Applied Statistical Analysis

EDUC 6050

Week 10

Finding clarity using data

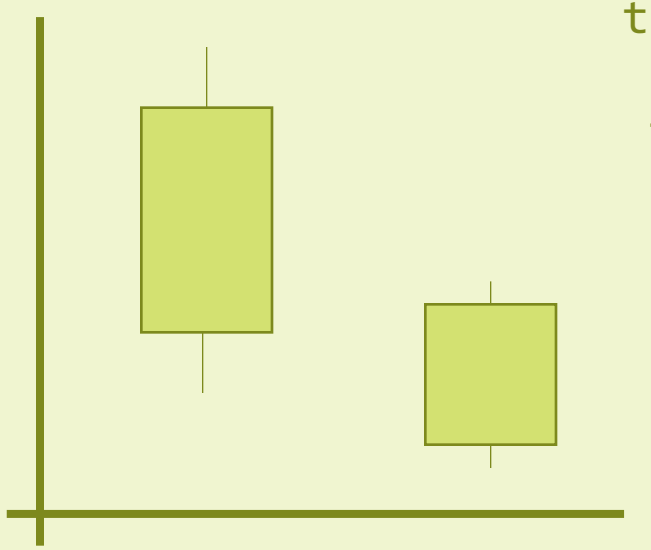
Today

REGRESSION!

Comparing Means

Is one group different than the other(s)?
- Z-tests
- T-tests
- ANOVA

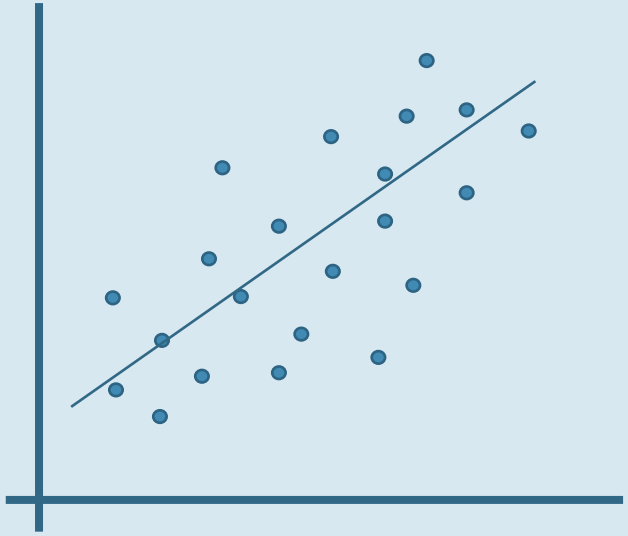
We compare the means and use the variability to decide if the difference is significant



Assessing Relationships

Is there a relationship between the two variables?
- Correlation
- Regression

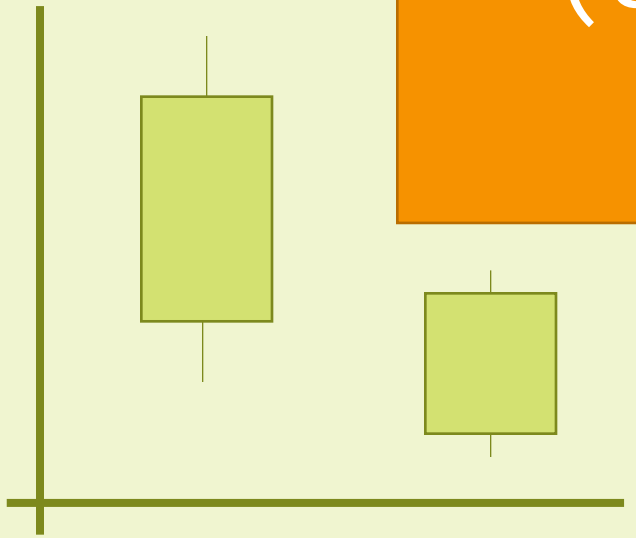
We look at how much the variables “move together”



Comparing Means

Is one group different than the other(s)?

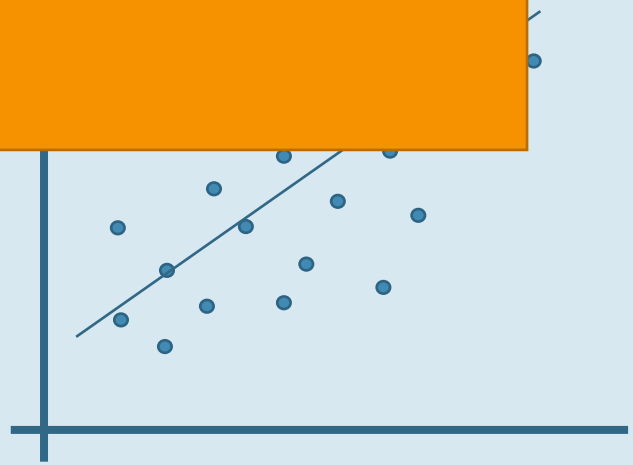
- Z-tests
- T-tests
- ANOVA



Assessing Relationships

Is there a relationship between the two variables?

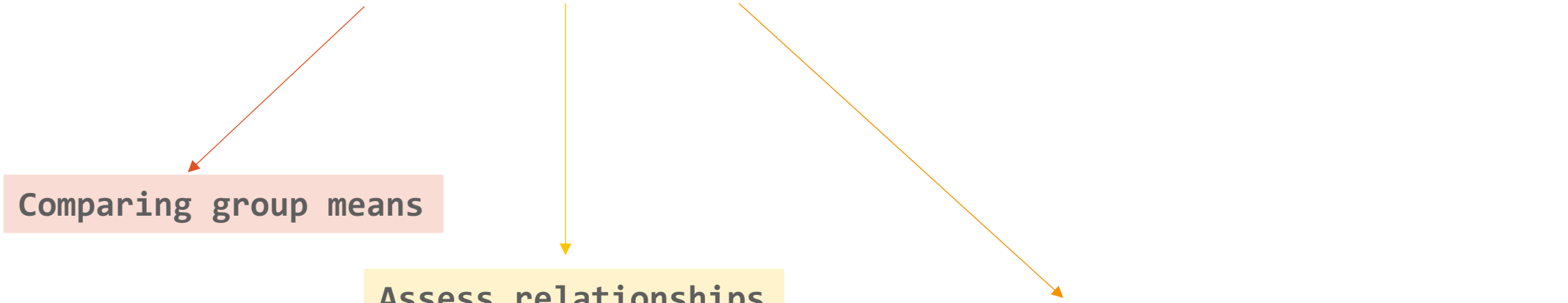
We look at how much the variables “move together”



Regression does both (can be at the same time)

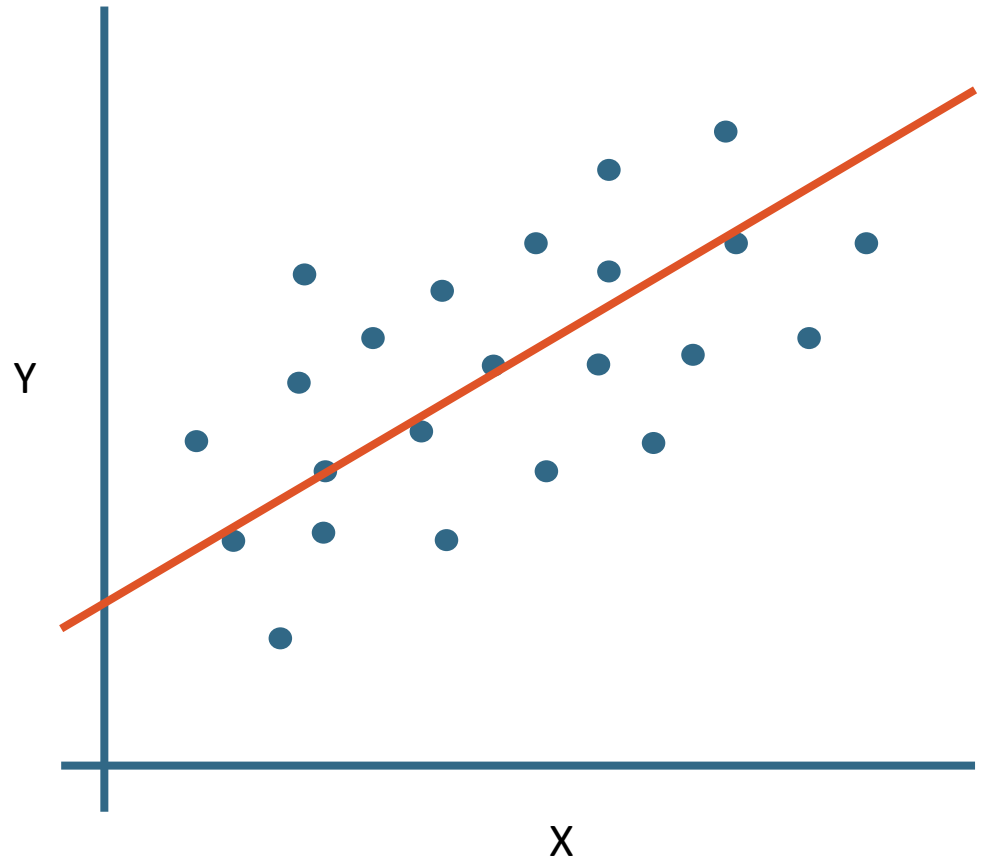
Intro to Regression

The **foundation** of almost everything we do in statistics



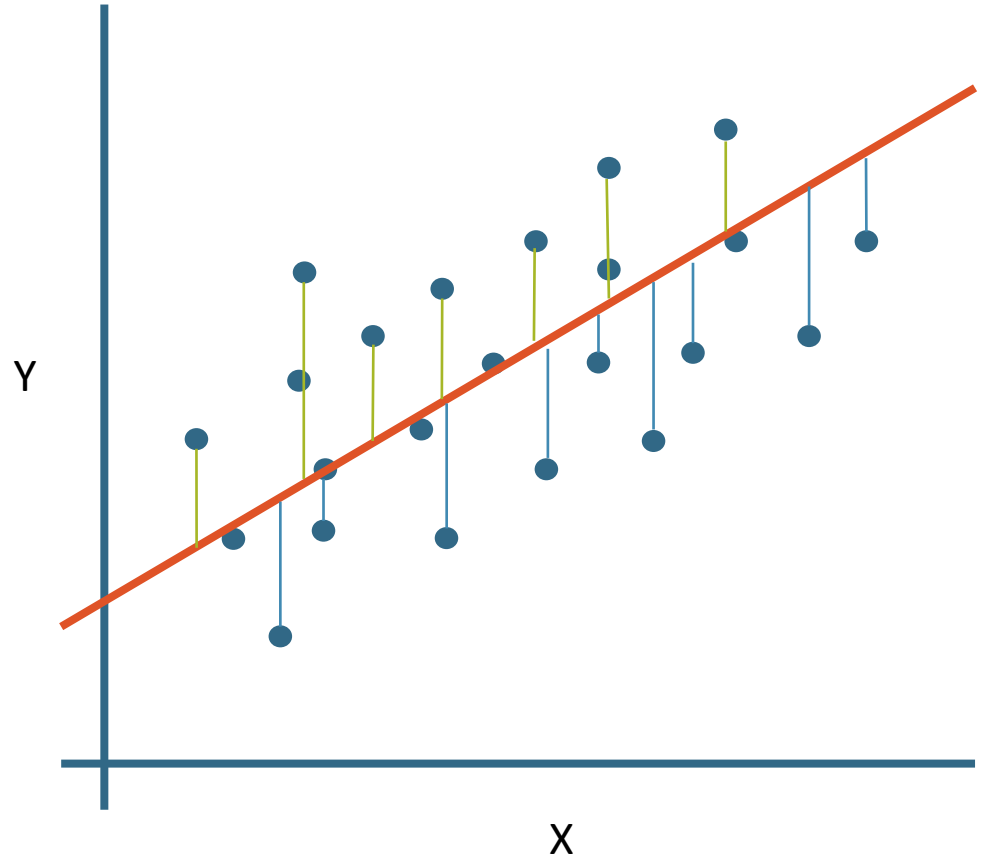
Can handle many types of outcome and predictor data types
Results are interpretable

Logic of Regression



We are trying to find the best fitting line

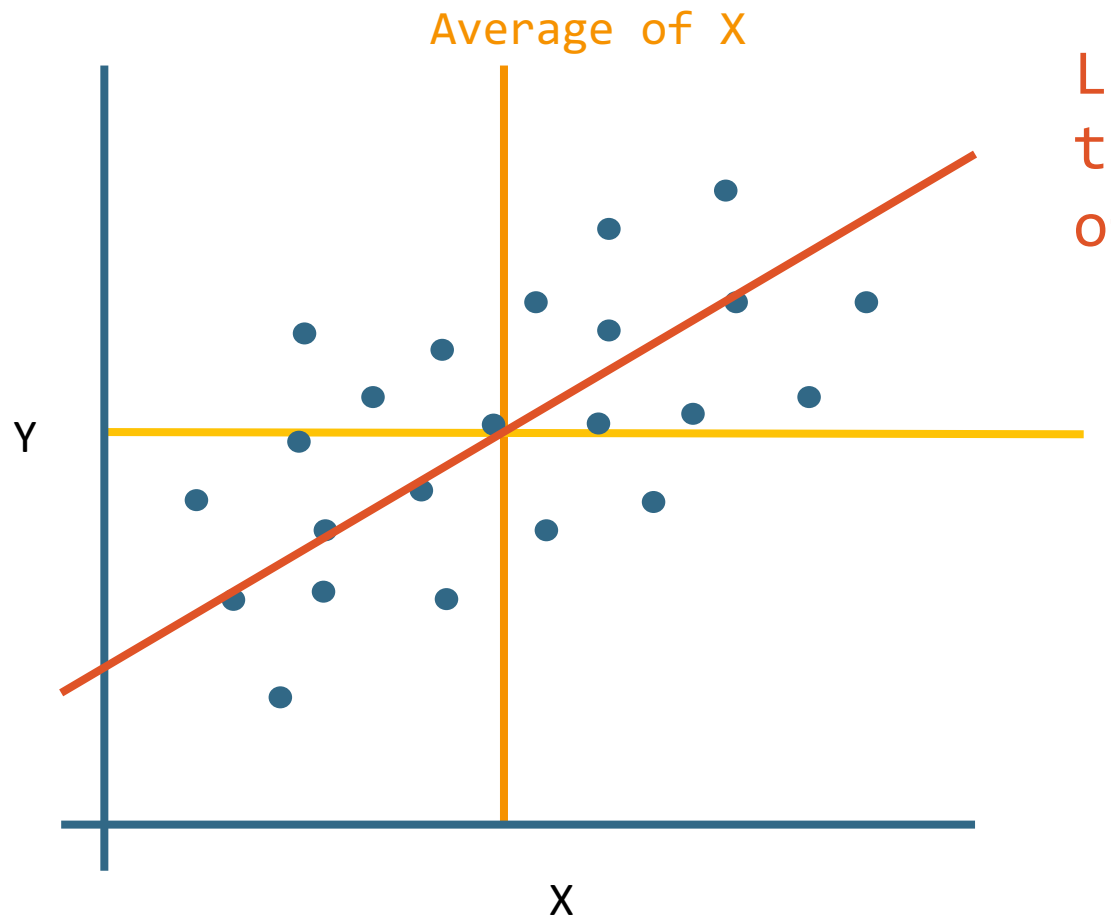
Logic of Regression



We are trying to find the best fitting line

We do this by minimizing the difference between the points and the line (called the residuals)

Logic of Regression



Line always goes through the averages of X and Y

Two Main Types of Regression

Simple

- Only one predictor in the model
- When variables are standardized, gives same results as correlation
- When using a grouping variable, same results as t-test or ANOVA

Multiple

- More than one variable in the model
- When variables are standardized, gives “partial” correlation
- Predictors can be any combination of categorical and continuous

Two Main Types of Regression

Simple

- Only one predictor in the model
- When variables are standardized, gives same results as correlation
- When using a grouping variable, same results as t-test or ANOVA

Multiple

- More than one variable in the model
- When variables are standardized, gives "partial" correlation
- Predictors can be any combination of categorical and continuous

Simple Linear Regression

- Only one predictor in the model
- When variables are standardized, gives same results as correlation
- When using a grouping variable, same results as t-test or ANOVA

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

Simple Linear Regression

- Only one predictor in the model
- When variables are standardized, gives same results as correlation
- When using a grouping variable, same results as t-test or ANOVA

slope

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

intercept

Simple Linear Regression

- Only one predictor in the model
- When variables are standardized, gives same results as correlation
- When using a grouping variable, same results as t-test or ANOVA

slope

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

intercept

unexplained stuff in Y

Simple Linear Regression

- Only one predictor in the model
- When variables are standardized, gives same results as correlation
- When using a grouping variable, same results as t-test or ANOVA

Example

We have two variables, X and Y, the predictor and outcome. We want to know if increases/decreases in X are associated (or predict) changes in Y.

Simple Linear Regression

- Only one predictor in the model
- When variables are standardized, gives same results as correlation
- When using a grouping variable, same results as t-test or ANOVA

Example

X	Y
3	9
2	7
4	8
4	6
5	9

Regression vs. Correlation

- Very related
- In simple regression, when variables are standardized, they are the same thing
  - (just with directionality in regression)
- Jamovi provides both standardized and non-standardized results

Quick Note: Models

- Models are just simplifications of the world that help us describe it
- “All models are wrong, but some models are useful.” - George E.P. Box (1979)
- A model is useful when it represents reality and is concise enough to understand and act on it

General Requirements

1. Two or more variables,
2. Outcome needs to be continuous
3. Others can be continuous or categorical

ID	X	Y
1	8	7
2	6	2
3	9	6
4	7	6
5	7	8
6	8	5
7	5	3
8	5	5

Hypothesis Testing with Simple Regression

The same 6 step approach!

1. Examine Variables to Assess Statistical Assumptions
2. State the Null and Research Hypotheses (symbolically and verbally)
3. Define Critical Regions
4. Compute the Test Statistic
5. Compute an Effect Size and Describe it
6. Interpreting the results

1 Examine Variables to Assess Statistical Assumptions

Basic Assumptions

1. Independence of data
2. Appropriate measurement of variables for the analysis
3. Normality of distributions
4. Homoscedastic

1 Examine Variables to Assess Statistical Assumptions

Basic Assumptions

1. **Independence** of data
2. Appropriat
   for the al
3. Normality
4. Homoscedastic

Individuals are independent of each other (one person's scores does not affect another's)

1 Examine Variables to Assess Statistical Assumptions

Basic Assumptions

1. Independence of data
2. **Appropriate measurement** of variables for the analysis
3. Normality of distributions
4. Homogeneity

Here we need interval/ratio outcome

1 Examine Variables to Assess Statistical Assumptions

Basic Assumptions

1. Independe
2. Appropria
   for the analysis
3. Normality of distributions
4. Homoscedastic

Residuals should be normally distributed

1 Examine Variables to Assess Statistical Assumptions

Basic Assumptions

1. Independence of data
2. Appropri
   for
3. Normalit
4. Homoscedastic

Variance around the line should be roughly equal across the whole line

1 Examine Variables to Assess Statistical Assumptions

Basic Assumptions

1. Independence of data
2. Appropriate measurement of variables for the analysis
3. Normality of distributions
4. Homoscedastic
5. **Linear Relationships**
6. **No omitted variables**

1 Examine Variables to Assess Statistical Assumptions

Basic Assumptions

1. Independence of data
2. Appropri… for the …
3. Nor…
4. Hom…scedastic
5. **Linear Relationships**
6. No omitted variables

Relationships between the outcome and the continuous predictors should be linear

1 Examine Variables to Assess Statistical Assumptions

Basic Assumptions

1. Independence of data
2. Appropriate measurement of variables for the
3. Normalit
4. Hom
5. Lin ar R
6. No omitted variables

Any variable that is related to both the predictor and the outcome should be included in the regression model

1 Examine Variables to Assess Statistical Assumptions

Examining the Basic Assumptions

1. **Independence:** random sample
2. **Appropriate measurement:** know what your variables are
3. **Normality**: Histograms, **Q-Q**, skew and kurtosis
4. **Homoscedastic:** Scatterplots
5. **Linear**: Scatterplots
6. **No Omitted**: check correlations, know the theory

2 State the Null and Research Hypotheses (symbolically and verbally)

Hypothesis Type	Symbolic	Verbal	Difference between means created by:
Research Hypothesis	$\beta \neq 0$	X predicts Y	True relationship
Null Hypothesis	$\beta = 0$	There is no *real* relationship.	Random chance (sampling error)

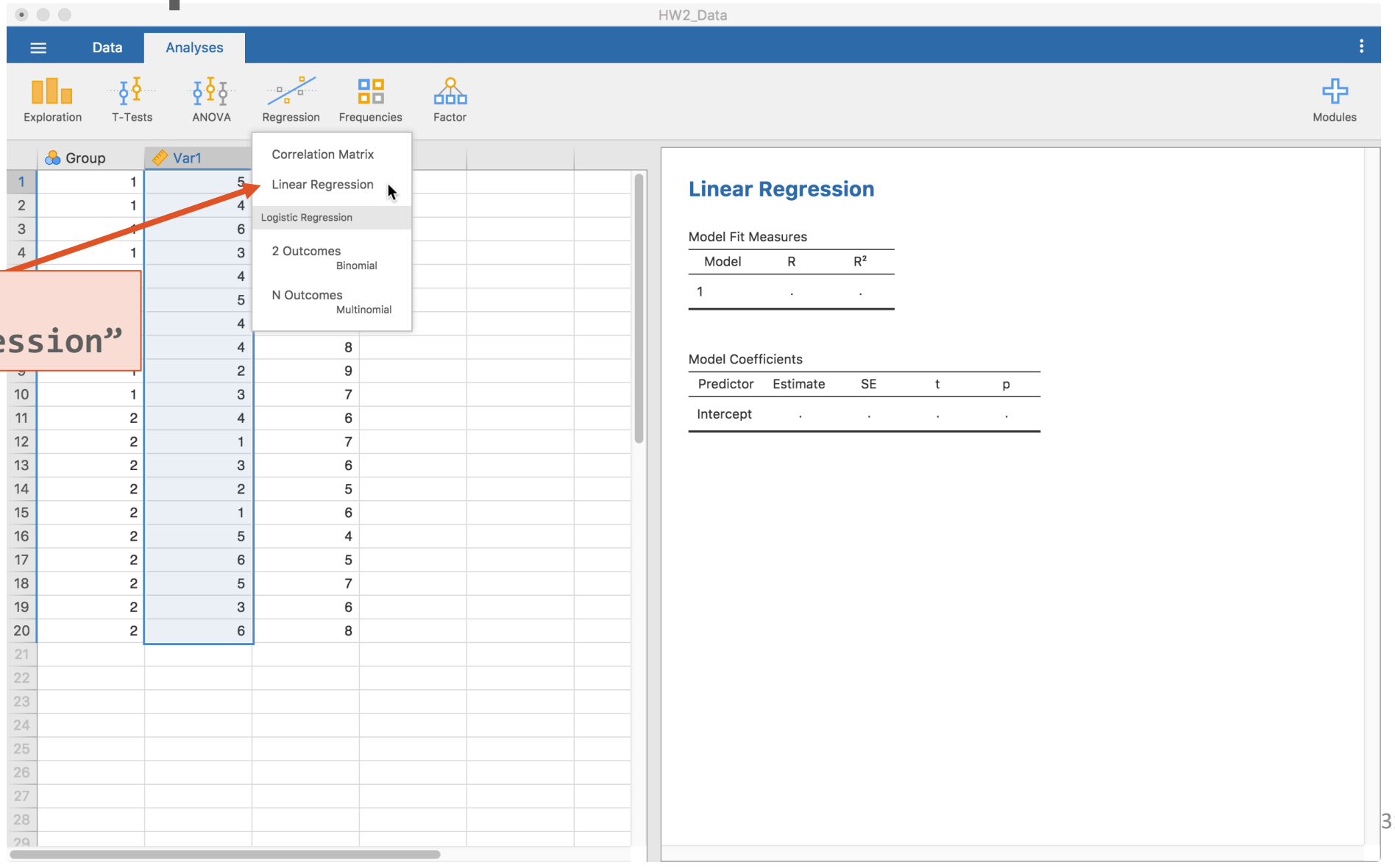
3 Define Critical Regions

How much evidence is enough to believe the null is not true?

- generally based on an alpha = .05

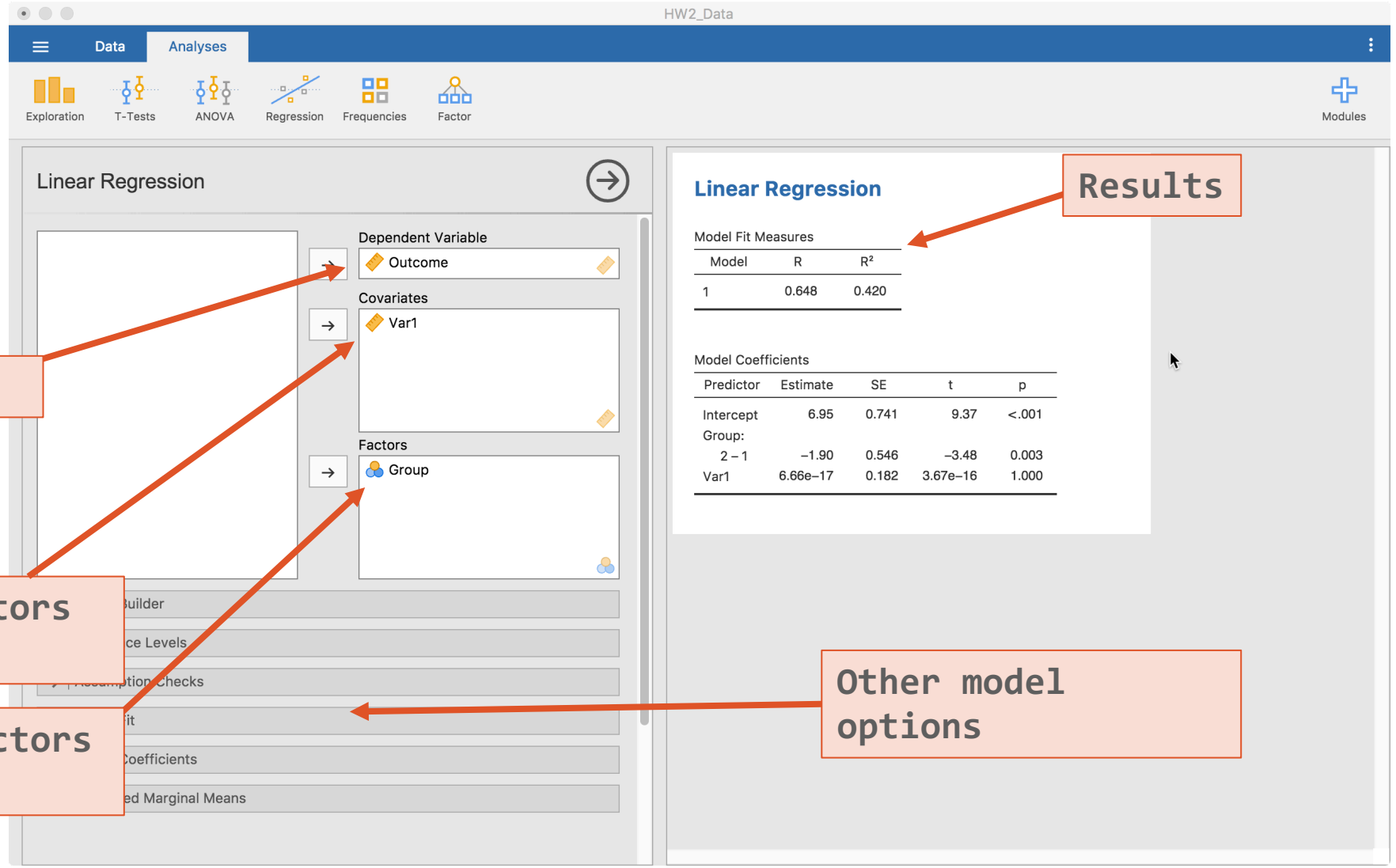
Use software's p-value to judge if it is below .05

4 Compute the Test Statistic



Click on "Linear Regression"

4 Compute the Test Statistic



4 Compute the Test Statistic

Intercept = What Y is when X is zero

$$\text{Slope} = \frac{\text{Covariation of X and Y}}{\text{Variation of X}}$$

4 Compute the Test Statistic

Intercept = What Y is when X is zero

$$\text{Slope} = \frac{\text{Covariation of X and Y}}{\text{Variation of X}}$$

The way the variables move together (just like in correlation)

4 Compute the Test Statistic

Intercept = What Y is when X is zero

Slope = The change in Y for a one unit change in X, on average.

5 Compute an Effect Size and Describe it

One of the main effect sizes for regression is R^2

$$R^2 = \frac{\text{Variation in Y we can explain}}{\text{Total Variation in Y}}$$

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

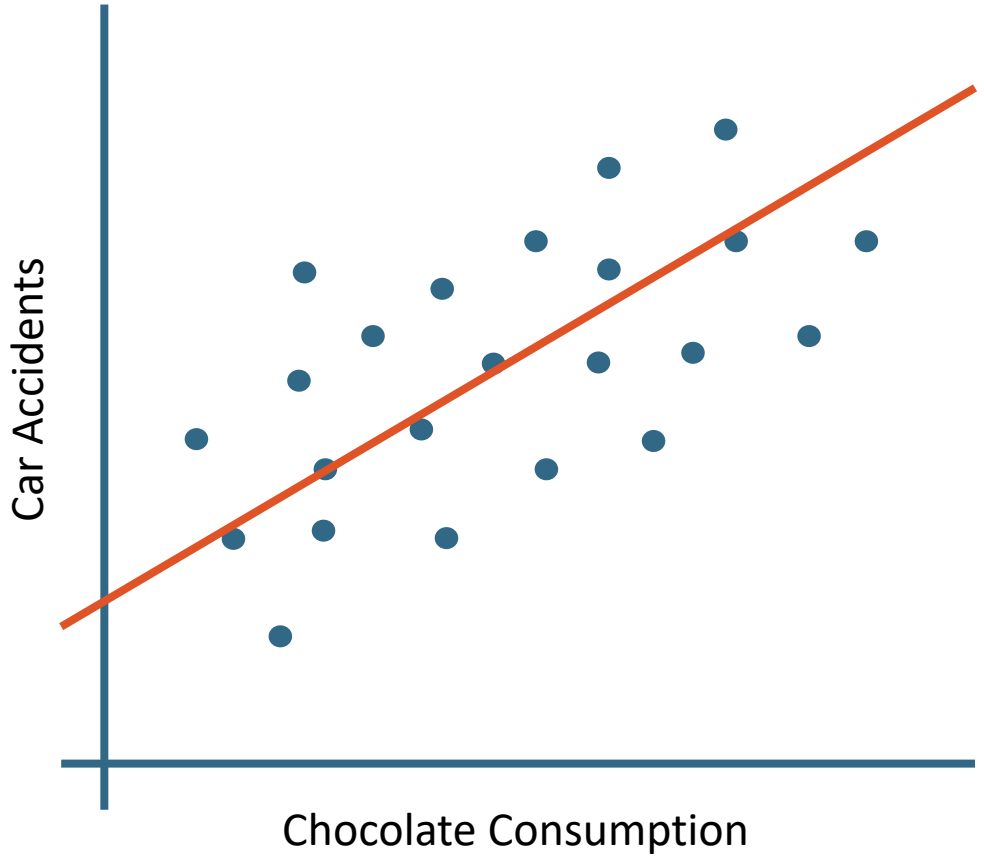
6 Interpreting the results

Put your results into words

The regression analysis showed that X significantly predicts Y (b = .5, p = .02). X accounted for 32% of the variation in Y.

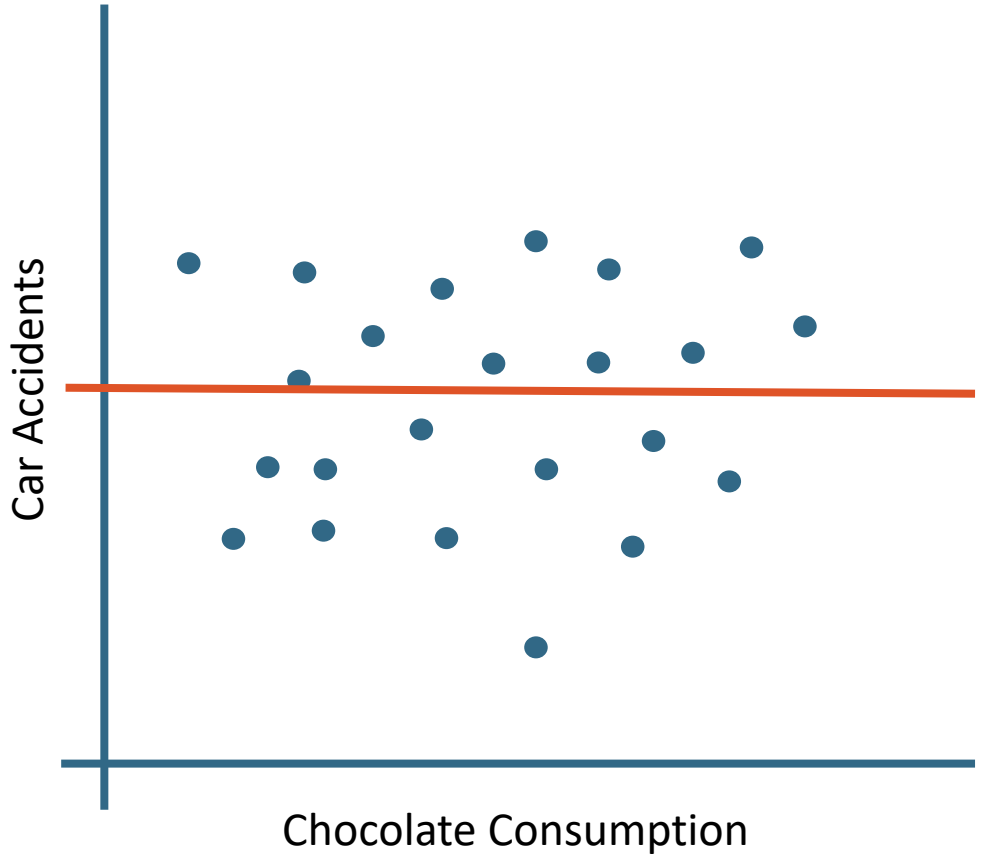
Multiple Regression

Example of Simple Regression



Chocolate consumption looks like it might cause car accidents. Is this accurate? What else could explain it?

What if we control for time of year?



There is no longer a relationship when we "take out" the part of the relationship that is related to time of the year

The two models

Simple Relationship

Model Fit Measures

Model	R	R^2
1	0.623	0.389

Model Coefficients - Car Accidents

Predictor	Estimate	SE	t	p
Intercept	2.316	0.877	2.64	0.014
Chocolate Consumption	0.643	0.158	4.07	<.001

Relationship *Controlling for Time of Year*

Model Fit Measures

Model	R	R^2
1	0.749	0.561

Model Coefficients - Car Accidents

Predictor	Estimate	SE	t	p
Intercept	3.149	0.803	3.922	<.001
Chocolate Consumption	0.185	0.200	0.922	0.365
Time of Year	3.291	1.051	3.132	0.004

The two models

Simple Relationship

Model Fit Measures

Model	R	R^2
1	0.623	0.389

Model Coefficients - Car Accidents

Predictor	Estimate	SE	t	p
Intercept	2.816	0.877	2.64	0.014
Chocolate Consumption	0.643	0.158	4.07	<.001

Relationship *Controlling for Time of Year*

Model Fit Measures

Model	R	R^2
1	0.749	0.561

Model Coefficients - Car Accidents

Predictor	Estimate	SE	t	p
Intercept	3.149	0.803	3.922	<.001
Chocolate Consumption	0.185	0.200	0.922	0.365
Time of Year	3.291	1.051	3.132	0.004

Two Main Types of Regression

Simple

- Only one predictor in the model
- When variables are standardized, gives same results as correlation
- When using a grouping variable, same results as t-test or ANOVA

Multiple

- More than one variable in the model
- When variables are standardized, gives “partial” correlation
- Predictors can be any combination of categorical and continuous

Multiple Regression

More than one predictor in the same model

This change the interpretation just a little:

- Slope is now the change in Y for a one-unit change in X, **while holding the other predictors constant.**

Multiple Regression

More than one predictor in the same model

This change the interpretation just a little

Also changes what we are estimating:

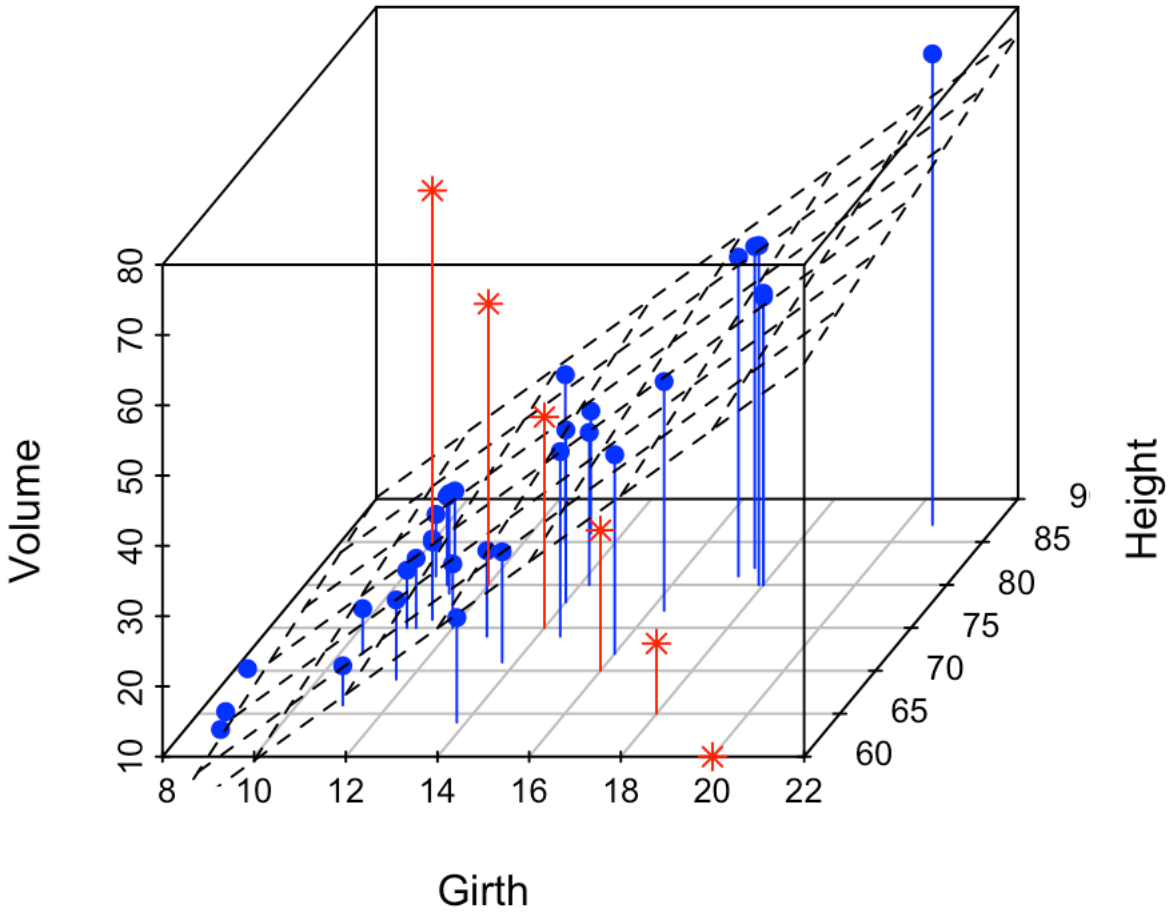
Multiple Regression

More than one predictor in the same model

This change the interp little

Also changes what we are estimating:

A plane instead of a line



Multiple Regression

Provides us with a few more things to think about

1. **Variable Selection**
2. **Assumption Checks**
3. **Multi-collinearity**
4. **Interactions**

Variable Selection When Theory Is Unclear

Several Approaches

1. Forward
2. Backward
3. Lasso
4. Covariates then predictor of interest

I'd recommend these two

Assumption Checks

Linearity and Homoskedasticity more difficult since it is now in 3+ dimensions

Jamovi makes these fairly straightforward

Multi-Collinearity

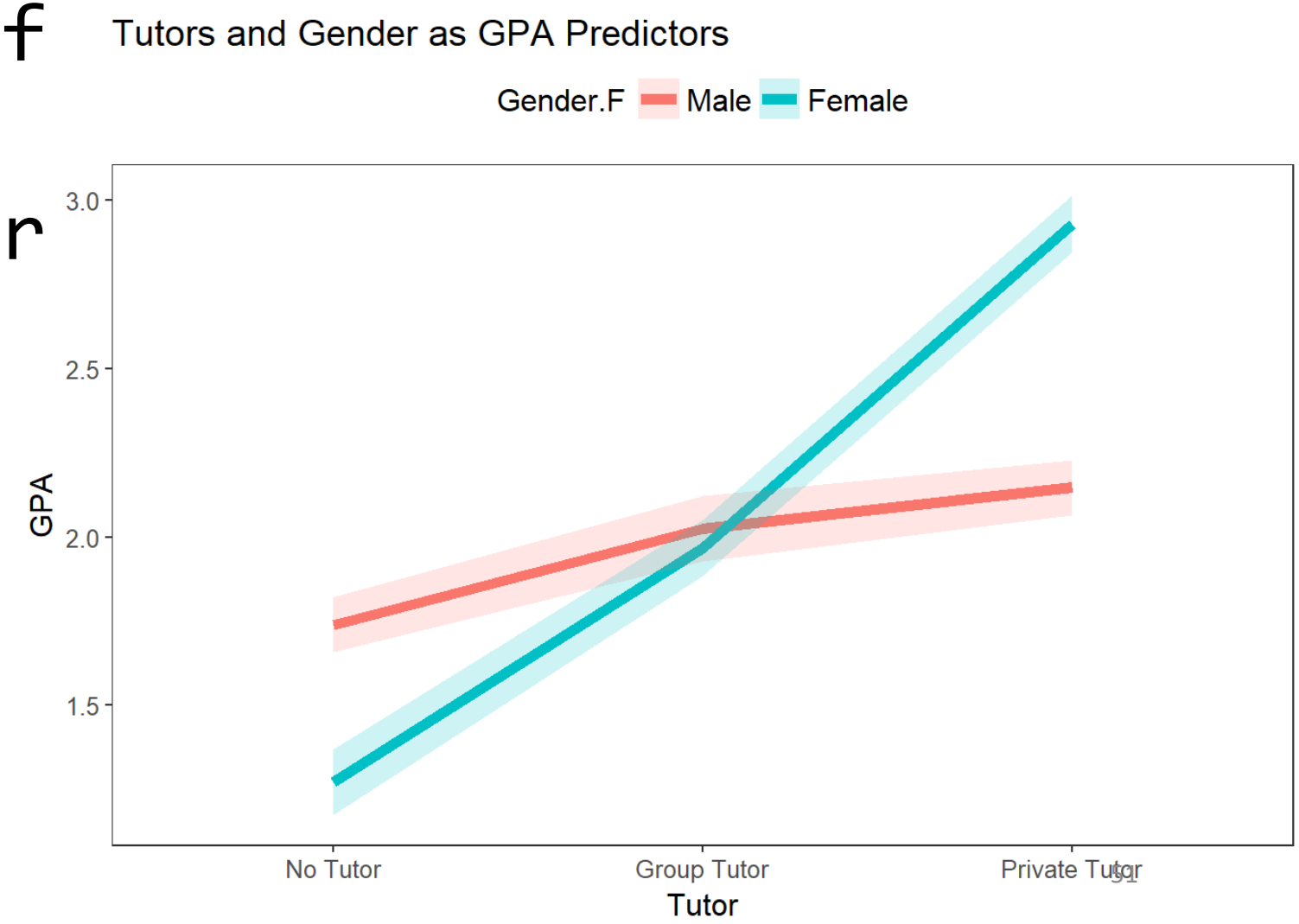
When two or more predictors are very related to each other or are linear combinations of each other

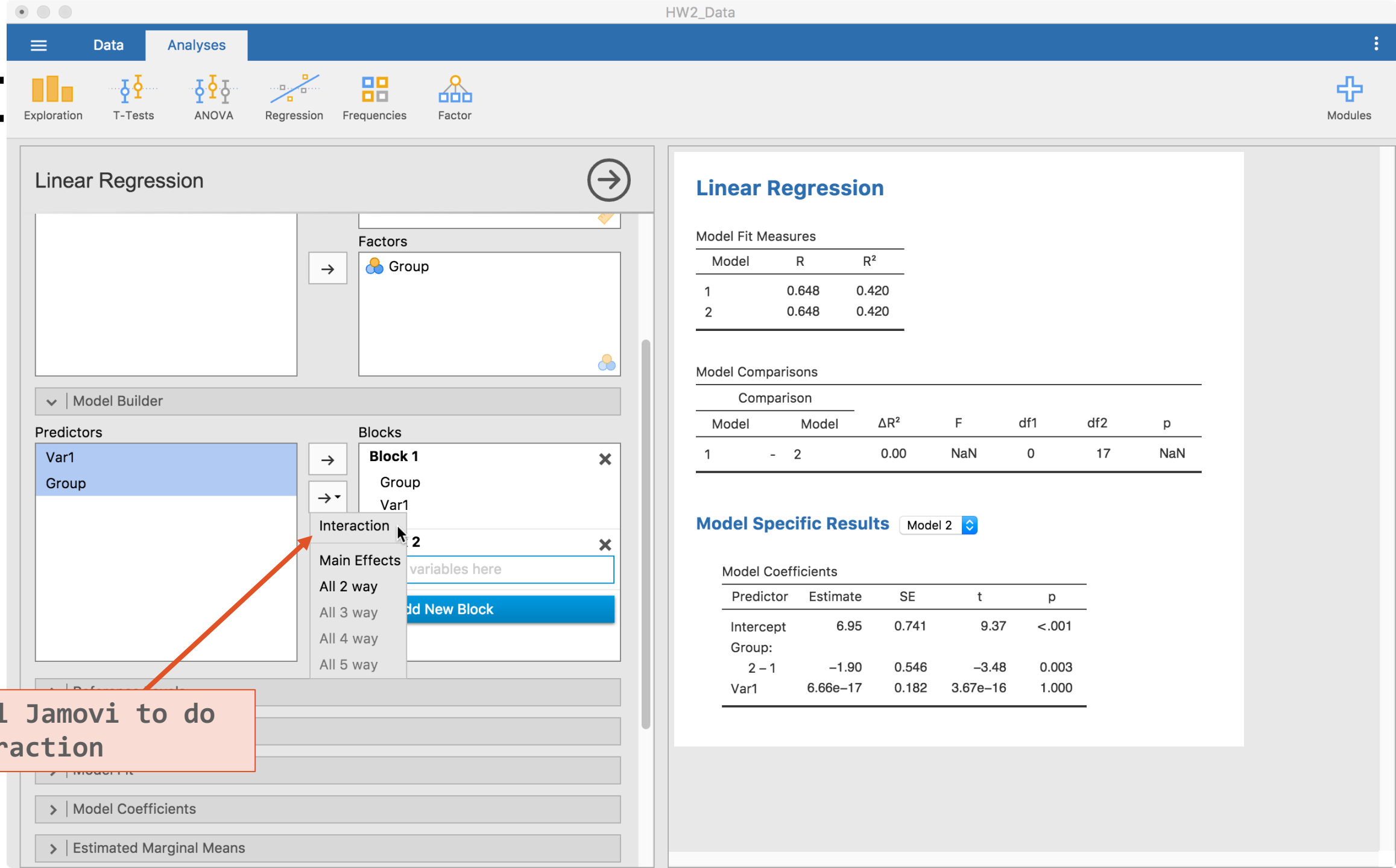
Check correlations
Dummy codes are correct (Jamovi does this automatically)

Interactions

When the effect of a predictor depends on another

Can have 2+ variables in the interaction




HW2_Data
Data
Analyses
Exploration
T-Tests
ANOVA
Regression
Frequencies
Factor
Modules
Linear Regression
Factors
Group
Model Builder
Predictors
Var1
Group
Blocks
Block 1
Group
Var1
Interaction
Main Effects
All 2 way
All 3 way
All 4 way
All 5 way
variables here
d New Block
Model Coefficients
Estimated Marginal Means
Linear Regression
Model Fit Measures
Model R R²
1 0.648 0.420
2 0.648 0.420
Model Comparisons
Comparison
Model Model ΔR² F df1 df2 p
1 - 2 0.00 NaN 0 17 NaN
Model Specific Results
Model 2
Model Coefficients
Predictor Estimate SE t p
Intercept 6.95 0.741 9.37 <.001
Group:
2 – 1 –1.90 0.546 –3.48 0.003
Var1 6.66e–17 0.182 3.67e–16 1.000
Can tell Jamovi to do an interaction

Challenge

For the following situations, describe what approach you would take and why:

You have data on life satisfaction and age and want to the know the relationship between them. They are both continuous.

Challenge

For the following situations, describe what approach you would take and why:

You have data on life satisfaction and age and want to the know the relationship between them. You believe that age causes an increase in life satisfaction. They are both continuous.

Challenge

For the following situations, describe what approach you would take and why:

You have data on life satisfaction and age and believe that the relationship between them depends on a third variable – social class. Social class is categorical while the others are continuous.

Challenge

For the following situations, describe what approach you would take and why:

You have multiple waves of data wherein the participants have received an intervention between times 1 and 2. There are a total of 3 time points.

Challenge

For the following situations, describe what approach you would take and why:

You have a binary outcome and you think that the continuous variable “var1” predicts which category of the outcome the individual belongs to.

In-class discussion slides

Application

Example Using
The Office/Parks and Rec Data Set

Hypothesis Test with Regression