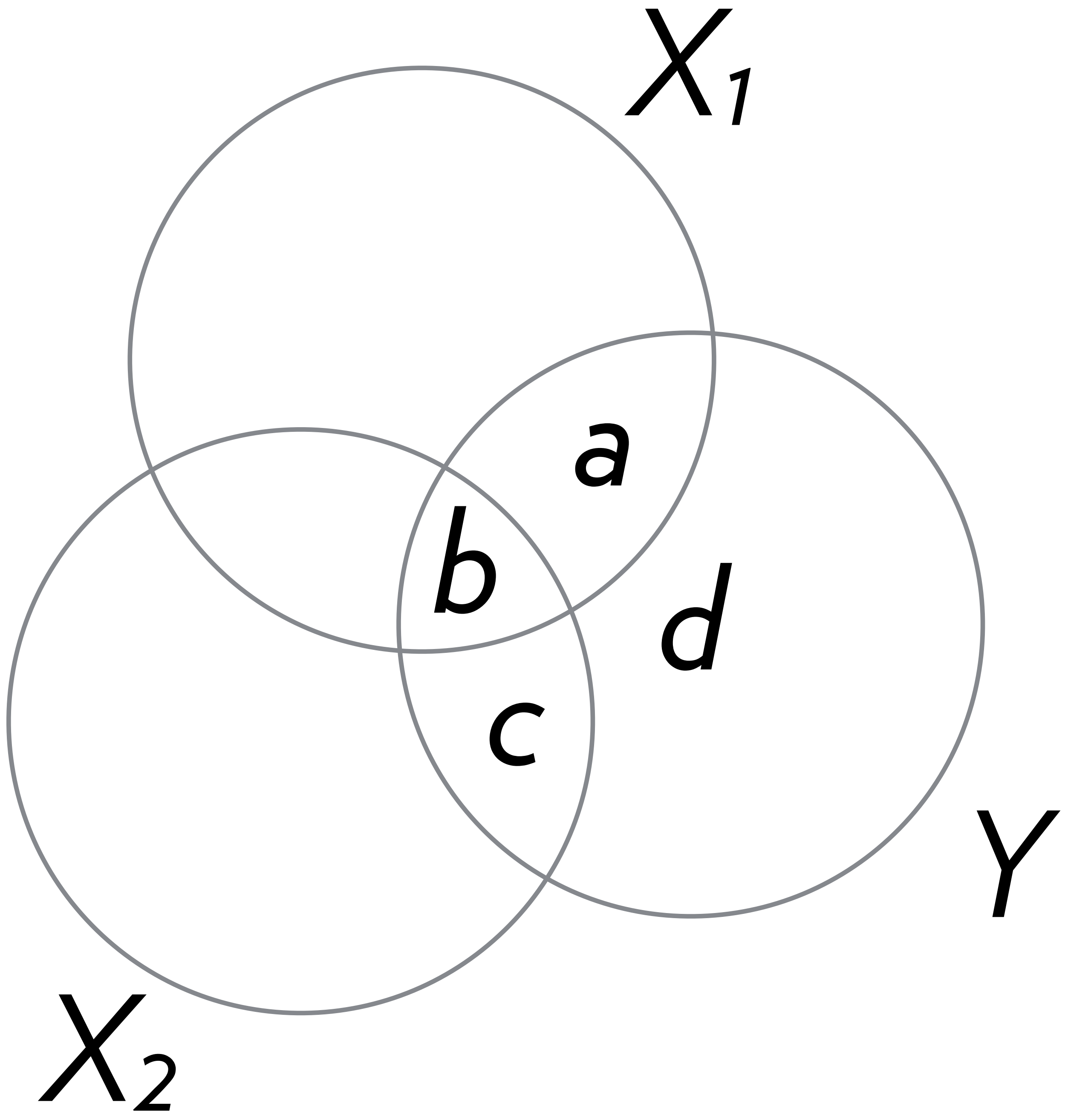
Partial correlation: a/(a+d)

Proportion of what is left after taking out X_2

Part correlation: a/(a+b+c+d)

Proportion of **total** after taking out X_2

Standardized regression coefficients are part correlations



$x \rightarrow y$

Linear regression

an explanation

$x \to y$ Linear regression

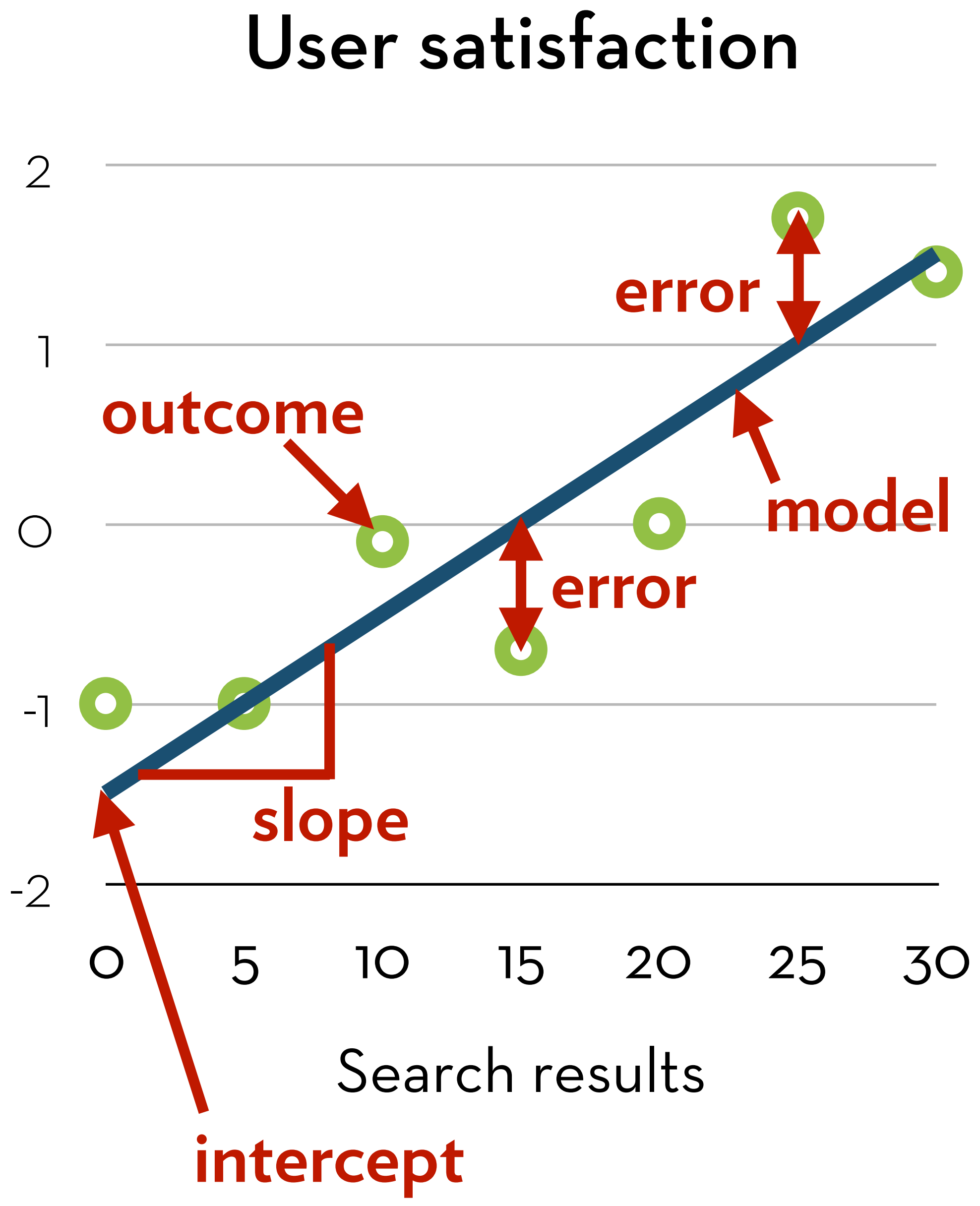
Any type of model:

outcome$_i$ = model + error$_i$

Linear regression:

The model is a line with an intercept (a) and a slope (b)

$Y_i = a + bX_i + e_i$



Finding the best line

$Y_i = a + bX_i + e_i$

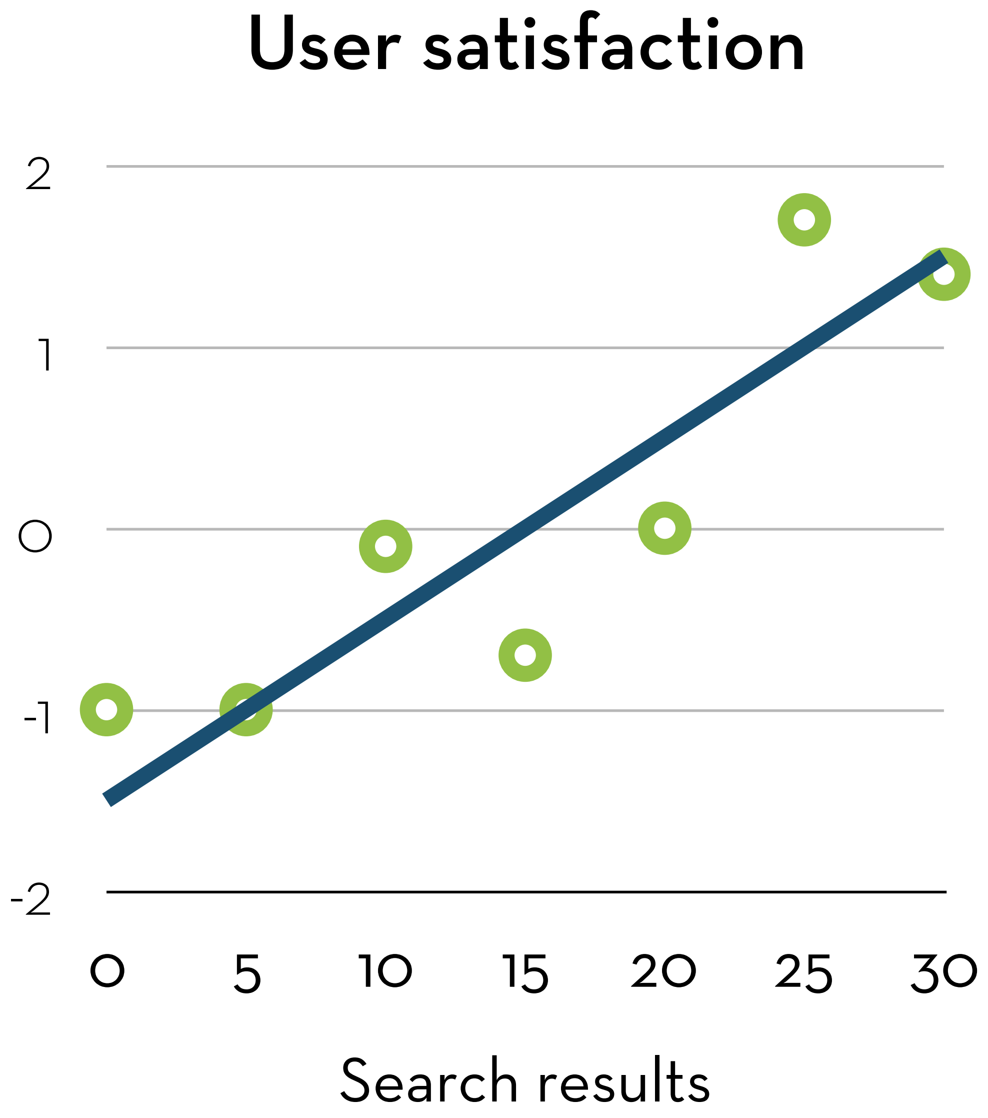
a and b are chosen so that the deviations (residuals) are minimized

We know this! General:

deviation =

$\sum(\text{observation}_i - \text{model})^2$

Goal: minimize sum of squared errors (SSr)



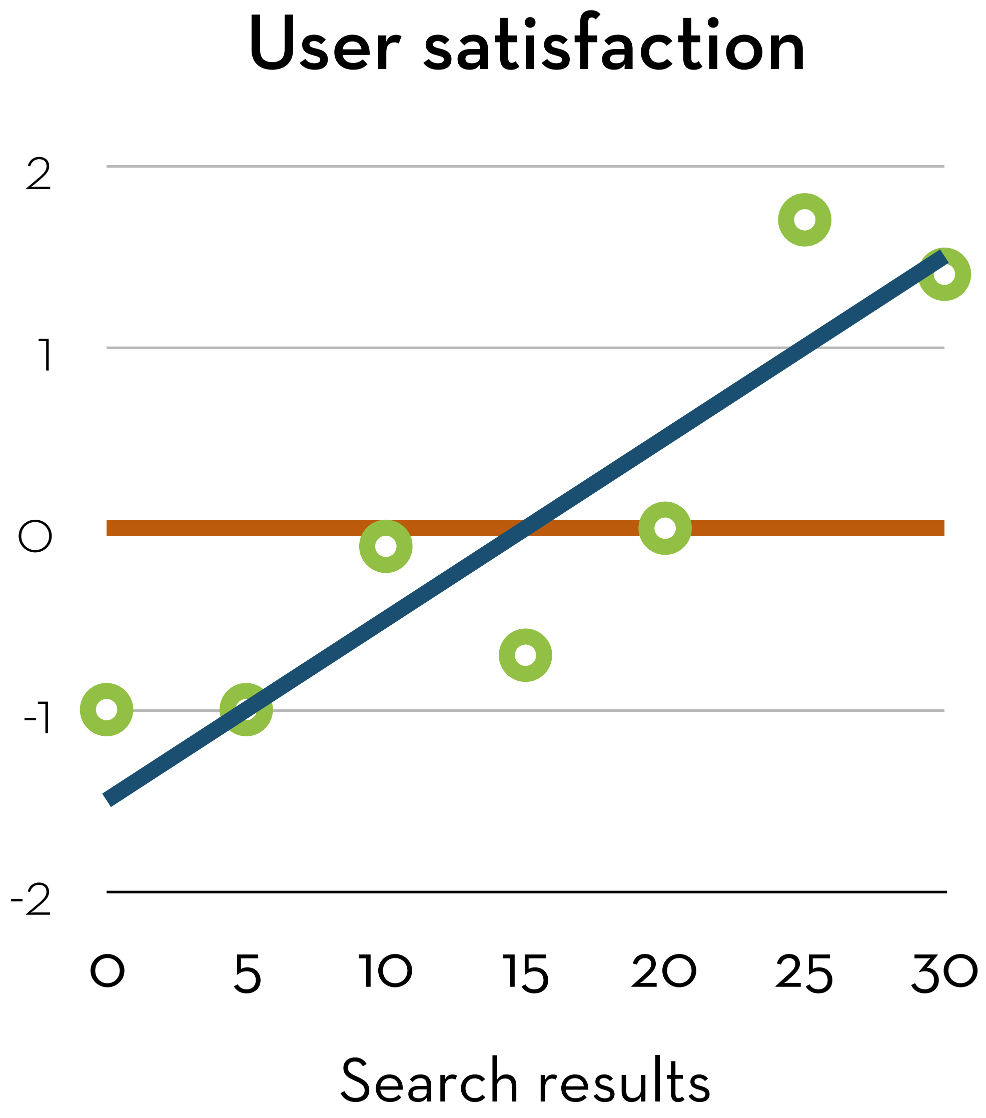
Goodness of fit

How good is the **model**?

We can use deviation for this as well!

Compare against the deviation of the simplest model

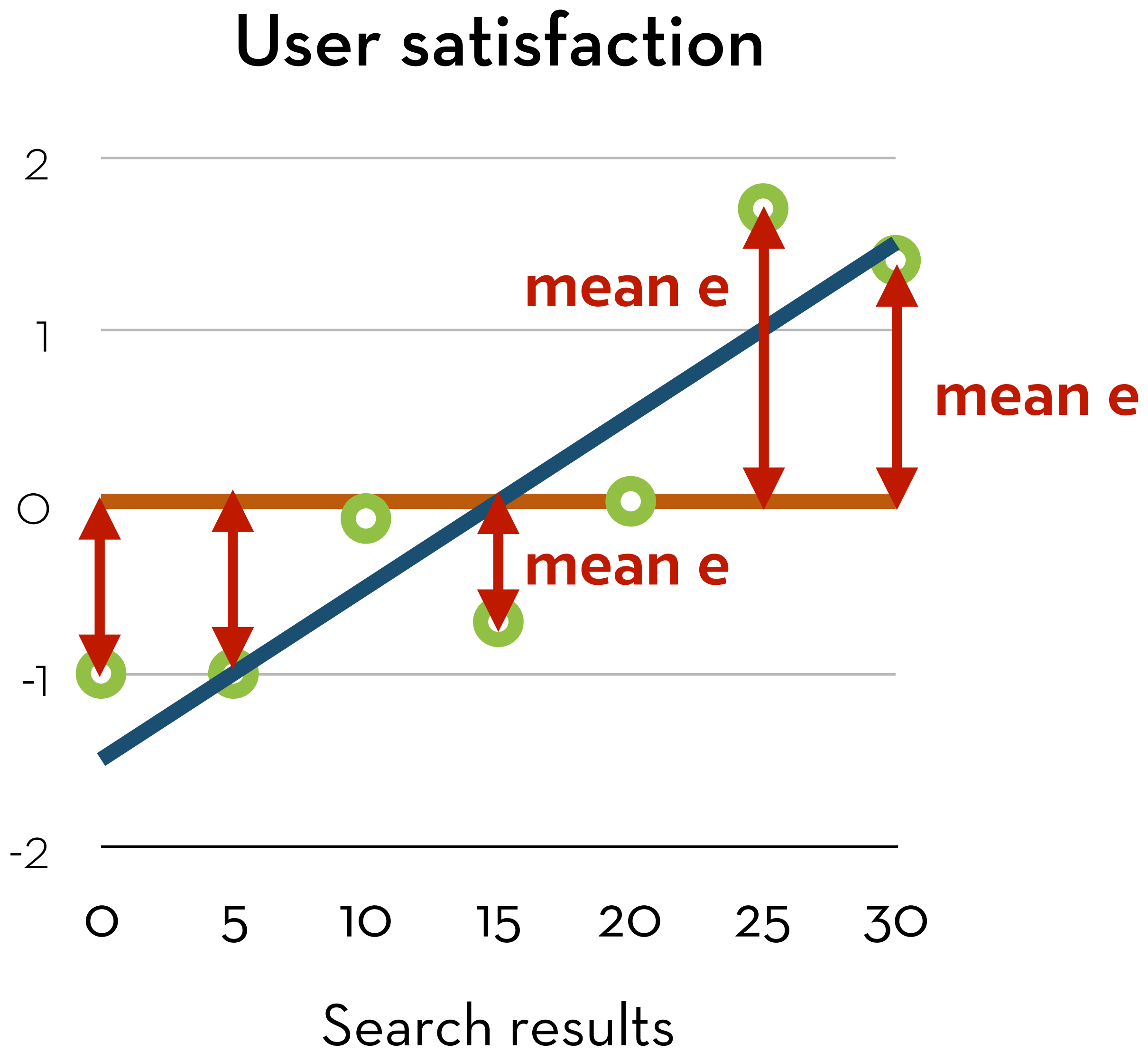
In this case: the mean



Goodness of fit

Total sum of squares (SSt)

Squared e from the mean



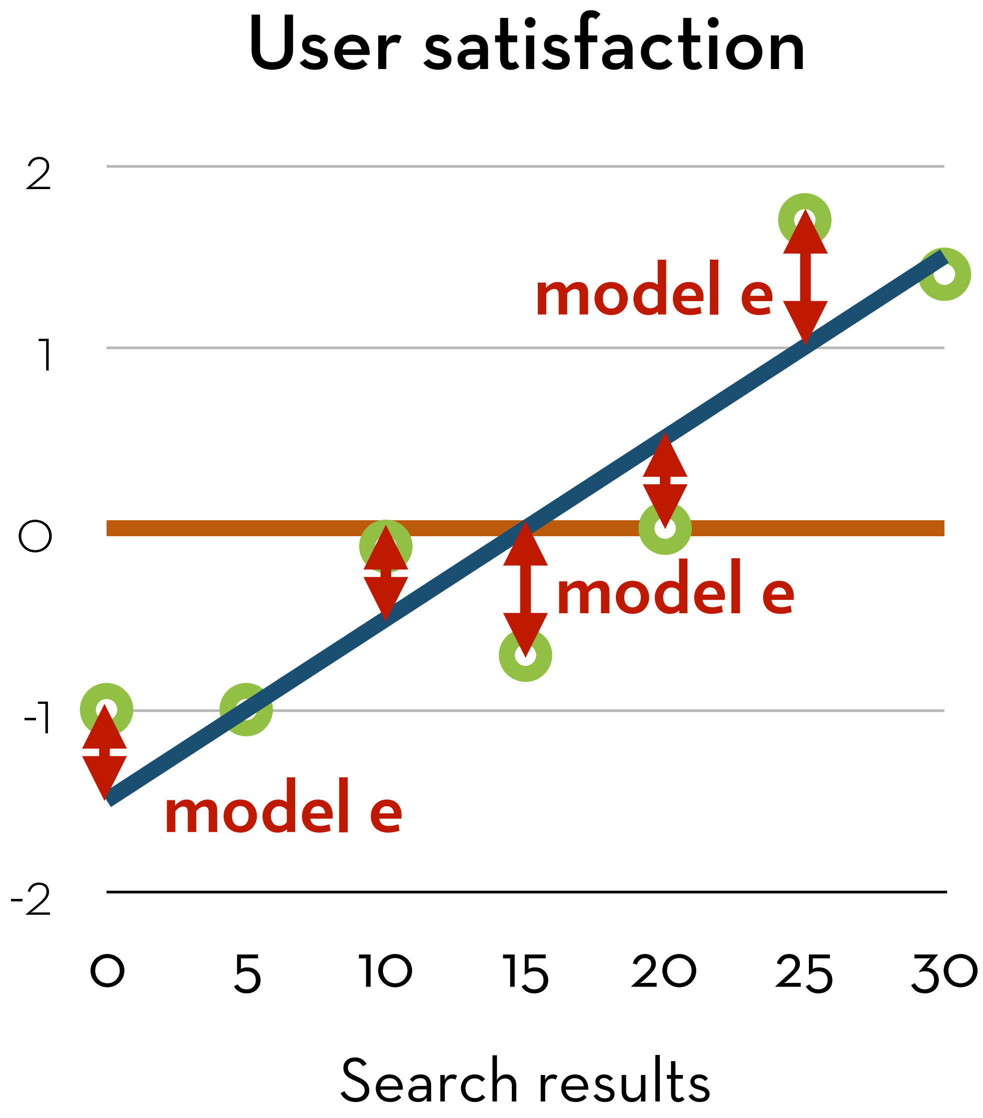
Goodness of fit

Total sum of squares (SSt)

Squared e from the mean

Residual sum of sq. (SSr)

Squared e from the model



$x \rightarrow y$ Goodness of fit

Total sum of squares (SSt)

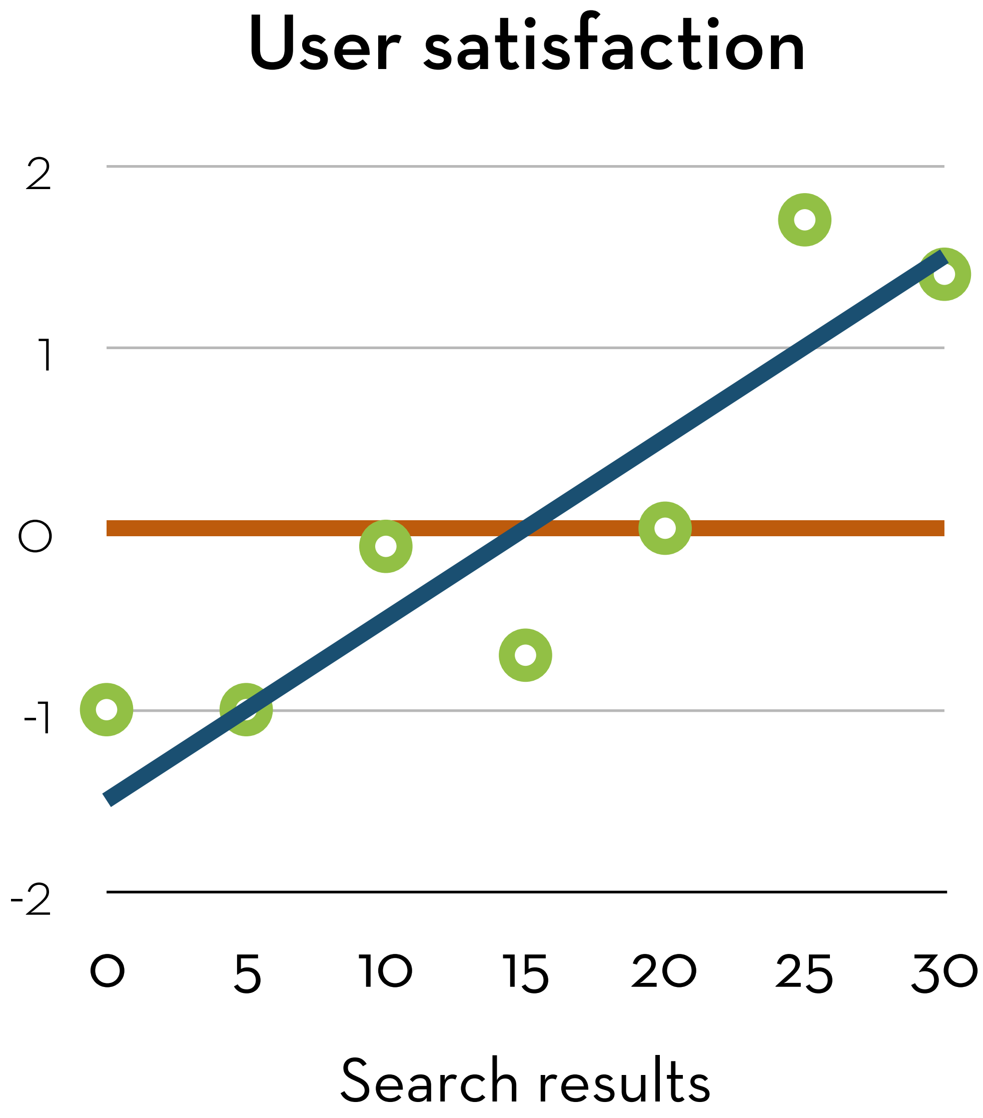
Squared e from the mean

Residual sum of sq. (SSr)

Squared e from the model

Model sum of squares (SSm)

SSt – SSr



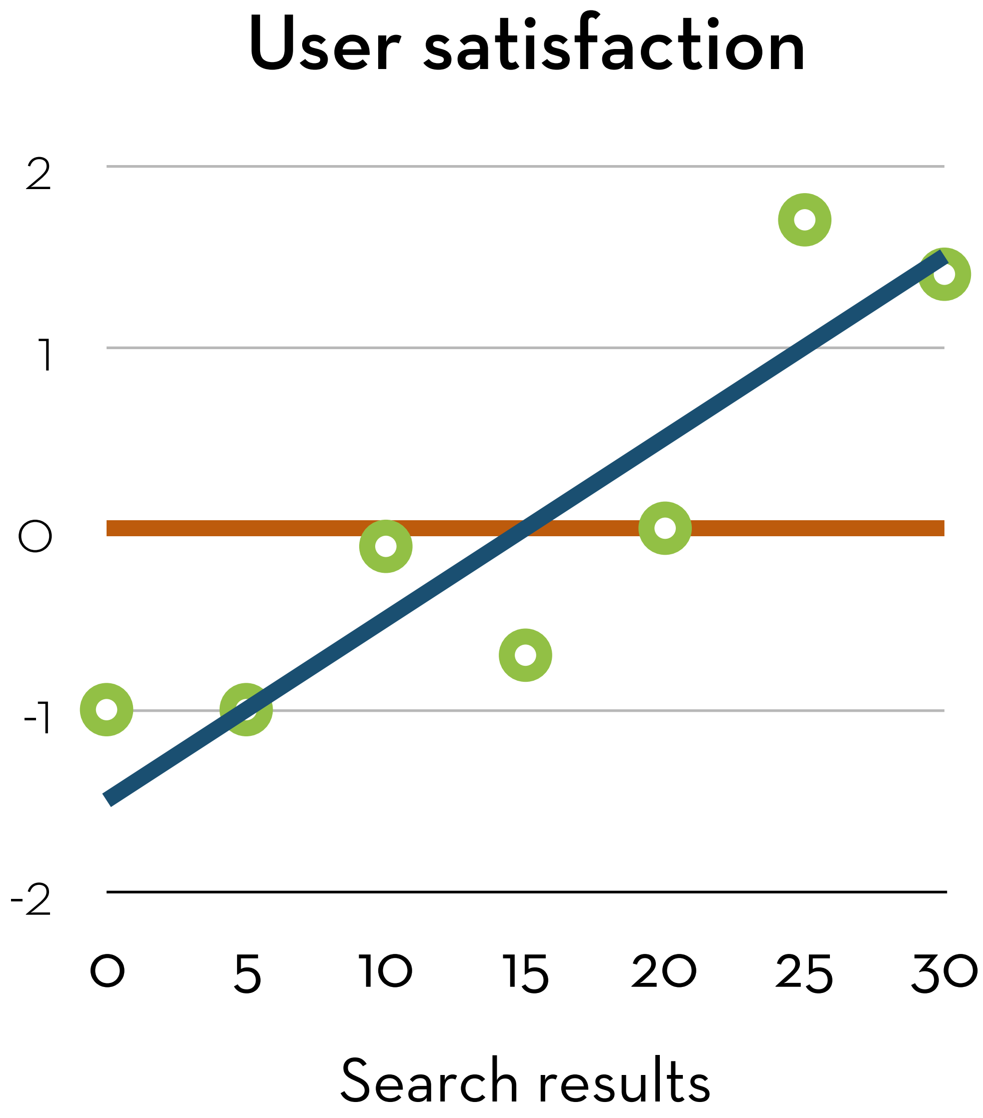
$x \rightarrow y$ Goodness of fit

R-square model fit

$R^2 = SSm / SSt$

Amount of variation in Y explained by the model

Note: In simple regression (one X), $R^2 = r_{xy}^2$



Multiple Regression

outcome$_i$ = model + error$_i$

Multiple regression:

The model is a line with an intercept (a) and **several** slopes ($b_1 \ldots b_n$)

$Y_i = a + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} + e_i$

This means you can predict satisfaction using usability **and** gender, in each case controlling for the other variable

Note: bs are partial correlations (not the same as r!)

Multiple Regression

E.g.: $\text{satisfaction}_i = 1.00 + 2.00{*}\text{usability}_i + 1.50{*}\text{gender}_i + e_i$

For every 1 point increase in usability, satisfaction is expected to increase by 2 points, controlling for gender

Controlling for usability, the satisfaction for males (1) is expected to be 1.5 points higher than for females (0)

Multiple Regression

Why “controlling for”?

- There may be a difference in usability between males and females!

E.g. if males’ usability is 0.5 points higher on average, then the males in our sample are expected to have a 1.5 + 2*0.5 = 2.5 point higher satisfaction

- 1.5 points because of their gender, 1 point because of their higher average usability

Interaction effects

$\text{satisfaction}_i = 1.00 + 2.00{*}\text{usability}_i + 1.50{*}\text{gender}_i + 1.20{*}\text{usability}_i{*}\text{gender}_i + e_i$

For females (0), for every 1 point increase in usability, satisfaction is expected to increase by 2 points

At usability = 0, the satisfaction for males (1) is expected to be 1.5 points higher than for females (0)

For males (1), the effect of usability is 1.2 points higher than for females (i.e. 3.2 points per one point of usability)

For a 1-point higher usability, the effect of gender is 1.2 points higher (i.e. 2.7 points difference at usability = 1)

Comparisons

Say, $\text{satisfaction}_i = 1.40 + 2.00{*}\text{usability}_i + 1.30{*}\text{control}_i + e_i$

Can you compare the effect of usability with the effect of control?

Depends on the scale!

Solution: standardize!

b*SDy/SDx

relative predictive power

Goodness of fit

R^2 = SSm / SSt

- Same as before, but R^2 is now called the "multiple R^2"
- Combined effect of all predictors
- Total variance in Y explained by all Xes in the model

Note: R^2 is the sum of part correlations (plus overlap)

- It is also the square of the correlation between Y and the predicted value of Y

$x \rightarrow y$ Testing a predictor

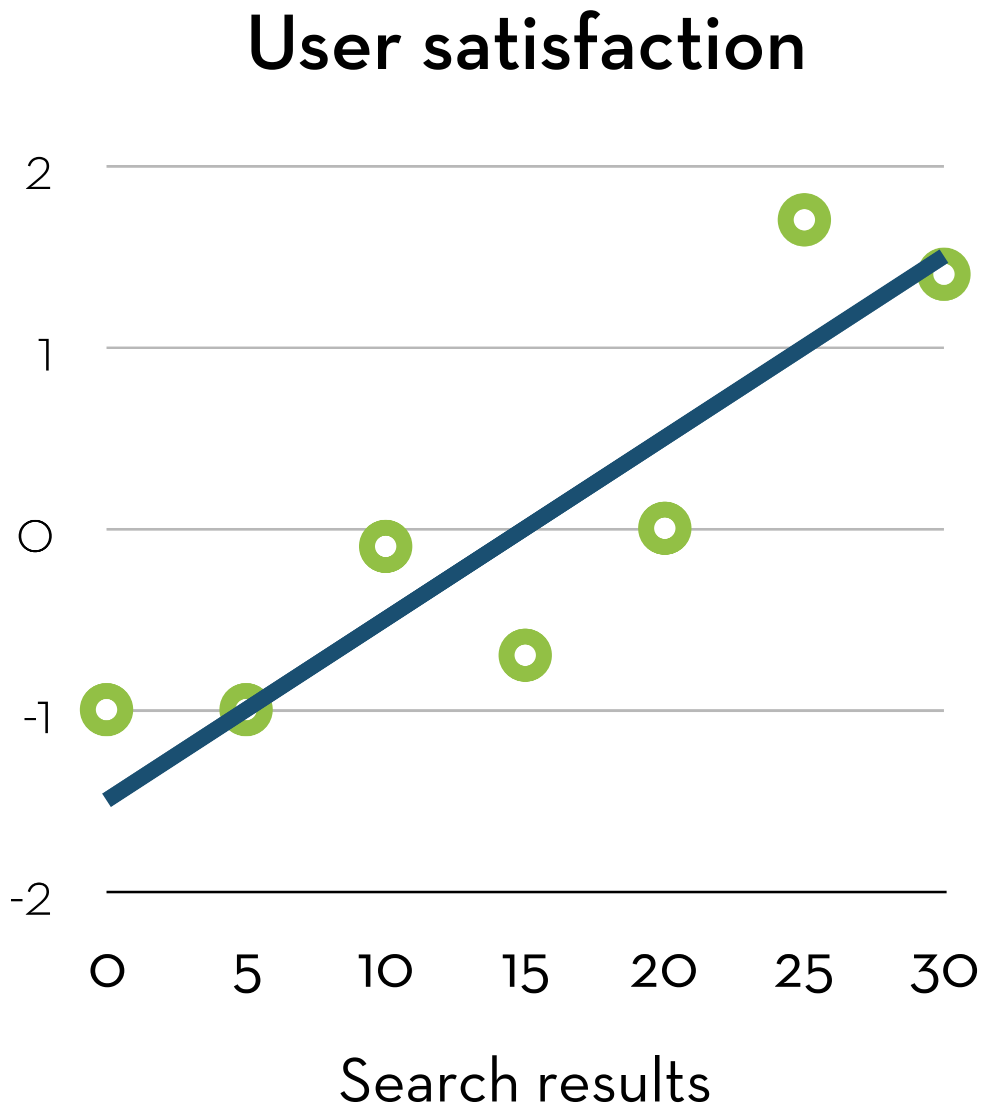
If a predictor is bad, its slope (b) will be almost zero (like the mean)

A good predictor has a slope that is significantly different from zero

Compare slope (b) against variability of slope (SEb):

$t = b/SEb$

with $df = N - p - 1$



p-value

p is the percentage of times you'd expect a result of this magnitude or larger if there was actually no effect

Not the probability the effect happened by chance

Not the probability that you incorrectly rejected H0 (or that H0 is true)

Not 1 - the probability that H1 is true

Not 1 - the probability that you find the same effect again

Note: any statistical test assumes that the null hypothesis is true (and then checks how unlikely the observed sample is).

Some assumptions

No measurement error (this problem can be fixed with SEM)

- Errors in Xs result in biased predictors (can be either direction)
- Error in Y results in lower R-square, larger SD, so larger error and lower beta (not B)

Eg.: we have a variable S measuring trait Y with 36% error

- This is random noise that does not measure Y

Result: no regression with S as dependent can have an R-squared > 0.64!

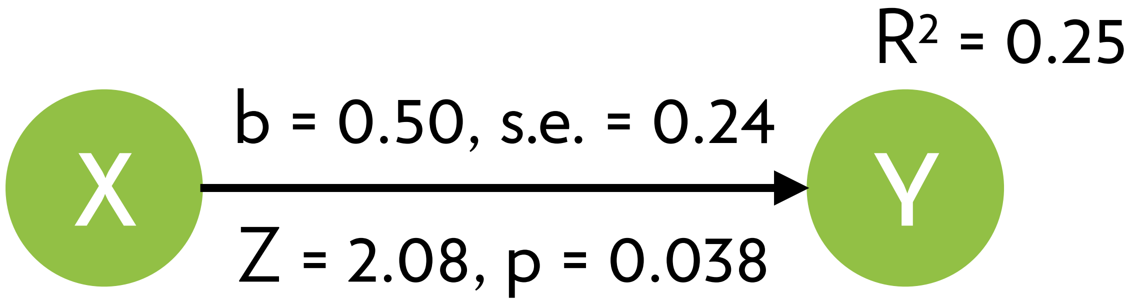
Measurement error

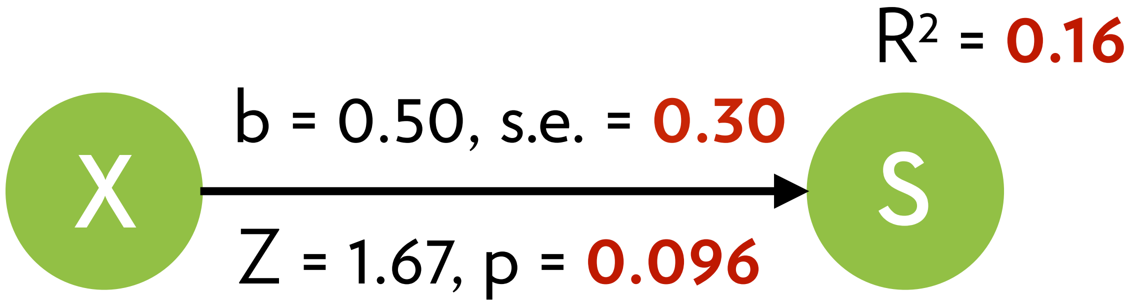
Any s.e. will be **attenuated** by the error of S!

Take for instance this X, which potentially explains 25% of the variance of Y...

...it only explains 16% of the variance of S!

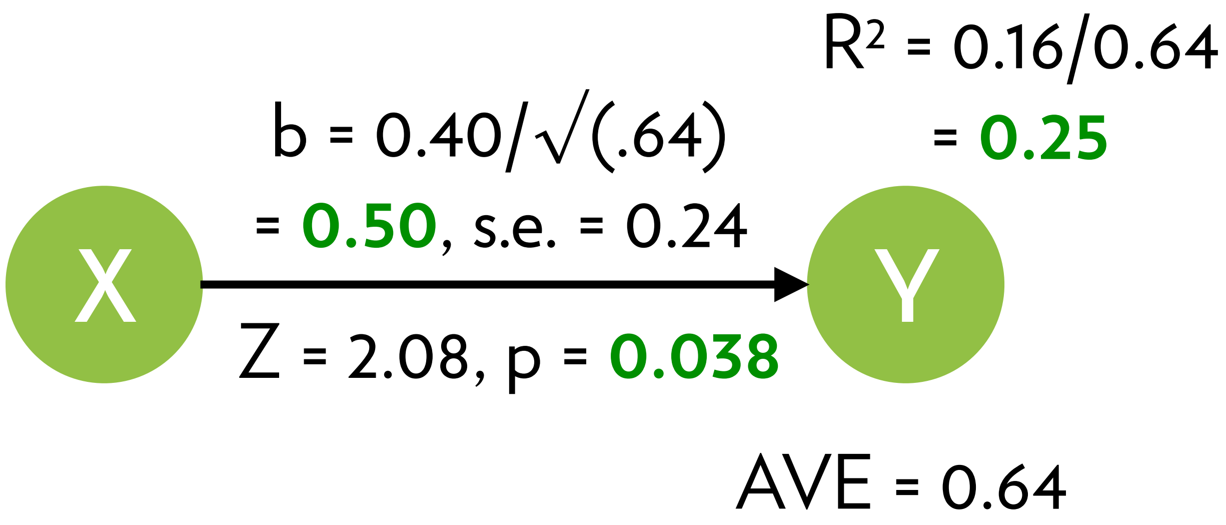
...and the effect is non-significant!





Measurement error

If we could use Y instead of S, we can get much more precise tests



$R^2 = 0.16/0.64$

= **0.25**

$b = 0.40/\sqrt{(.64)}$

= **0.50**, s.e. = 0.24

Z = 2.08, p = **0.038**

AVE = 0.64

Some assumptions

No variables correlated with both X and Y should be left out (still a problem in SEM)

This is called “suppression”

- Usually results in overestimation, e.g. shoe size and intelligence (omitted: age)
- Can also result in underestimation, e.g. ice cream sales and date (omitted: hemisphere)
- Can even result in sign switching (negative suppression): race and arrests (omitted: police profiling)

Some assumptions

Outcome should be quantitative, continuous, unbounded

Not true for:

- yes-no questions
- counts (e.g. of clicks)
- 5-point rating scales

Otherwise:

- Non-linear regression
- Bootstrapping

$x \rightarrow y$

Non-linear regression

logistic, poisson, ordered-logistic

Logistic regression

Linear regression:

$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki} + e_i$

What if Y is binary (0 or 1)?

We can try to predict the **probability** of Y=1 — P(Y)

However, this probability is a number between 0 and 1

For linear regression, we want an unbounded linear Y!

Can we find some transformation that allows us to do this?

Yes: $P(Y) = 1 / (1+e^{-U})$

$x \rightarrow y$ Logistic regression

$P(Y) = 1 / (1+e^{-U})$

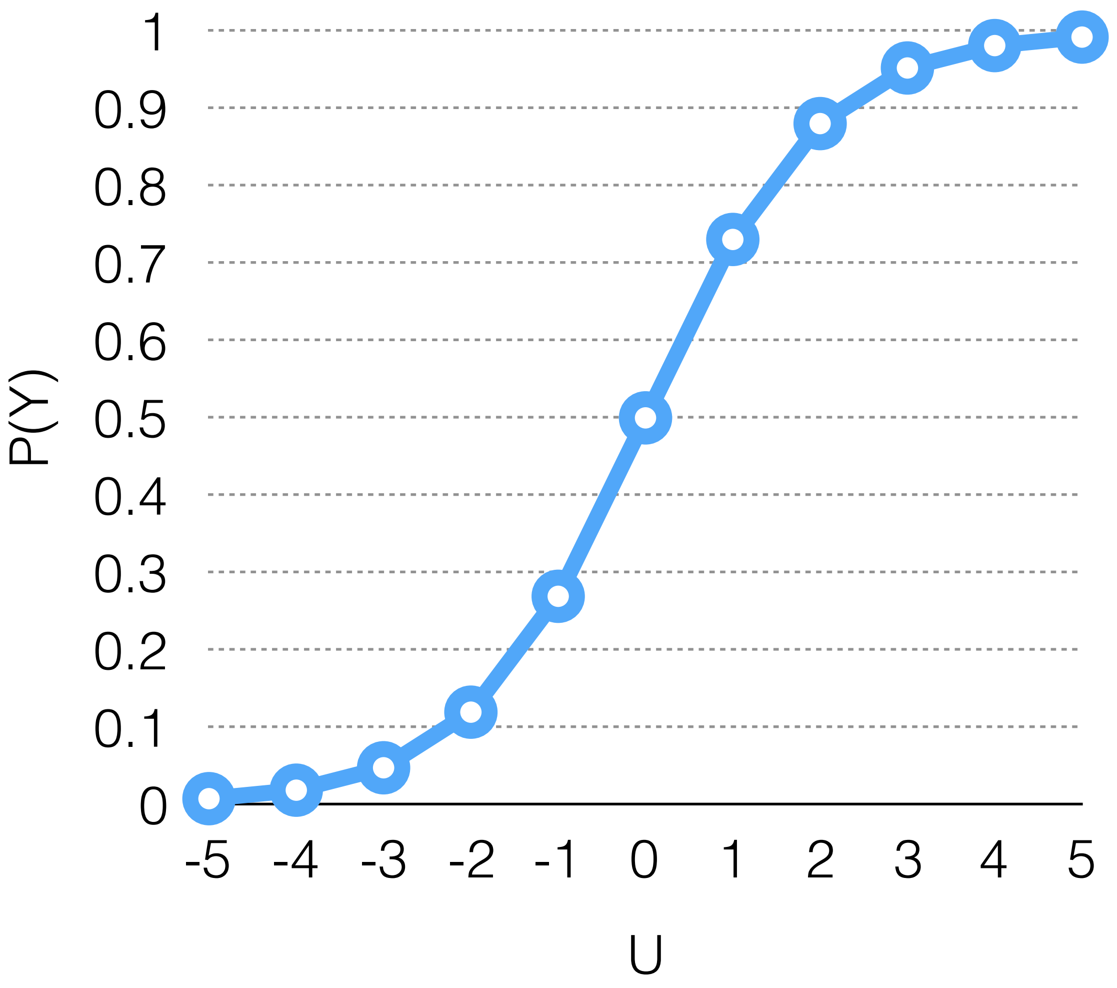
Conversely:

$U = \ln(P(Y)/(1-P(Y)))$

Interpretation:

$P(Y)/(1-P(Y))$ is the **odds** of Y

Therefore, U is the log odds, or **logit** of Y



Logistic regression

Since U is unbounded, we can treat it as our regression outcome:

$$U_i = \ln(P(Y_i)/(1–P(Y_i))) = Y_i = a + b_1X_{1i} + b_2X_{2i} + ... + b_kX_{ki} + e_i$$

We can always transform it back to $P(Y_i)$ if we want to:

$$P(Y_i) = 1 / (1+e^{–(a + b1X1i + b2X2i + ... + bkXki + ei)})$$

Log-likelihood

How do we assess the fit of a logistic regression?

We calculate the **log-likelihood**, which is a type of residual

Log-likelihood = $\sum(Y_i * \ln(P(Y_i)) + (1 - Y_i) * \ln(1 - P(Y_i)))$

where Y_i is the observed value, and $P(Y_i)$ is the predicted value

Log-likelihood

Log-likelihood = $\sum(Y_i{*}\ln(P(Y_i)) + (1-Y_i){*}\ln(1-P(Y_i)))$

If $Y_i = 1$, then this simplifies to $\ln(P(Y_i))$

> which is zero when the prediction is correct ($P(Y_i)=1$) but gets a large (negative) value if the prediction is incorrect ($P(Y_i)$ is closer to 0)

If $Y_i = 0$, then this simplifies to $\ln(1-P(Y_i))$

> which is zero when the prediction is correct ($P(Y_i)=0$) but gets a large (negative) value if the prediction is incorrect ($P(Y_i)$ is closer to 1)

Deviance (-2LL)

A more useful measure is deviance (a.k.a. –2LL)

–2 * log-likelihood

Difference can be used to compare nested models

Likelihood ratio: $\chi^2 = -2LL_{baseline} - -2LL_{new}$

Chi-square distribution with $k_{new} - k_{baseline}$ df

We will use this a lot in CFA and SEM to compare different models!

Coefficients

How to interpret the b coefficients?

- b is the increase in U for each increase of X
- b is the increase in ln(P(Y)/(1–P(Y))) for each increase in X
- e^b is the ratio of P(Y)/(1–P(Y)) for each increase in X
- e^b is the **odds ratio**

Coefficients

Odds ratio examples:

If $e^b > 1$: The odds of Y are e^b times as high for each increase in X

> E.g. $e^b = 3$: The odds of Y are 3 times as high for each increase in X

If $e^b < 1$: The odds of Y are $1/e^b$ times as low for each increase in X

> E.g. $e^b = .333$: The odds of Y are 3 times as low for each increase in X

Coefficients

If e^b = 1.xx: each 1 pt increase in X leads to a xx% increase in the odds of Y

E.g. e^b = 1.30: The odds of Y are 30% higher for each increase in X

If e^b = 0.xx: each 1pt increase in X leads to a (100-xx)% decrease in the odds of Y

e^b = 0.70: The odds of Y are 30% lower for each increase in X

Problems

Some times a logistic regression does not converge

You will get weirdly large standard errors

1. You have no or little data for some combinations of Xs

This is especially problematic when Xs are nominal

2. One or a combination of Xs are a perfect predictor of Y

The odds ratios are infinite!

Solution:

Collect more data, or use a simpler model!

Poisson regression

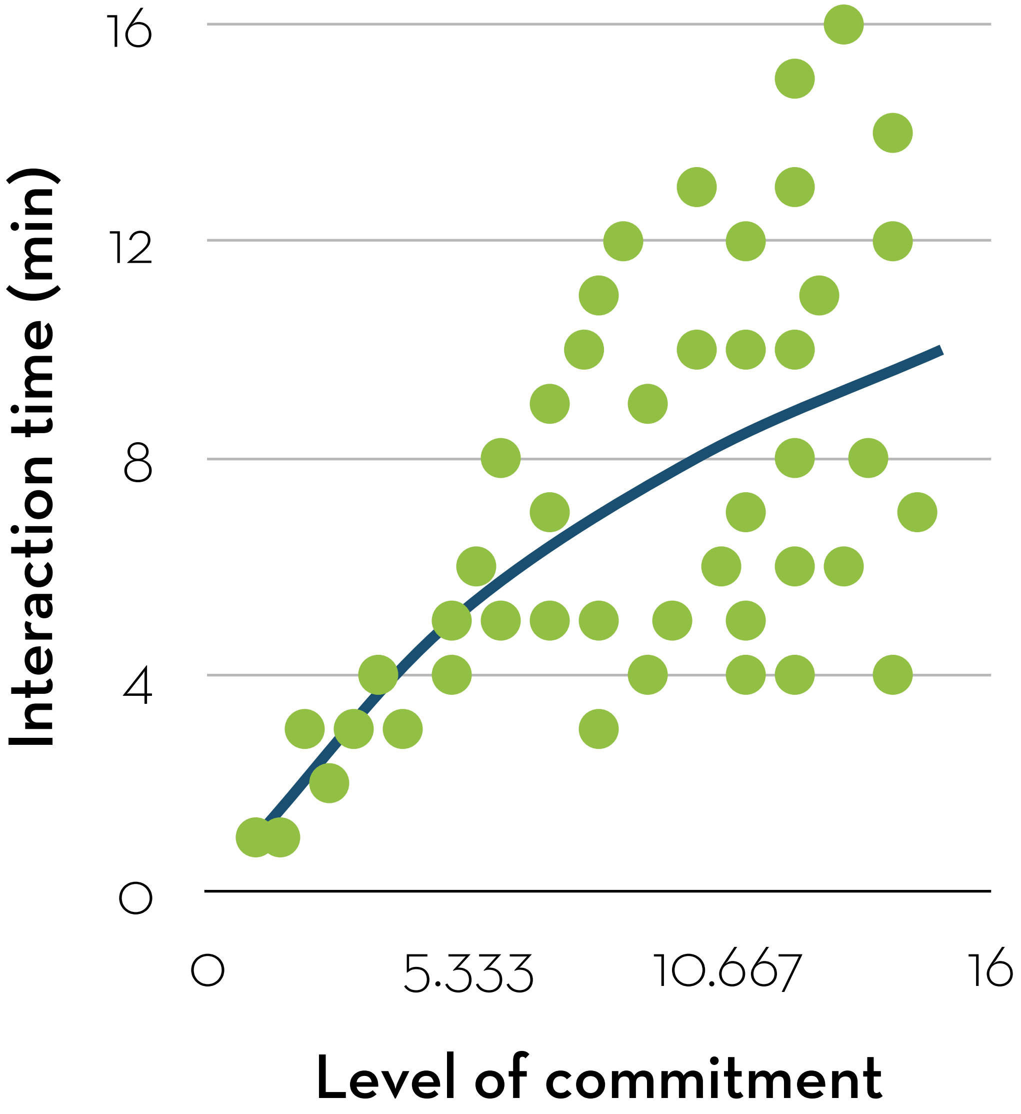
Count variables often look like this

Examples: # of purchases, # of clicks, time*, price*

Not normal, heteroscedastic!

Can we find some transformation that makes this work?

Yes: $Y = e^U$



Coefficients

How to interpret the b coefficients?

b is the increase in U for each increase of X

b is the increase in the **log rate** of Y for each increase in X

e^b is the ratio of rate Y for each increase in X

e^b is the **rate ratio**

Why the ratio?

$b = \log(\text{rate}_{x+1}) - \log(\text{rate}_x) = \log(\text{rate}_{x+1} / \text{rate}_x)$

therefore, $e^b = \text{rate}_{x+1} / \text{rate}_x$

Ordered logistic

Question: “I only act to satisfy immediate concerns, figuring the future will take care of itself.”

Answer categories:

1=extremely uncharacteristic

2=somewhat uncharacteristic

3=uncertain

4=somewhat characteristic

5=extremely characteristic

A problem...

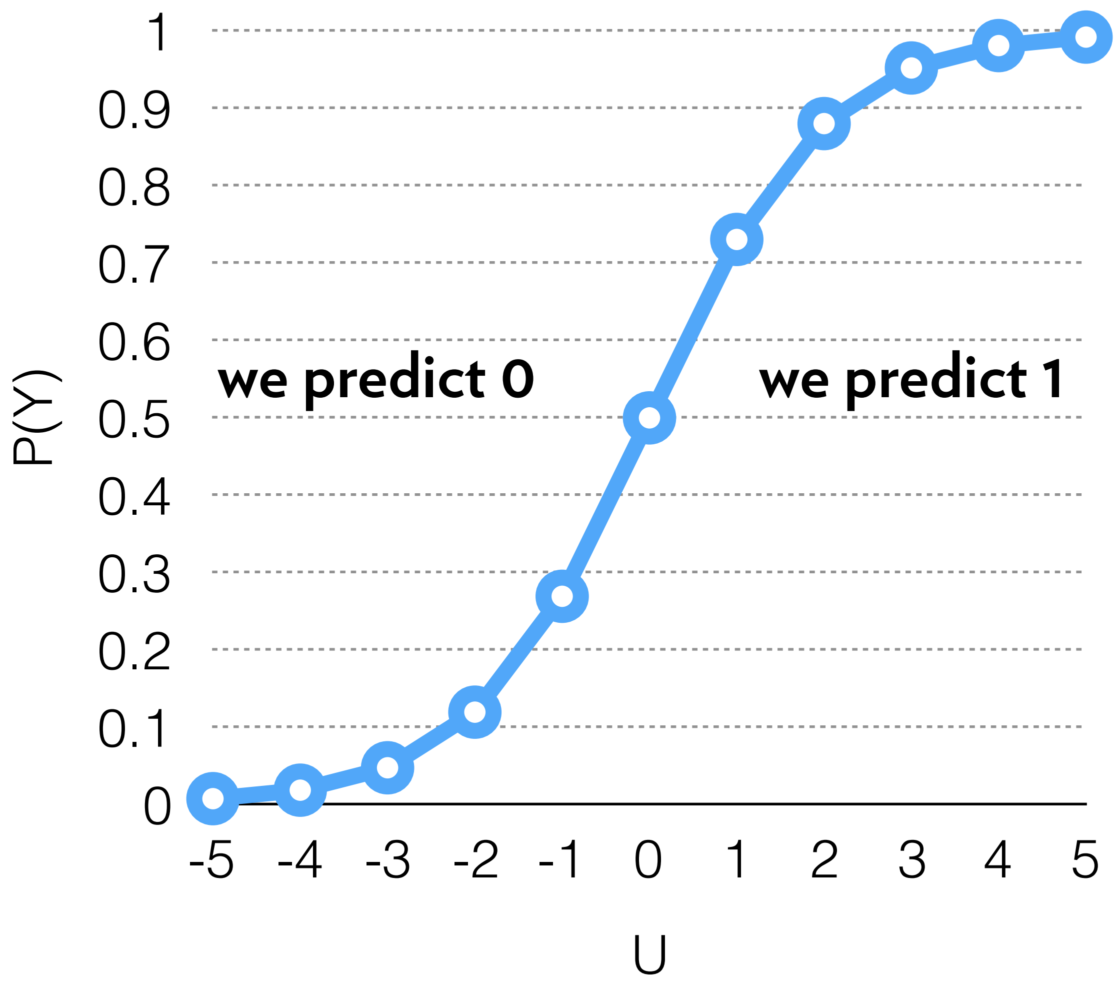
This is ordinal, not interval!

> Is the difference between “extremely uncharacteristic” and “somewhat uncharacteristic” the same as the difference between “uncertain” and “somewhat characteristic”?

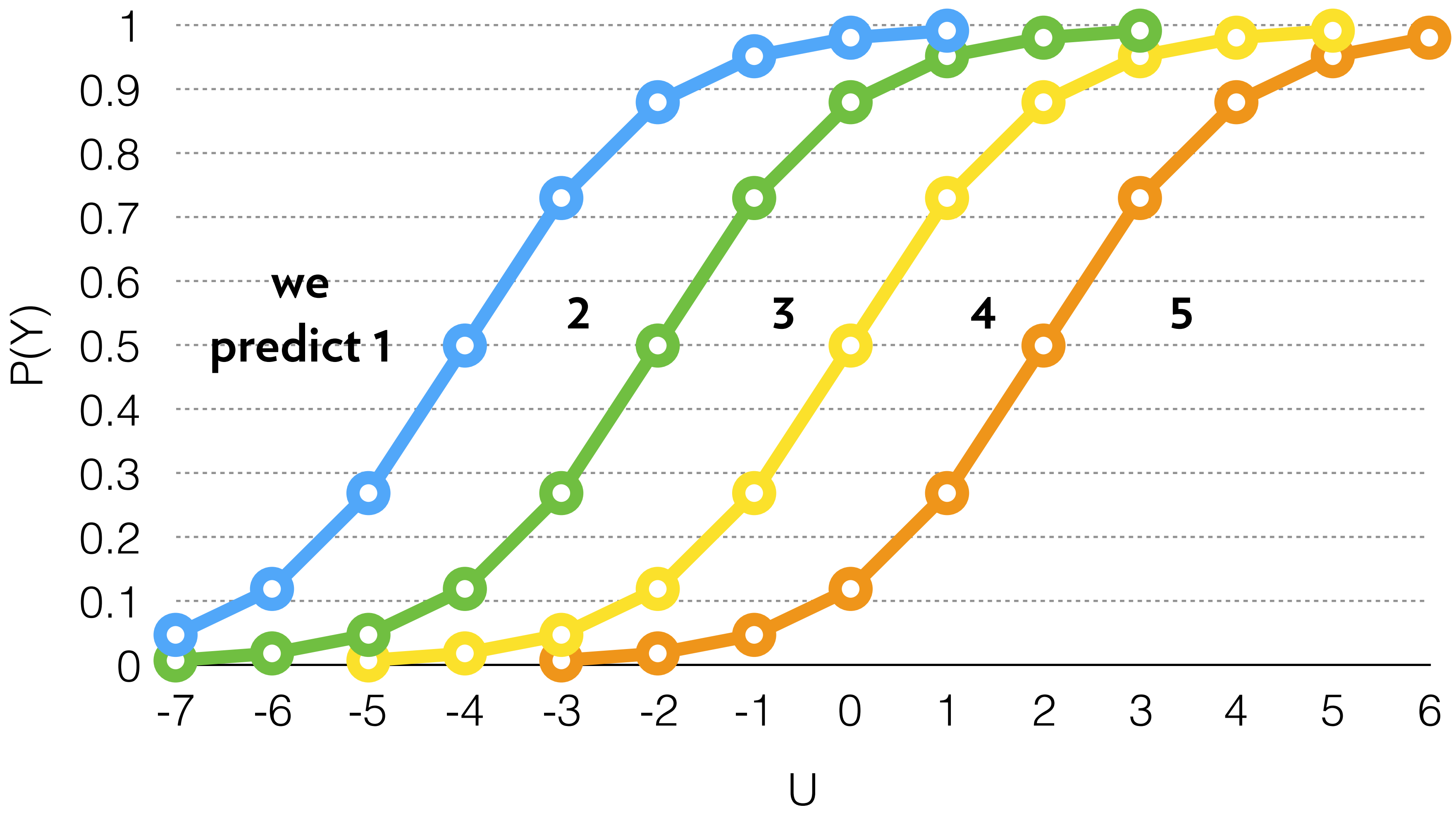
Also, likely not very normally distributed!

How can we solve these problems?

Logistic regression



Ordered logistic



Coefficients

The model estimates intercepts for each threshold

1|2, 2|3, 3|4, 4|5

These thresholds are the **log odds** of any person having **at least** this value

How to interpret the b coefficients?

e^b is the **odds ratio** for a 1pt increase in X

e.g. if the odds ratio is 1.40, then the odds of a higher value increase by 40% if X is 1 higher

Bootstrapping

as a way to solve problems with normality

Bootstrapping

What if we have problems? (e.g. heteroscedasticity, non-normality, outliers, non-linearity)

- Use bootstrapping!

In CFA and SEM, we are going to use bootstrapped results by default

Bootstrapping

1. Treat your sample like a population of size N

2. Sample N items (with replacement!) from this population

3. Calculate the test statistic on this sample

4. Repeat MANY times (like, 1000 times)

5. Get the SE and confidence interval from the data

   For 95% CI: order the data, take the value of item # 25 (lower bound) and item # 975 (upper bound)

“It is the mark of a truly intelligent person to be moved by statistics.”

THANKS!

George Bernard Shaw