Statistical Procedures Covered in Stats II

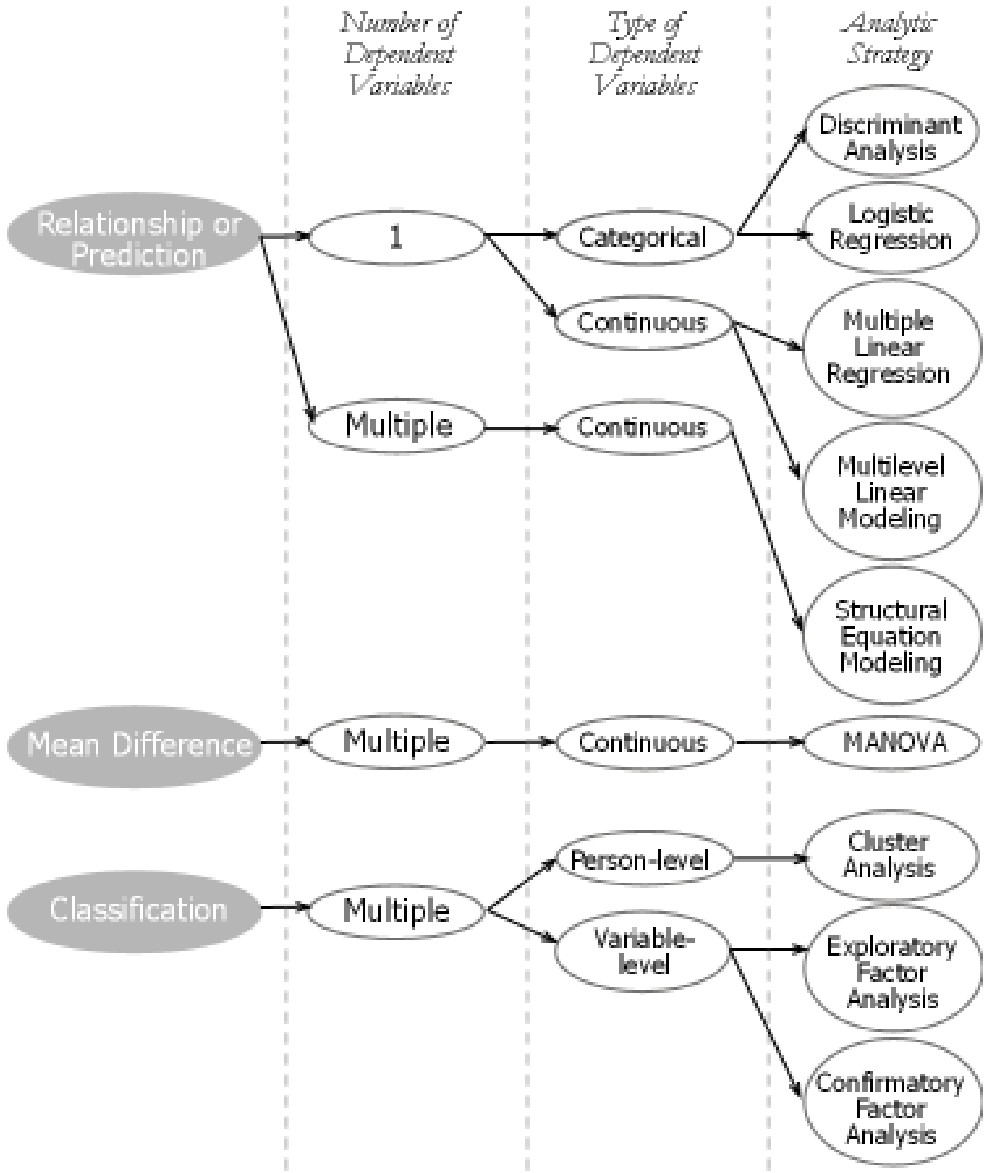
- Ordinary Least Squares regression
- Logistic regression
- Cluster analysis
- Exploratory factor analysis
- Path analysis, confirmatory factor analysis, and structural equation modeling

Correlation & Simple Linear Regression

Statistics I
Week 10

Basic Statistics for Continuous Variables

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated (and small sample size):
	independent	correlated	
Continuous	**T-test:** compares means between two independent groups **ANOVA:** compares means between more than two independent groups **Pearson's correlation coefficient** (linear correlation): shows linear correlation between two continuous variables **Linear regression:** multivariate regression technique used when the outcome is continuous; gives slopes	**Paired ttest:** compares means between two related groups (e.g., the same subjects before and after) **Repeated-measures ANOVA:** compares changes over time in the means of two or more groups (repeated measurements) **Mixed models/GEE modeling:** multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time	Non-parametric statistics **Wilcoxon sign-rank test**: non-parametric alternative to the paired ttest **Wilcoxon sum-rank test** (=Mann-Whitney U test): non-parametric alternative to the ttest **Kruskal-Wallis test:** non-parametric alternative to ANOVA **Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient


Number of Dependent Variables
Type of Dependent Variables
Analytic Strategy
Relationship or Prediction
1
Multiple
Categorical
Continuous
Continuous
Discriminant Analysis
Logistic Regression
Multiple Linear Regression
Multilevel Linear Modeling
Structural Equation Modeling
Mean Difference
Multiple
Continuous
MANOVA
Classification
Multiple
Person-level
Variable-level
Cluster Analysis
Exploratory Factor Analysis
Confirmatory Factor Analysis

Correlation analysis: a good intro to regression

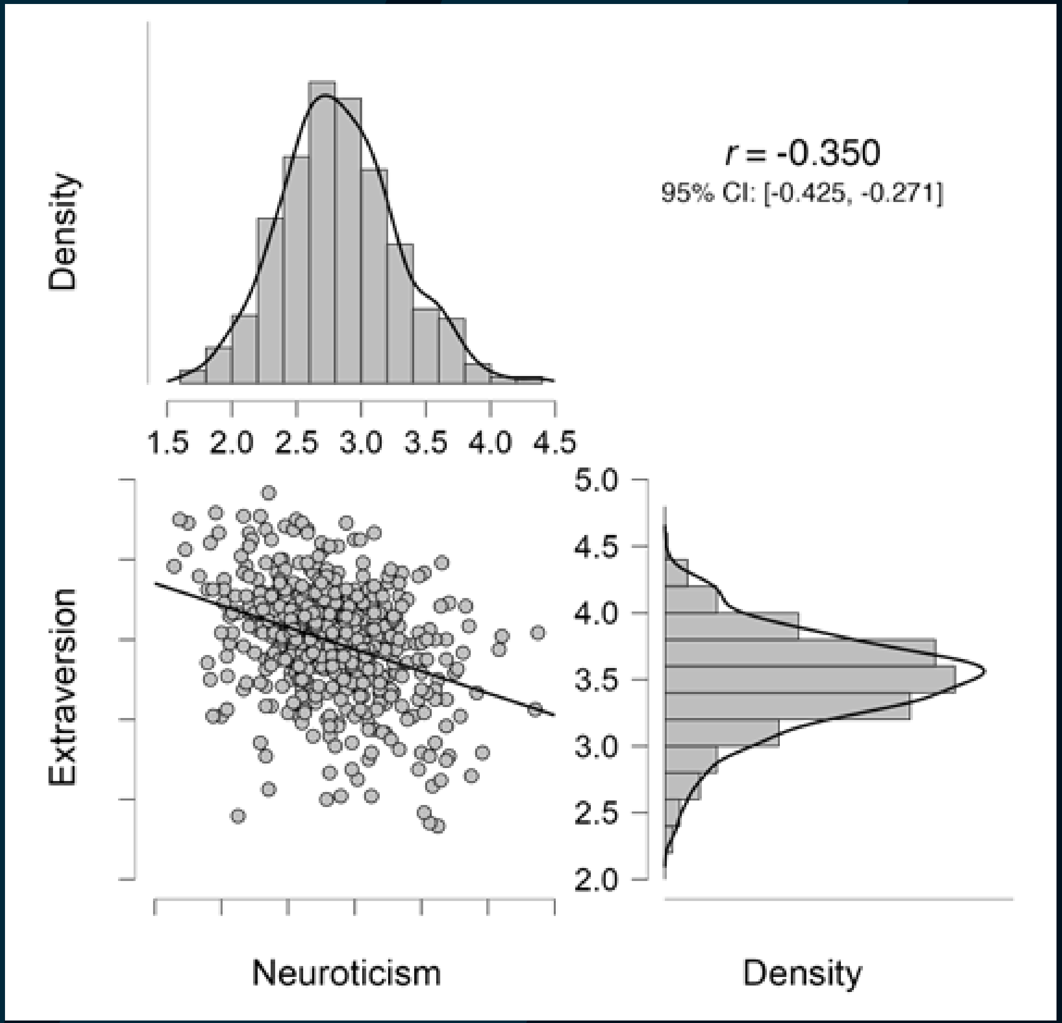
- Measures the degree of **linear relationship** between two continuous variables, x and y
- We have a **linear relationship** between x and y if a straight line drawn through the points provides the most appropriate approximation to the observed relationship
- We measure how close the observations are to the straight line that best describes their linear relationship by calculating the **Pearson product moment correlation coefficient**, usually simply called the **correlation coefficient, denoted "*r*" or "rho" (ρ)**

Faulty Analyses

- Correlation does not equal causation
- There are rules that determine causal effects
- Data is often incorrectly analyzed in addition to being presented in different ways

Confusing Correlation and Causation

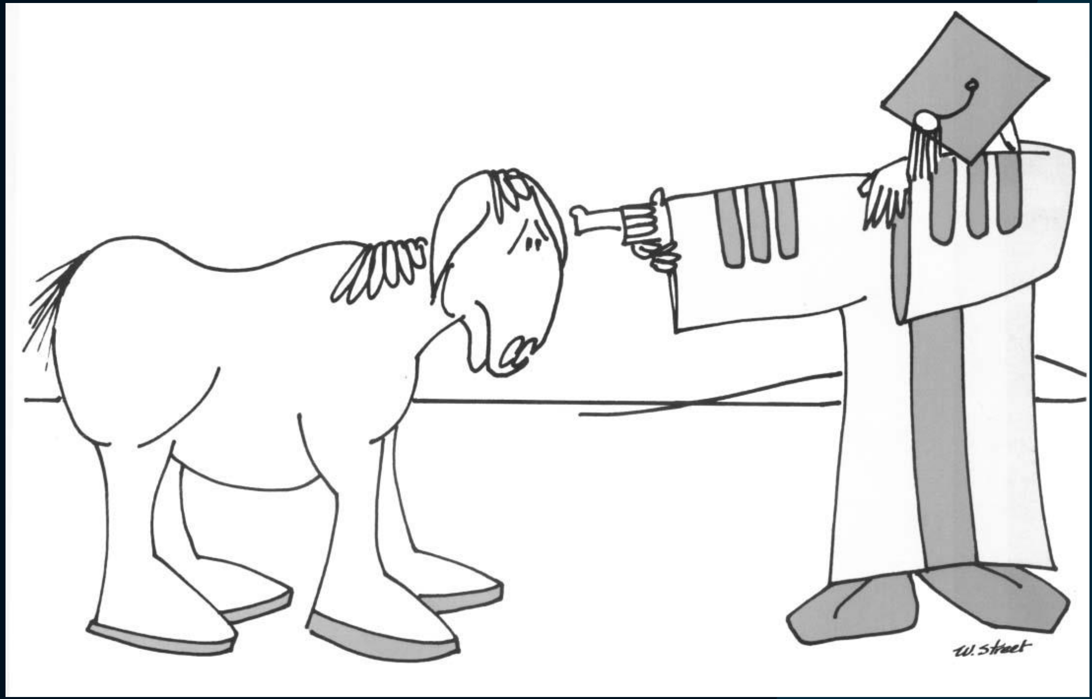
- Correlation = covariation (co-occurrence of change on two variables)
  - This tells us nothing about cause (why the two variables changed)
- Causation: Change in one variable RESULTS IN change in another



Two of the "Big 5"

Correlation & Causation (use common sense)

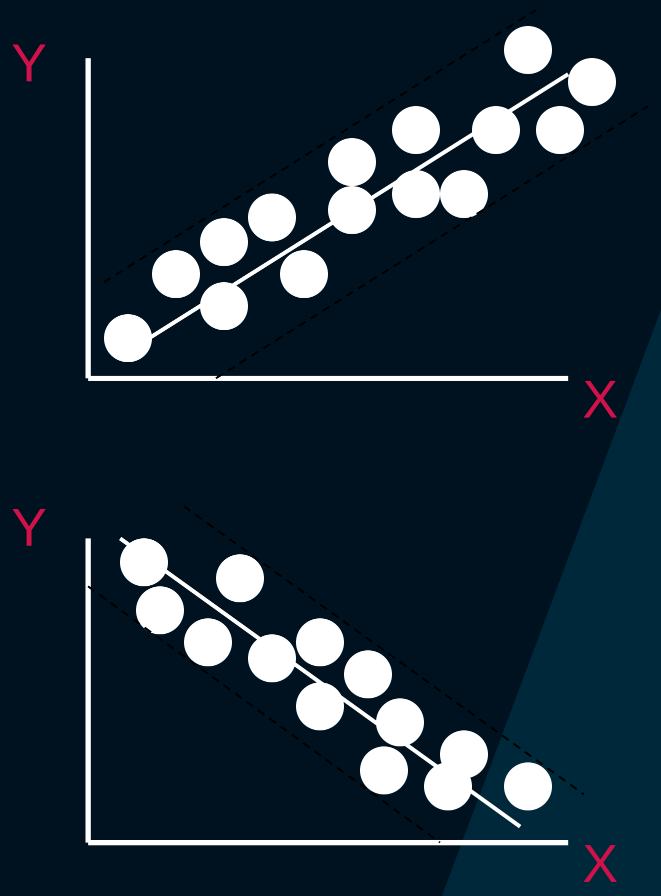
"There's a significant NEGATIVE correlation between the number of mules and the number of academics in a state, but remember, correlation is not causation"



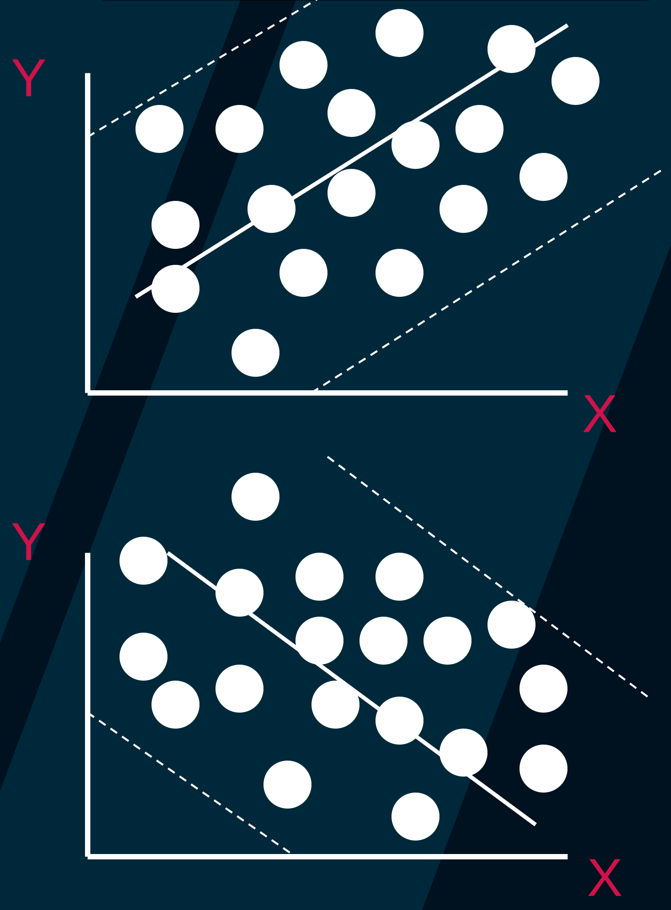
- Often used as means for prediction, correlation tells us how related two variables are
- **Note** : the 'correlation dne causation' applies more broadly than people assume
  - Ex. Regression analysis is a method of prediction, but it does not imply causation, but merely correlation (albeit a partial correlation)
- The question becomes: what other variable is 'causing' the observed relationship
- When might we be more confident that an observed correlation is causal?

Scatter Plots of Various Correlations

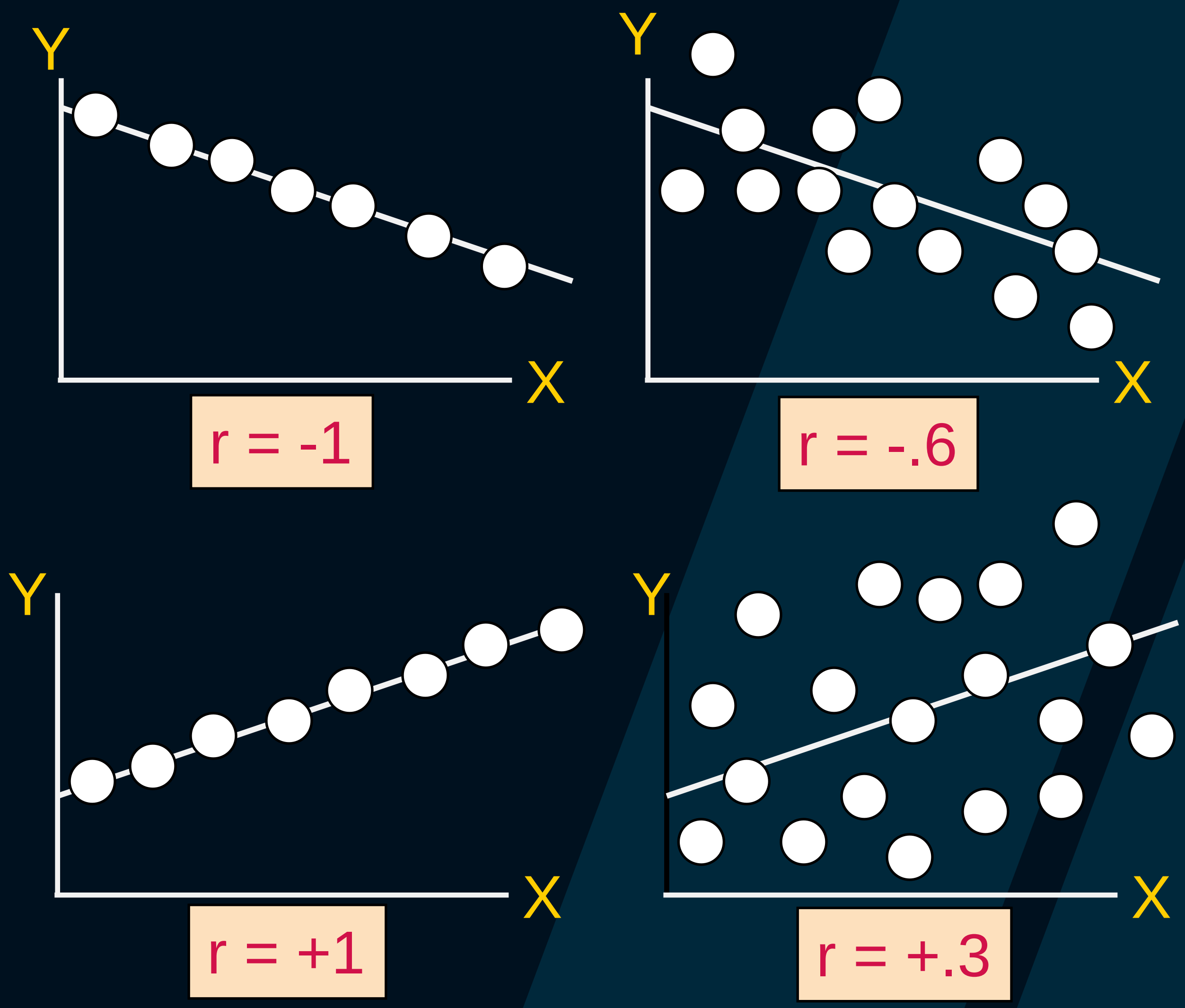
Strong relationships



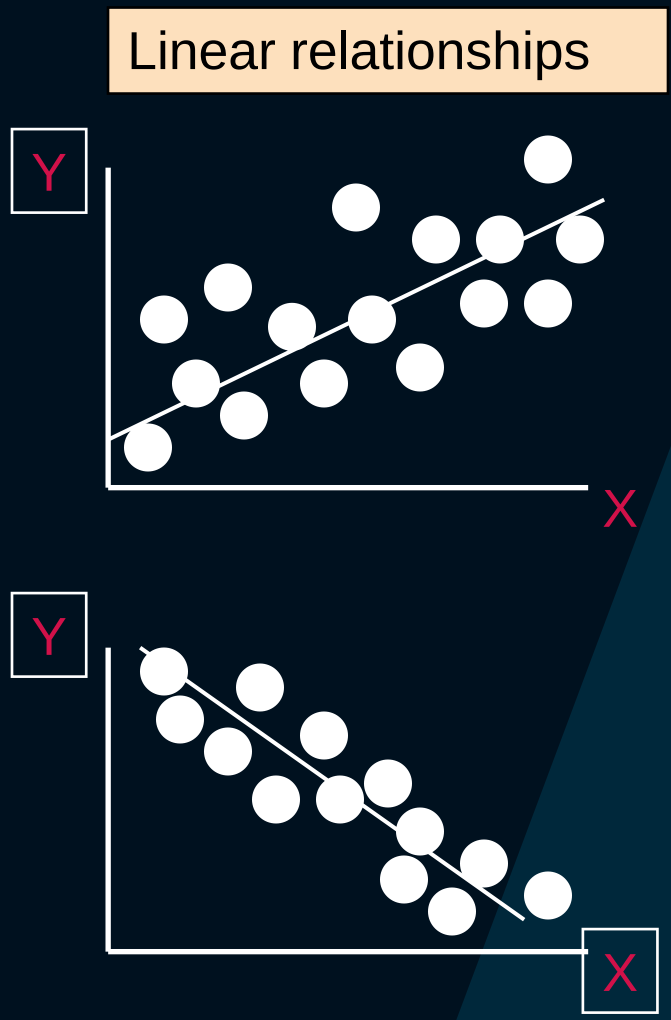
Weak relationships

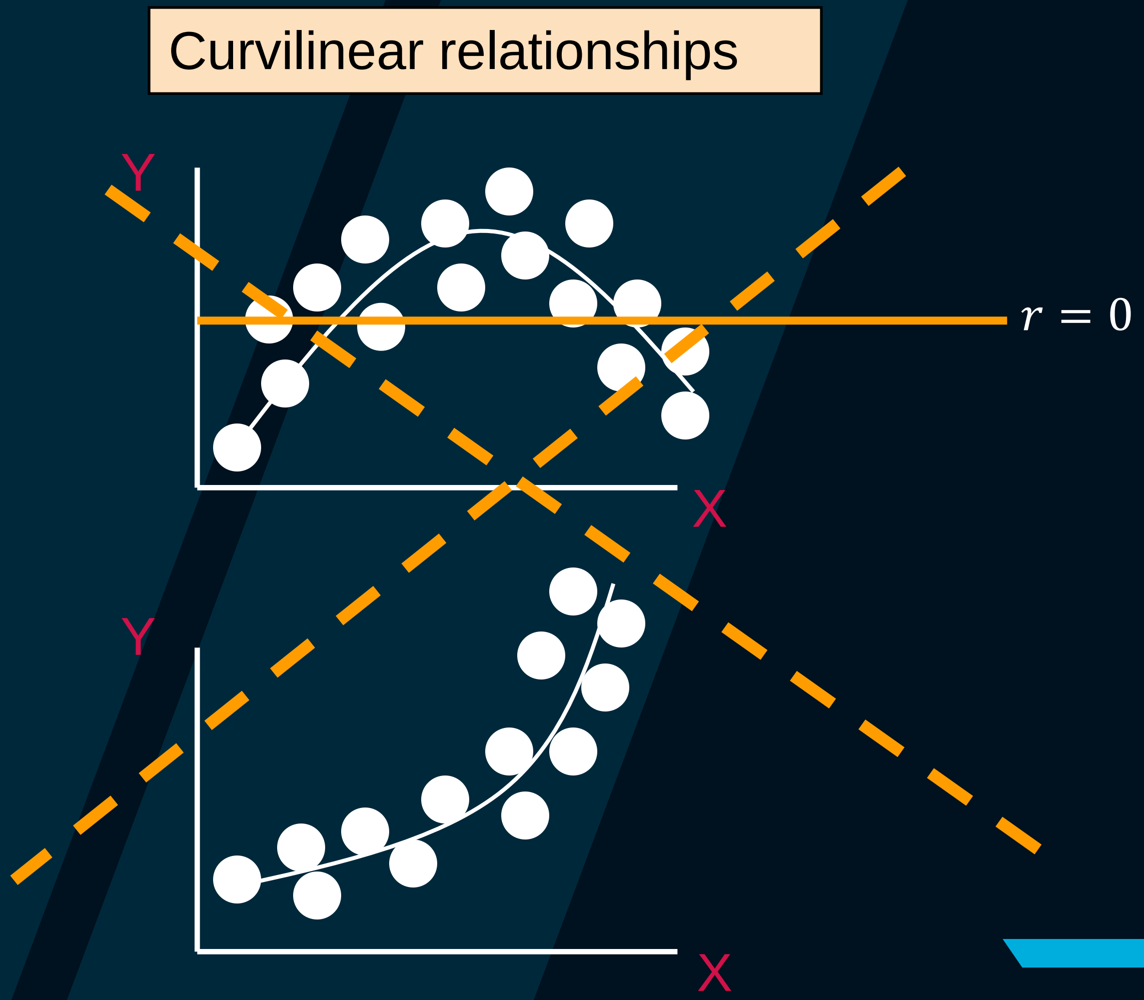


Scatter Plots of Various Correlations



Scatter Plots of Various Correlations





Demonstrating Causation

- 4 requirements to logically infer a causal relationship

1. Covariation--statistical association: if A changes, B must also change. This is <u>necessary</u>, but not <u>sufficient</u>.
   - Not enough alone to show cause.
2. Time order--IV must come before DV
   - big problem in cross-sectional surveys.
   - Note: sometimes you can make reasonable inferences
3. Nonspurious--no third factor can explain the covariation
   - Ice cream and violent crime
4. Theory--logical explanation of the relationship

Common misinterpretations I see in my work

- Poverty is a major source of child welfare involvement
- Children who are abused have higher risk of adult criminal behavior
- Victims of domestic violence are more likely to be housing insecure (i.e., be evicted)
- Homeless individuals are likely to have physical and mental health problems
- People who are highly educated make more money
- Drug overdoses increased during the COVID-19 pandemic

General Requirements

- Two or more continuous variables,
- Not necessarily directional (one causes the other)
- Linear Relationship (or at least ordinal)

ID	Age	Score
1	8	7
2	6	2
3	9	6
4	7	6
5	7	8
6	8	5
7	5	3
8	5	5

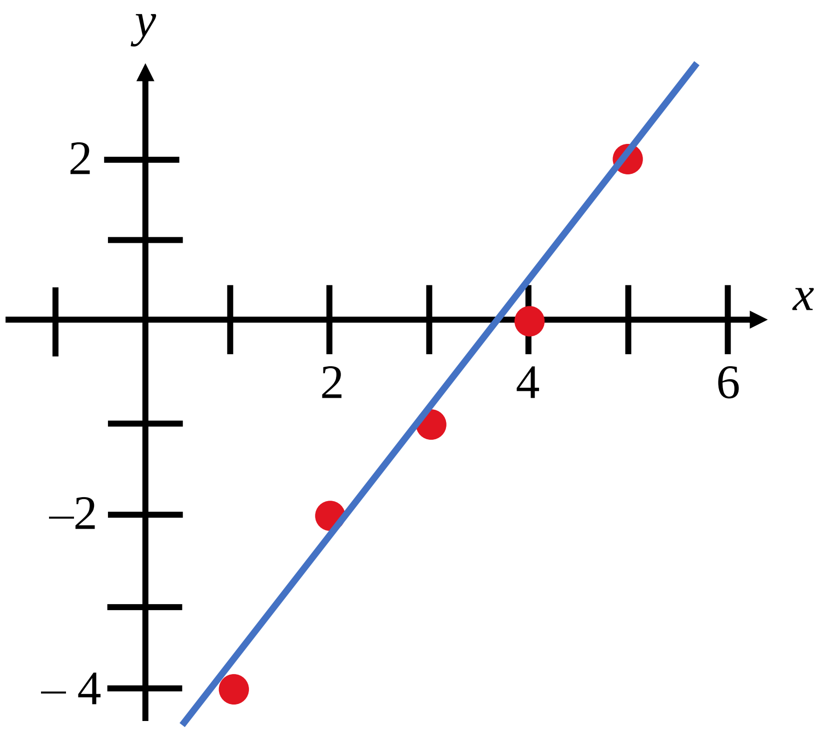
Correlation

A **correlation** is a relationship between two variables. The data can be represented by the ordered pairs (x, y) where x is the **independent** (or **explanatory**) **variable**, and y is the **dependent** (or **response**) **variable**.

A **scatter plot** can be used to determine whether a **linear** (straight line) v. some other relationship exists between two variables.

Example:

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



Covariance

$$cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

Calculate the covariance between these two variables

x	1	2	3	4	5
y	– 4	– 2	– 1	0	2

cov(X,Y) > 0 → X and Y are positively correlated

cov(X,Y) < 0 → X and Y are inversely correlated

cov(X,Y) = 0 → X and Y are independent, or uncorrelated

Correlation = Standardized Covariance

$$\hat{r} = \frac{COV(x,y)}{\sqrt{\operatorname{var} x}\,\sqrt{\operatorname{var} y}} =$$

$$\frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

Simpler calculation formula…

$$\hat{r} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\cancel{n-1}}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\cancel{n-1}}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\cancel{n-1}}}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerator of covariance

$$\boxed{\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}}$$

Numerators of variance

Yet another…

The formula can also be expressed as follows, by manipulating the terms

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$$

The range of the correlation coefficient is –1 to 1. If *x* and *y* have a strong positive linear correlation, *r* is close to 1. If *x* and *y* have a strong negative linear correlation, *r* is close to –1. If there is no linear correlation or a weak linear correlation, *r* is close to 0.

Calculating a Correlation Coefficient

Calculating a Correlation Coefficient

In Words	*In Symbols*
1. Find the sum of the x-values.	$\sum x$
2. Find the sum of the y-values.	$\sum y$
3. Multiply each x-value by its corresponding y-value and find the sum.	$\sum xy$
4. Square each x-value and find the sum.	$\sum x^2$
5. Square each y-value and find the sum.	$\sum y^2$
6. Use these five sums to calculate the correlation coefficient.	$r = \dfrac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$

Correlation Coefficient

Example:

Calculate the correlation coefficient *r* for the following data.

x	y	xy	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} = \frac{5(9) - (15)(-1)}{\sqrt{5(55) - 15^2}\sqrt{5(15) - (-1)^2}}$$

$$= \frac{60}{\sqrt{50}\sqrt{74}} \approx 0.986$$

There is a strong positive linear correlation between *x* and *y*.

Correlation Coefficient

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Display the scatter plot.

b.) Calculate the correlation coefficient *r*.

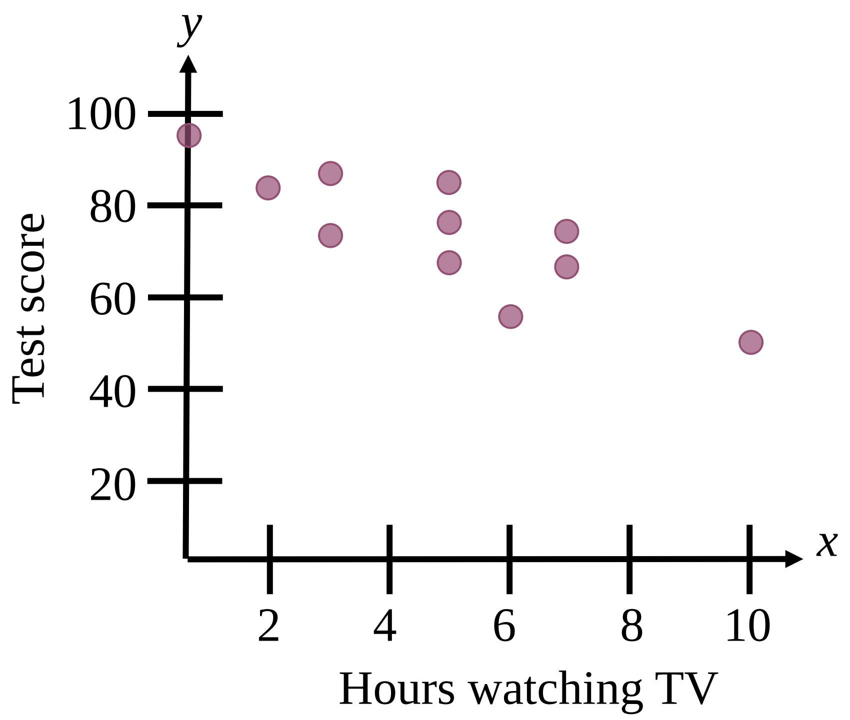
Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Continued.

Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50



Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$\sum x = 54$

$\sum y = 908$

$\sum xy = 3724$

$\sum x^2 = 332$

$\sum y^2 = 70836$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{12(3724) - (54)(908)}{\sqrt{12(332) - 54^2}\sqrt{12(70836) - (908)^2}} \approx -0.831$$

There is a strong negative linear correlation.

As the number of hours spent watching TV increases, the test scores tend to decrease.

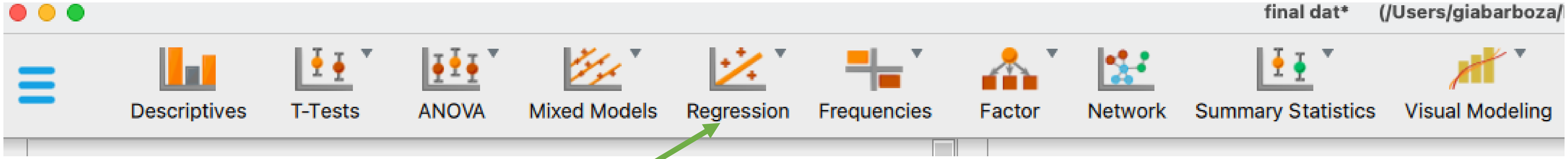
Let’s apply

- correlation-example-big5.JASP
- Open the document handout-bringing-it-full-circle.doc
- We will re-do these analyses in SPSS (nscaw-subset.sav)
- Then we will do them using JASP (nscaw-subset.jasp)

Correlation in SPSS/JASP

- We are going to explore correlations between symptoms of PTSD following a traumatic experience
- The data asked over 2000 foster youth about Potentially Traumatic Experiences (PTEs)
  - PTEs are sexually molested, kidnapping, badly injured or killed, watched badly injured or killed, physically attacked, threatened with a weapon
- Symptoms are *re-experiencing* (intrusive thoughts, distressing dreams, flashbacks, psychological distress, physiological reactivity), *arousal problems* (sleep, irritability, concentration, hypervigilance, startle), *avoidance* (avoid thoughts and places) and *emotional numbing* (diminished interest, feeling detached, restricted affect and foreshortened future)
- **File: PTSD_foster_care.csv**

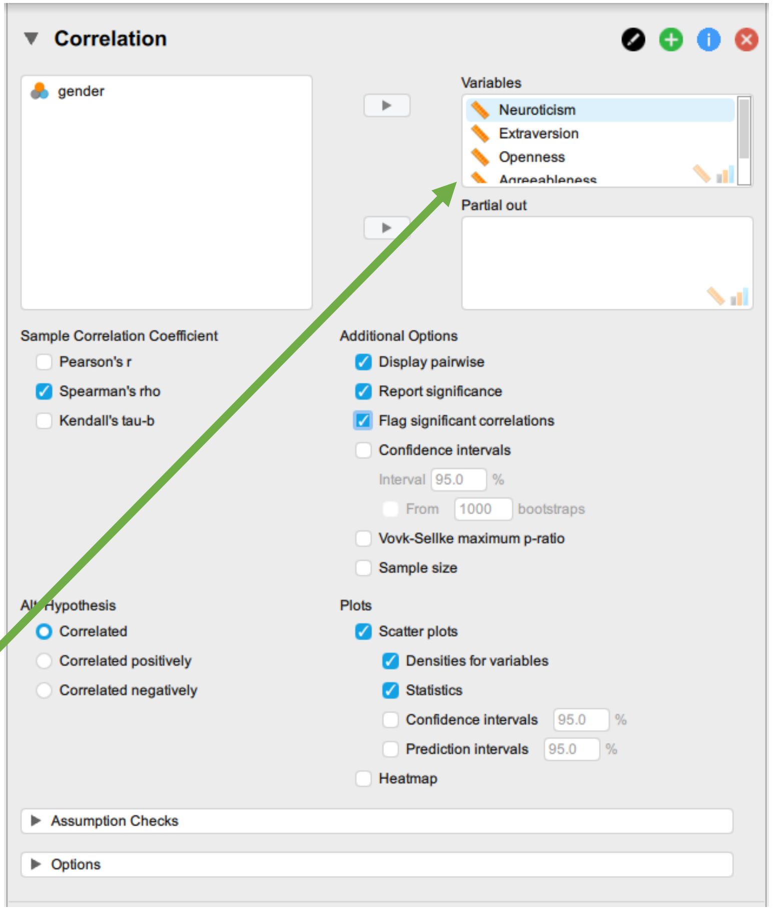
Compute the Test Statistic



Click on
“Regression Tab → Correlation”

Compute the Test Statistic

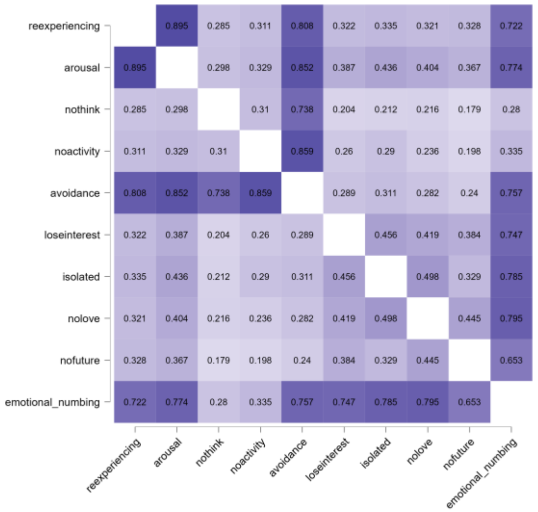
Results



Bring variables to be correlated over here

Spearman's Correlations

			Spearman's rho	p
reexperiencing	-	arousal	0.895	< .001
reexperiencing	-	nothink	0.285	< .001
reexperiencing	-	noactivity	0.311	< .001
reexperiencing	-	avoidance	0.808	< .001
reexperiencing	-	loseinterest	0.322	< .001
reexperiencing	-	isolated	0.335	< .001
reexperiencing	-	nolove	0.321	< .001
reexperiencing	-	nofuture	0.328	< .001
reexperiencing	-	emotional_numbing	0.722	< .001
arousal	-	nothink	0.298	< .001
arousal	-	noactivity	0.329	< .001
arousal	-	avoidance	0.852	< .001
arousal	-	loseinterest	0.387	< .001
arousal	-	isolated	0.436	< .001
arousal	-	nolove	0.404	< .001
arousal	-	nofuture	0.367	< .001
arousal	-	emotional_numbing	0.774	< .001



Compute an Effect Size and Describe it

One of the main effect sizes for correlation is r^2

$$r^2 = (r)^2$$

r^2	Estimated Size of the Effect
Close to .01	Small
Close to .09	Moderate
Close to .25	Large

Interpreting the results

Example write-up

Correlations were computed among symptoms are *re-experiencing*, *arousal problems*, *avoidance* and *emotional numbing* in over 2,000 foster care youth who experienced a potentially traumatic event. The correlations between all symptom pairs were statistically significant at the alpha = .05 level. For example, the correlation coefficient between re-experiencing and arousal was positive and highly significant indicating that as re-experiencing increases so do symptoms of arousal (r = .832, p < .001). Etc.

Correlation

Pearson's Correlations

			Pearson's r
reexperiencing	-	arousal	0.832***
reexperiencing	-	avoidance	0.767***
reexperiencing	-	emotional_numbing	0.700***
arousal	-	avoidance	0.799***
arousal	-	emotional_numbing	0.750***
avoidance	-	emotional_numbing	0.696***

* p < .05, ** p < .01, *** p < .001

Overview of next 3-4 weeks

- Ordinary Least Squares
  - Simple linear regression
    - Mechanics (today)
  - Multiple linear regression
  - Power
  - Effect Size
  - Assumptions
- Examples and Applications

Continuous Variables

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated (and small sample size):
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	**T-test:** compares means between two independent groups **ANOVA:** compares means between more than two independent groups **Pearson's correlation coefficient** (linear correlation): shows linear correlation between two continuous variables **Linear regression:** multivariate regression technique used when the outcome is continuous; gives slopes	**Paired ttest:** compares means between two related groups (e.g., the same subjects before and after) **Repeated-measures ANOVA:** compares changes over time in the means of two or more groups (repeated measurements) **Mixed models/GEE modeling**: multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time	<u>Non-parametric statistics</u> **Wilcoxon sign-rank test**: non-parametric alternative to the paired ttest **Wilcoxon sum-rank test** (=Mann-Whitney U test): non-parametric alternative to the ttest **Kruskal-Wallis test:** non-parametric alternative to ANOVA **Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient

Uses of Regression Analysis

- Regression analysis serves 3 major purposes:
  1. Description
  2. Prediction
  3. Control → ceteris paribus

- The several purposes of regression analysis frequently overlap in practice

Mechanics

- Simple regression analysis is a statistical tool that gives us the ability to estimate the relationship between a dependent variable (usually called *y*) and an independent variable (usually called *x*).
  - The dependent variable is the variable for which we want to make a prediction
  - While various non-linear forms may be used, simple linear regression models are the most common (and the most useless)

Introduction: Linear Regression

- Aims to predict the value of an outcome (e.g., violence, PTSD, etc), *Y*, based on the value of one or more explanatory variables, $X_1, \ldots X_n$.
- Two main questions
  - What is the relationship between the Y and Xs, on <u>average</u>?
    - The analysis “models” this as a line
    - We care about “slope”—size, direction
    - Slope = 0 corresponds to “no association” (just as in correlation)
  - How precisely can we predict Y conditional on certain values of the independent variables?
    - Another way to say this: How much of the variation in Y is being explained by X?

Linear regression –Terminology

- Outcome, Y
  - Dependent variable
  - Response variable
- Explanatory variable (predictor), X
  - Independent variable
  - Covariate

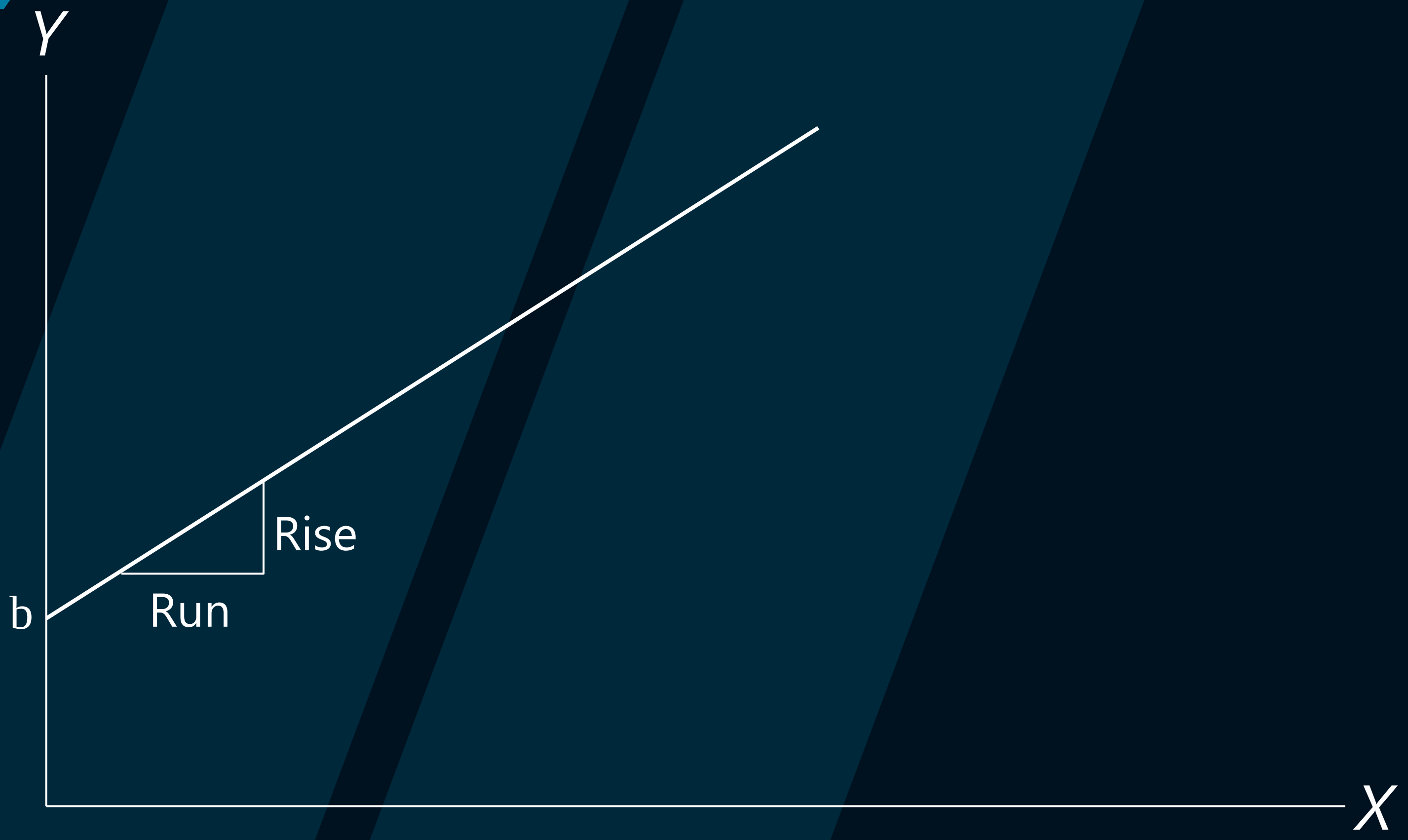
Functional Form

- The first order linear model

$$Y = mX + b$$

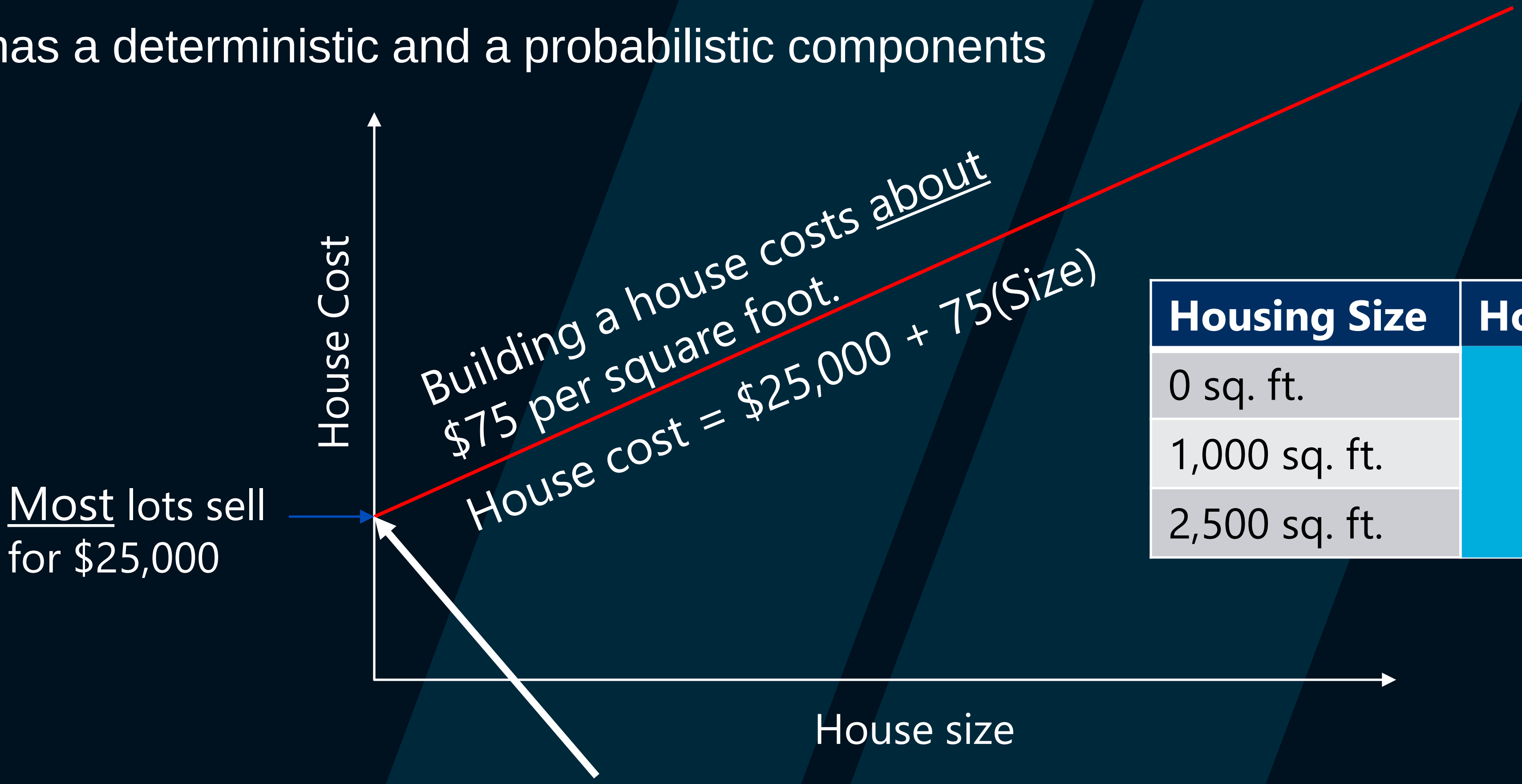
Y = dependent variable
X = independent variable
b = Y-intercept
m = slope of the line

Define b, and m



Functional Form

The model has a deterministic and a probabilistic components



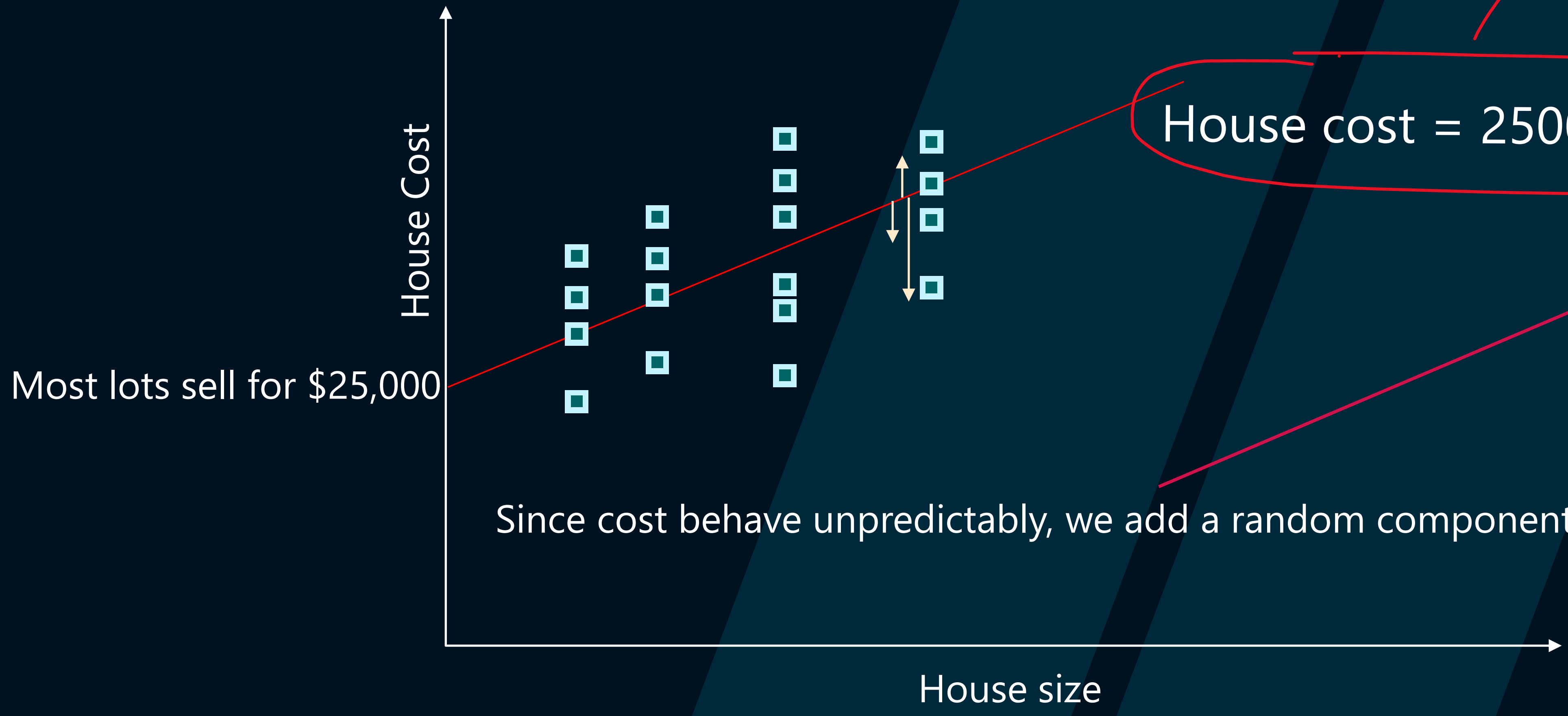
Housing Size	Housing Cost
0 sq. ft.	?
1,000 sq. ft.	
2,500 sq. ft.	

What is the expected value of the house when the house size is 0?

However, house cost vary even among same size houses!

Deterministic

$$\text{House cost} = 25000 + 75\,(\text{Size}) + \varepsilon$$



Since cost behave unpredictably, we add a random component.

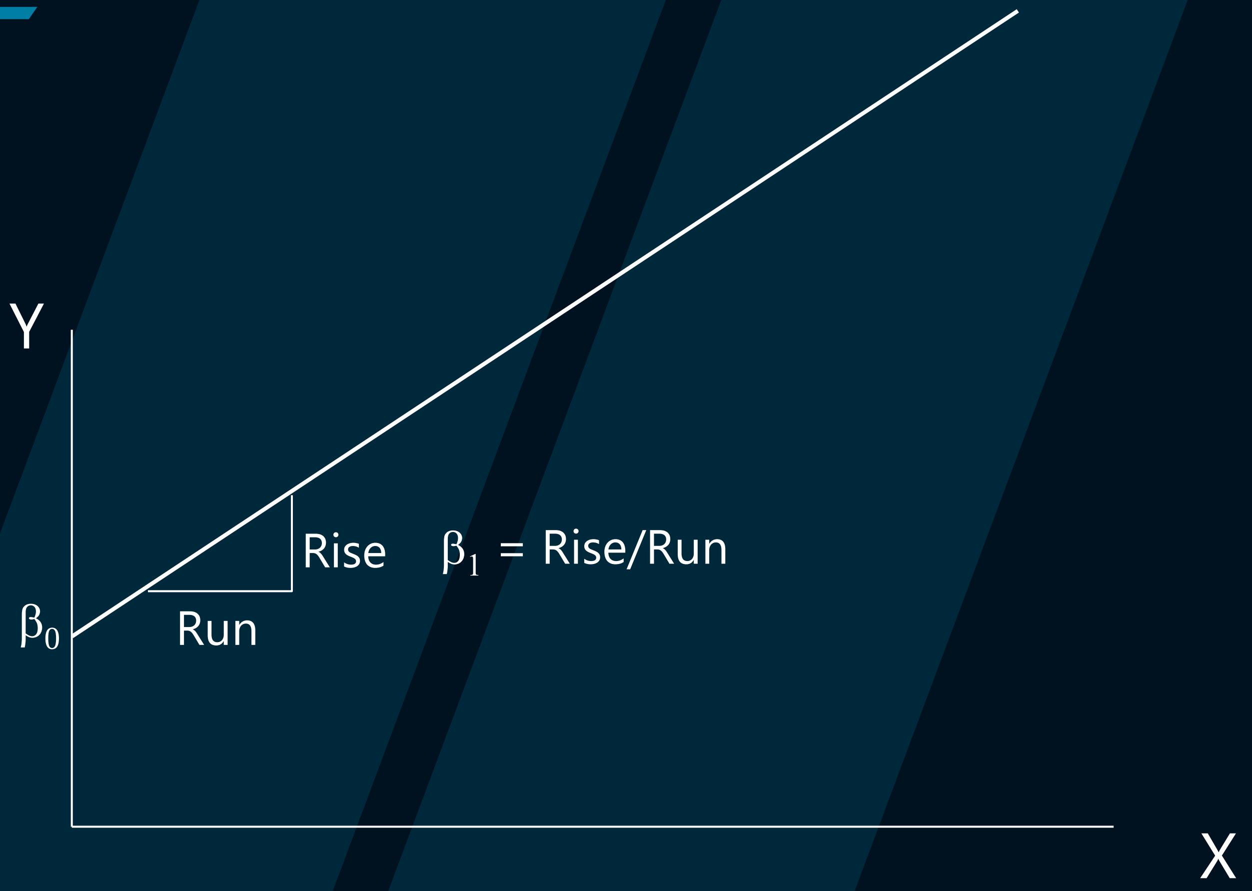
What types of variables might be part of the error term?

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0 and β_1 are unknown population parameters, therefore are estimated from the data.

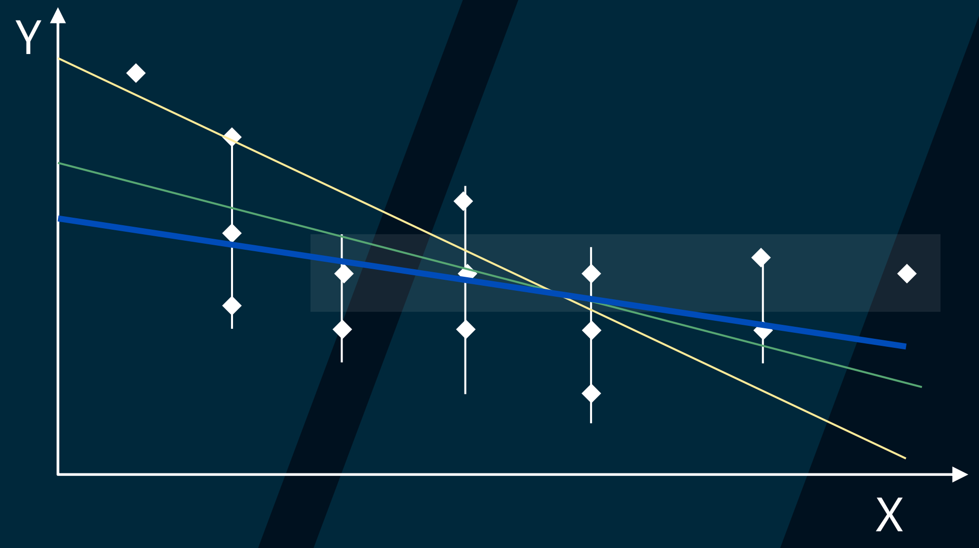
- The first order linear model

Y = dependent variable
X = independent variable
β_0 = Y-intercept (b)
β_1 = slope of the line (m)
ε = error variable



Estimating the model parameters

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts through the data.



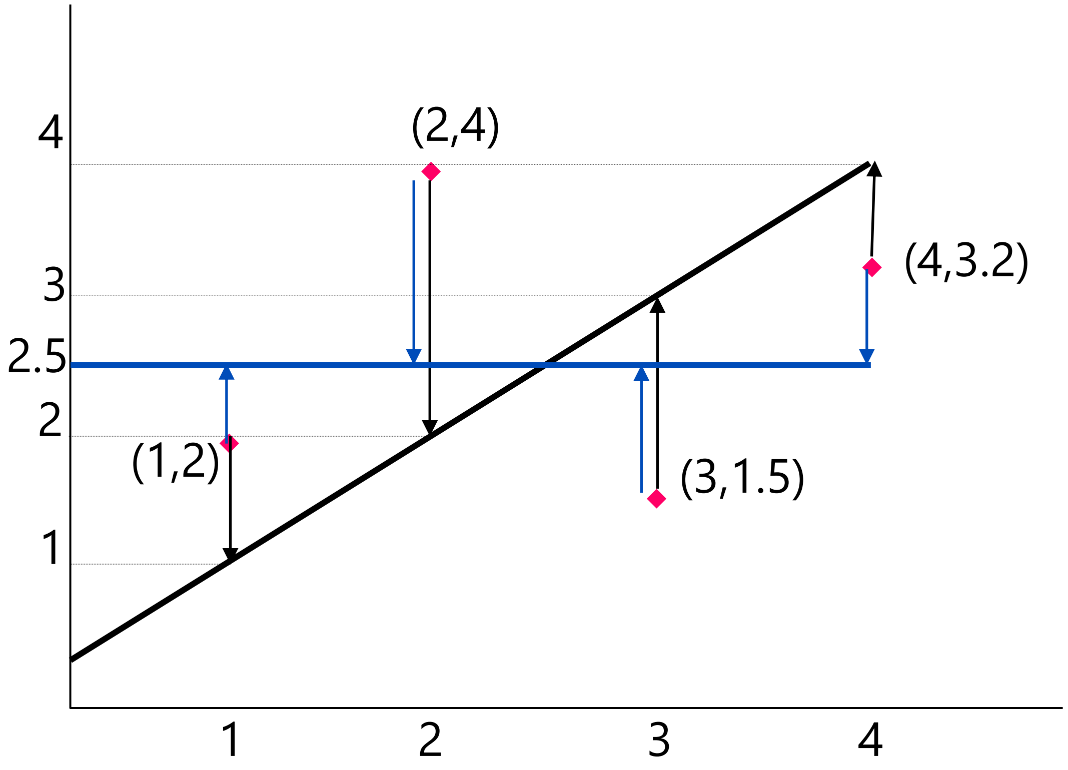
Question: What should be considered a good line?

The Least Squares (Regression) Line

A good line is one that minimizes the sum of squared differences between the points and the line.

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$



The smaller the sum of squared differences the better the fit of the line to the data.

Formal statement of the model

- If the scatter plot of our sample data suggests a linear relationship between two variables then we can summarize the relationship by drawing a straight line between the variables:

$$Y = \beta_0 + \beta_0 X + \varepsilon$$

- The values of the regression parameters β_0, and β_1 are not known. We estimate them from data.
  - β_0 indicates the intercept of Y
  - β_1 indicates the change in the mean response in Y per unit increase in X
- X is a known independent variable
- Deviations ε are independent, identically distributed $N(0, \sigma^2)$
- **Least squares** method give us the "best" estimated line for our set of sample data.

Let’s see how this is applied

- simple-linear-regression.JASP

What Multiple Linear Regression Is and How It Works: Characteristics

- Two or more predictor variables and one criterion variable
- Additional variables are incorporated via partial and semipartial correlations
  - *Partial correlation*: As a three-variable example, represents the linear relationship between two variables, say X_1 and X_2, independent of the linear influence of X_3

What Multiple Linear Regression Is and How It Works: Characteristics

- Unstandardized regression model
  - $Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + \ldots + B_m X_{mi} + e_i$
  - Y is the criterion variable
  - X_k's are the predictor (or independent) variables where $k = 1, \ldots, m$
  - B_k is the sample partial slope of the regression line for Y as predicted by X_k
  - a is the sample intercept of the regression line for Y as predicted by the set of X_k's
  - e_i are the residuals or errors of prediction (the part of Y not predictable from the X_k's
  - i represents an index for an individual or object
- Sample prediction model is
  - $Y'_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + \ldots + B_m X_{mi}$
- Difference in the models:
  - Regression model explicitly includes prediction error as e_i
  - Prediction model incorporates error as part of prediction, it is not explicitly modeled

What Multiple Linear Regression Is and How It Works: Characteristics

- Standardized regression model
- Sample standardized linear prediction model

$$z(Y_i') = b_1^* z_{1i} + b_2^* z_{2i} + \ldots + b_m^* z_{mi}$$

  - b_i^* = sample standardized partial slope
  - no intercept term is necessary as the mean of the *z* scores for all variables is 0
- Unstandardized or standardized model?
  - Standardized model is not stable from sample to sample

What Multiple Linear Regression Is and How It Works: Characteristics

- Coefficient of multiple determination and multiple correlation
  - Tells the proportion of total variation in the dependent variable *Y* that is *explained by* the set of predictor variables
  - R-squared = .45 → 45% of the variation in *Y* is explained by *X*s
    - <u>Example</u>: if we predict GGPA from UGPA and our R-squared is .45
  - 1 – R-squared

What Multiple Linear Regression Is and How It Works: Characteristics

- Significance tests
  - Overall: Test of significance of the overall regression model, or alternatively the test of significance of the coefficient of multiple determination
    - If H_0 is rejected, then *one or more of the individual regression coefficients is statistically significantly different from zero*
  - Ceteris paribus: Test of the statistical significance of each individual partial slope or regression coefficient, b_k

What Multiple Linear Regression Is and How It Works: Characteristics

- ~~Methods of entering predictors~~
  - ~~Simultaneous~~
  - ~~Backward elimination~~
    - ~~Forward selection~~
    - ~~Stepwise selection~~
    - ~~All possible subsets regression~~
    - ~~Hierarchical regression~~

What Multiple Linear Regression Is and How It Works: Characteristics

- Nonlinear relationships
  - Polynomial models
    - First degree polynomial (simple linear regression)
    - Second degree polynomial (quadratic)
    - Third degree polynomial (cubic)

    - <u>Example</u>: entering age (first degree), age-squared (quadratic) or age-cubed (3^{rd} degree)

What Multiple Linear Regression Is and How It Works: Characteristics

- Interactions
  - An interaction can be defined as occurring when the relationship between Y and X_1 depends on the level of X_2
    - In other words, X_2 is a moderator variable

What Multiple Linear Regression Is and How It Works: Characteristics

- Categorical predictors
  - Dummy coding
  - When there are more than two categories to the categorical predictor, multiple dummy coded variables must be created—*specifically one minus the number of levels or categories of the categorical variable*

What Multiple Linear Regression Is and How It Works

Sample Size

What Multiple Linear Regression Is and How It Works: Sample Size

- Sample size considerations differ depending on the research goal—testing a hypothesis test estimating a parameter (Algina & Olejnik, 2000; Maxwell, 2000)
  - Larger sample sizes needed for estimation (e.g., Pedhazur, 1997)
- Sample size increases as the squared multiple correlation coefficient diminishes (Knofszynski, 2008)
- Recommendation: estimate power using power software and to consult current advances based on simulation research such as Knofszynski (2008)

What Multiple Linear Regression Is and How It Works

Power

What Multiple Linear Regression Is and How It Works: Power

- In multiple regression, power is a function
  - sample size
  - number of predictors
  - level of significance
  - size of the population effect
- Recommendation: consult power table or power software

What Multiple Linear Regression Is and How It Works

Effect Size

What Multiple Linear Regression Is and How It Works: Effect Size

- Coefficient of multiple determination or multiple correlation coefficient
- Squared multiple correlation coefficient which can be used to compute a globalized f^2

What Multiple Linear Regression Is and How It Works

Assumptions

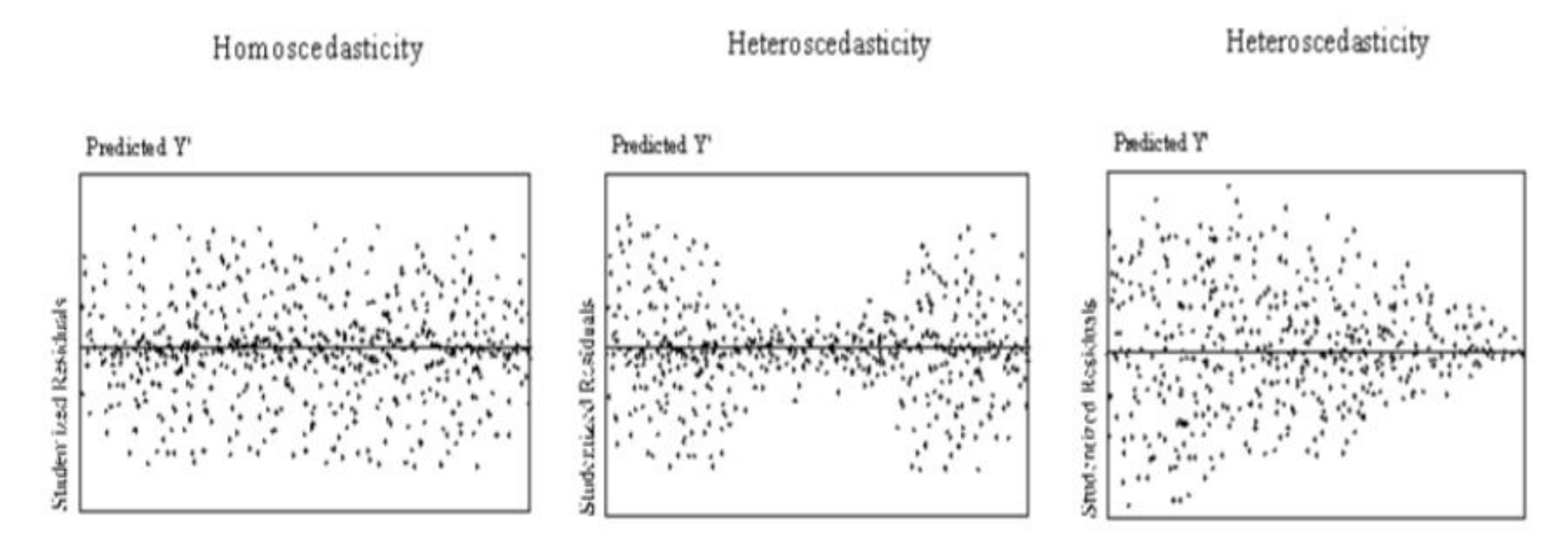
What Multiple Linear Regression Is and How It Works: Assumptions

- Independence
- Homoscedasticity
- Normality
- Linearity
- Fixed *X*
- Noncollinearity

Independence

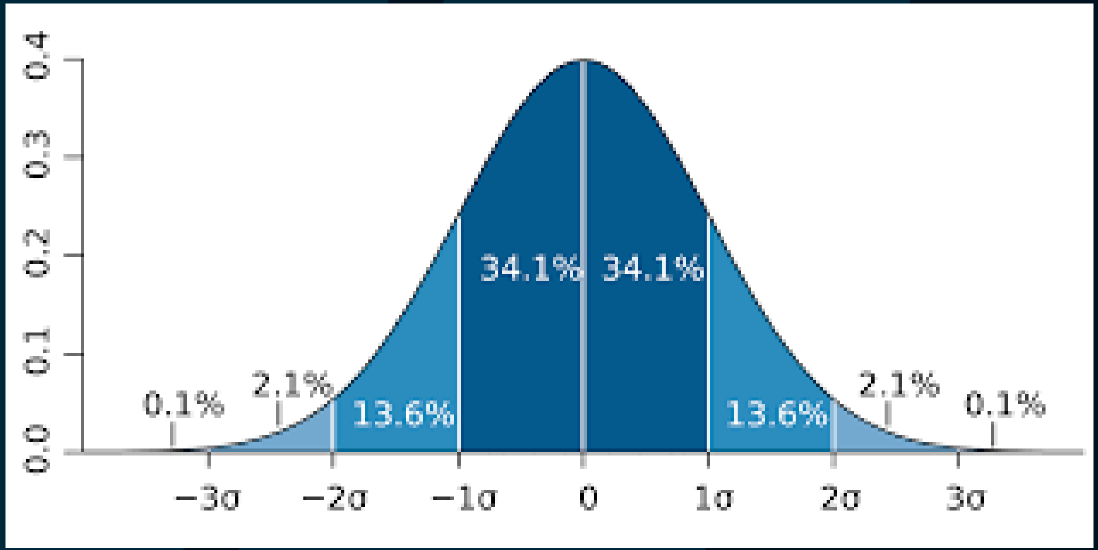
- Due to sampling scheme
- Example: I just reviewed a paper that was exploring the impact of county-level correlates and aggregate county-level perceptions of social workers on the outcomes of youth in EFC
- The authors used OLS
- This is incorrect, why?

Homoscedasticity



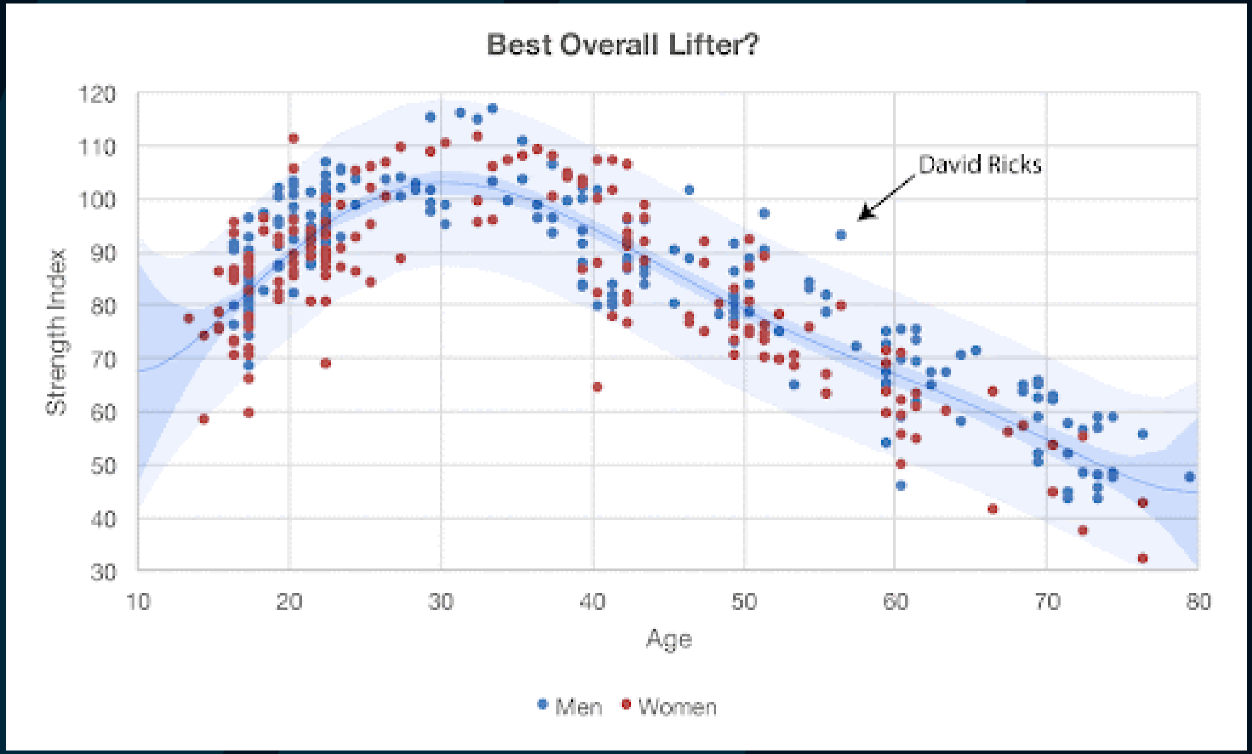
Normality

- Be careful, this assumption is about the prediction errors, not the independent variables
- May be due to outliers however (which are obviously related to the measurement of the independent variables



Linearity

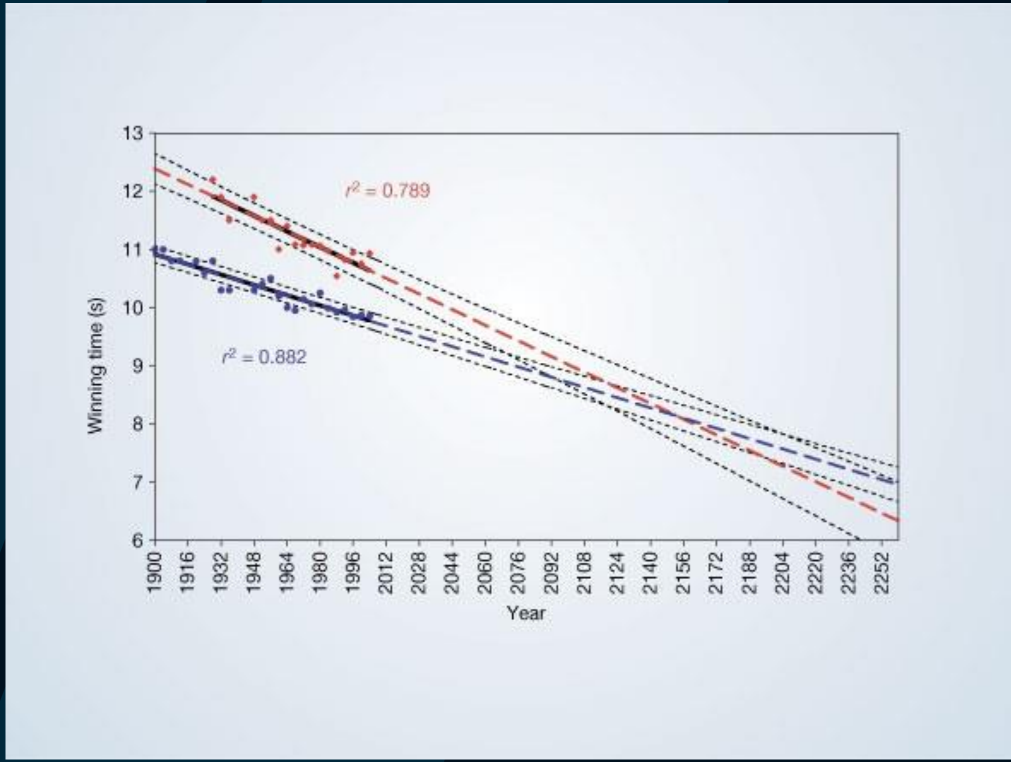
- There is a linear relationship between Y and X
    - Recall correlation
- But, we can model nonlinear relationships
- Why?



Fixed X

- Tatem, A. J., Guerra, C. A., Atkinson, P. M., & Hay, S. I. (2004). Momentous sprint at the 2156 Olympics?. Nature, 431(7008), 525-525.
    - Analyzed athletes' performance on the 100 m dash
    - Found an interesting pattern: the time it took to run it decreased steadily, such that males and females were getting faster and faster over the years
    - Around the year 2156, the lines crossed: sometime mid-century, they predicted, women would outrun men!
    - Even though the lines LOOK convincing, they are not truly capturing the underlying pattern of the data… **why**?

The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.



Noncollinearity

- Less is more
- Always examine correlations – substantial overlap means one should be taken out of the model
- But be weary – I learned that the forward/backwards elimination is not a good strategy, why?
- Start with theory → test hypotheses → draw conclusions

Mathematical Introduction Snapshot

Example: Can I predict your graduate GPA based on your undergraduate GPA & score on the GRE

Note: Be careful about what you CLAIM after the analysis. Your conclusions and policy recommendations can be very detrimental particularly to historically marginalized populations!

- You should also ask yourself: Who cares?

GRE-GPA Example Data

Student	GRE-Total (X_1)	Undergraduate GPA (X_2)	Graduate GPA(Y)
1	145	3.2	4.0
2	120	3.7	3.9
3	125	3.6	3.8
4	130	2.9	3.7
5	110	3.5	3.6
6	100	3.3	3.5
7	95	3.0	3.4
8	115	2.7	3.3
9	105	3.1	3.2
10	90	2.8	3.1
11	105	2.4	3.0

Sample Partial Slope and Intercept

$$b_1 = \frac{(r_{Y1} - r_{Y2}r_{12})s_Y}{(1 - r_{12}^2)s_1} = \frac{[.7845 - (.7516)(.3011)].3317}{(1 - .3011^2)16.3346} = .0125$$

$$b_2 = \frac{(r_{Y2} - r_{Y1}r_{12})s_Y}{(1 - r_{12}^2)s_2} = \frac{[.7516 - (.7845)(.3011)].3317}{(1 - .3011^2).4011} = .4687$$

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 3.5000 - (.0125)(112.7273) - (.4687)(3.1091) = .6337$$

Sample Multiple Linear Regression Model

- $Y_i = b_1 X_{1i} + b_2 X_{2i} + a + e_i$
- $Y_i = .0125\ X_{1i} + .4687\ X_{2i} + .6337 + e_i$
- If your score on the GRETOT was 130 and your UGPA was 3.5, then your predicted score on the GGPA would be computed as:
  - $Y'_i = .0125\ (130) + .4687\ (3.5000) + .6337 = 3.8992$

Overall *F*-Test Statistic

$$F = \frac{R^2/m}{(1 - R^2)/(n - m - 1)} = \frac{.9089/2}{(1 - .9089)/(11 - 2 - 1)} = 39.9078$$

- Or

$$F = \frac{SS_{reg}/df_{reg}}{SS_{res}/df_{res}} = \frac{0.9998/2}{.1002/8} = 39.9122$$

- The critical value, at the .05 level of significance, is ${}_{.05}F_{2,8} = 4.46$
- Test statistic exceeds the critical value, so we reject H_0 and conclude that all of the partial slopes are not equal to zero at the .05 level of significance