Applied Statistical Analysis

EDUC 6050

Week 12

Finding clarity using data

Today

Logistic Regression

Intro to Logistic Regression

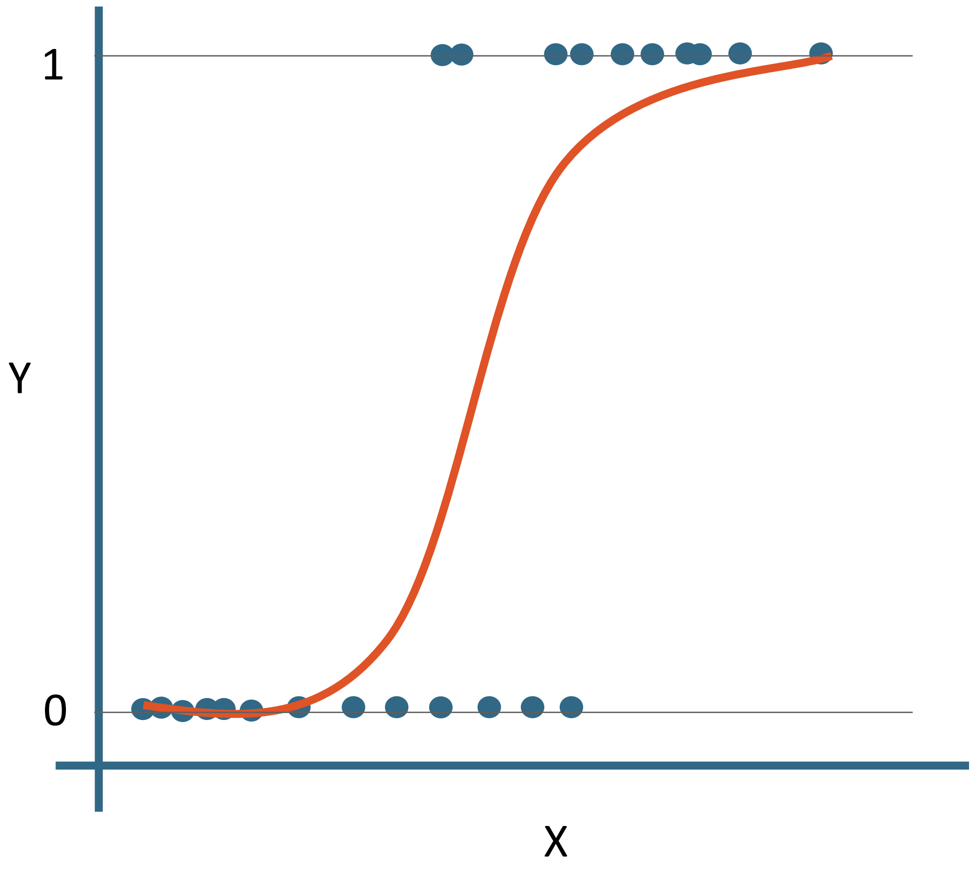
So far, we have always wanted continuous outcome variables

But what if our outcome is a categorical variable??

Logistic Regression is just like linear regression but works with binary (dichotomous) outcomes

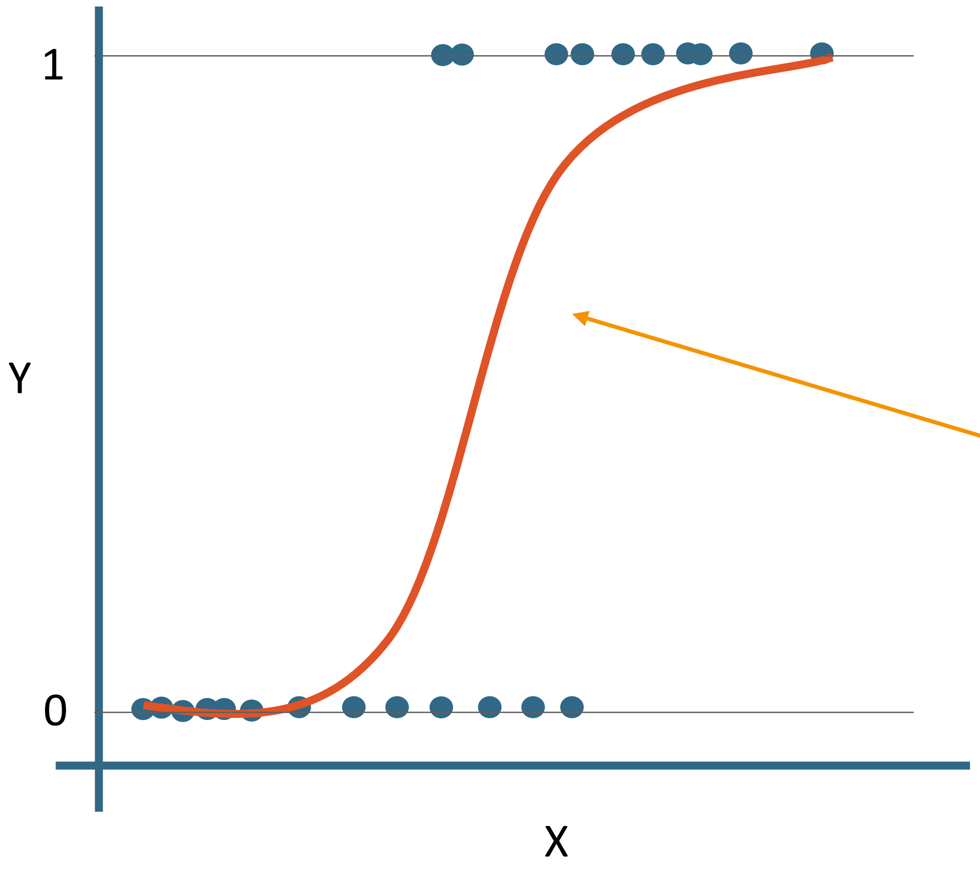
- Substance Use or Not
- Cancer or Not
- Buy it or Not

Logic of Logistic Regression



We are trying to find the best fitting S curve

Logic of Logistic Regression



We are trying to find the best fitting S curve

The curve is the model estimated probability of Y = 1

Logistic Regression

Simple

- Only one predictor in the model
- Tells you if that one predictor is associated with the odds of Y = 1

Multiple

- More than one variable in the model
- Tells you if, while holding the other variables constant, if that predictor is associated with the odds of Y = 1

Logistic Regression

- Logistic does what regression does but with a little bit of mathematical magic

$$\boxed{logit(Y)} = \beta_0 + \beta_1 X + \epsilon$$

Logistic Regression

- Logistic does what regression does but with a little bit of mathematical magic

slope

$$\boxed{logit(Y)} = \beta_0 + \beta_1 X + \epsilon$$

intercept

Logistic Regression

- Logistic does what regression does but with a little bit of mathematical magic

slope

$$\boxed{logit(Y)} = \beta_0 + \beta_1 X + \epsilon$$

intercept

unexplained stuff in the odds of Y

Logistic Regression

$$logit(Y) = \beta_0 + \beta_1 X + \epsilon$$

Example

We have two variables, X and Y. X is continuous, Y is binary. We want to know if increases/decreases in X are associated (or predict) changes in the chance of Y equaling 1.

Logistic Regression

- It is trying to predict the outcome accurately using the information from the predictor
- Better prediction tells us that the predictor(s) is/are more strongly related to the outcome

General Requirements

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

ID	X	Y
1	8	0
2	6	1
3	9	1
4	7	1
5	7	0
6	8	0
7	5	1
8	5	0

Hypothesis Testing with Logistic 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 a
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. Homo[illegible]

Here we need nominal 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. Nor alit
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. Logistic Relationship
6. No omitted variables

1 Examine Variables to Assess Statistical Assumptions

Basic Assumptions

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

The "S-shaped" curve should fit to the data

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. Logistic
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. Logistic: 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)

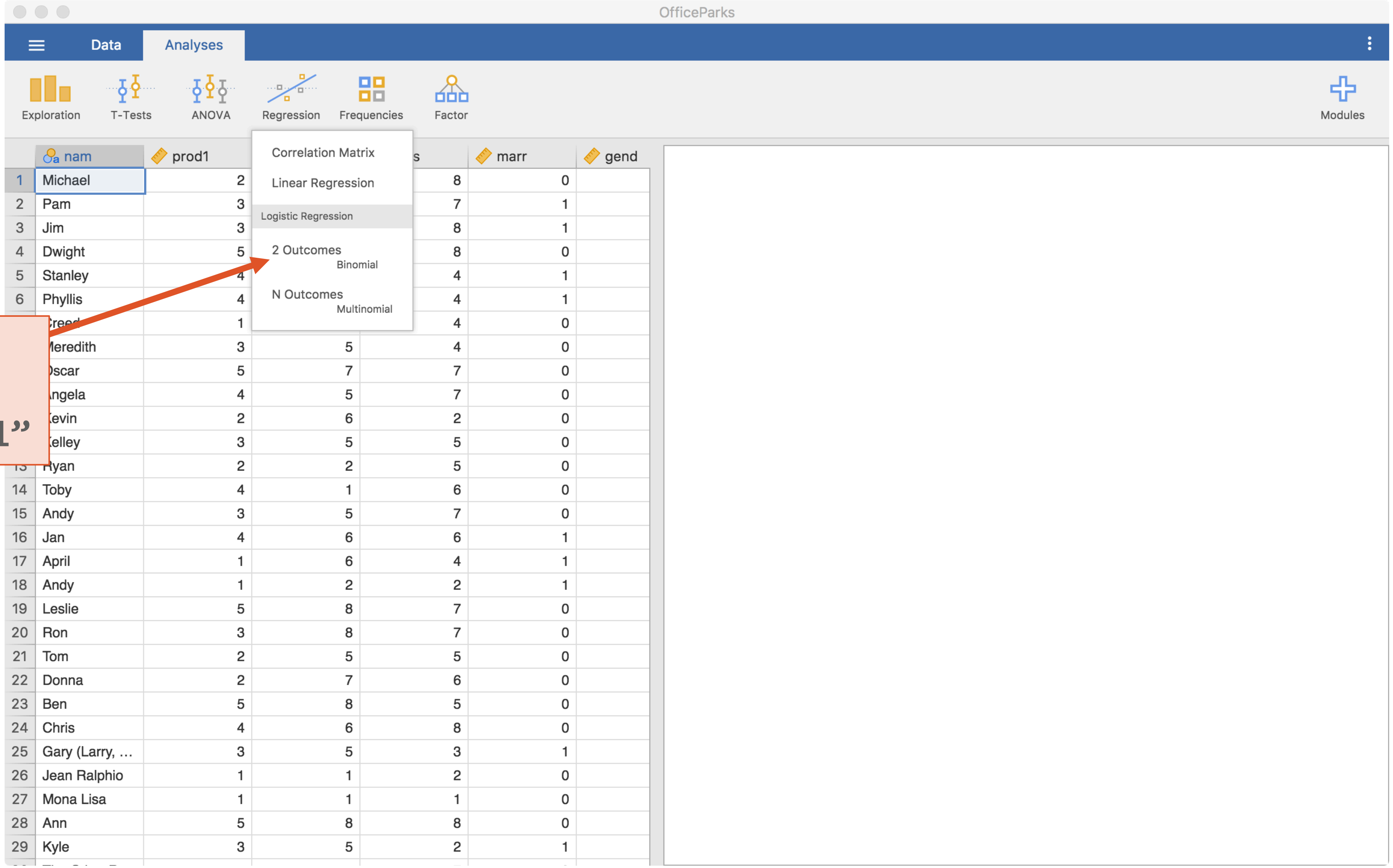
3 Define Critical Regions

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

generally based on an alpha = .05

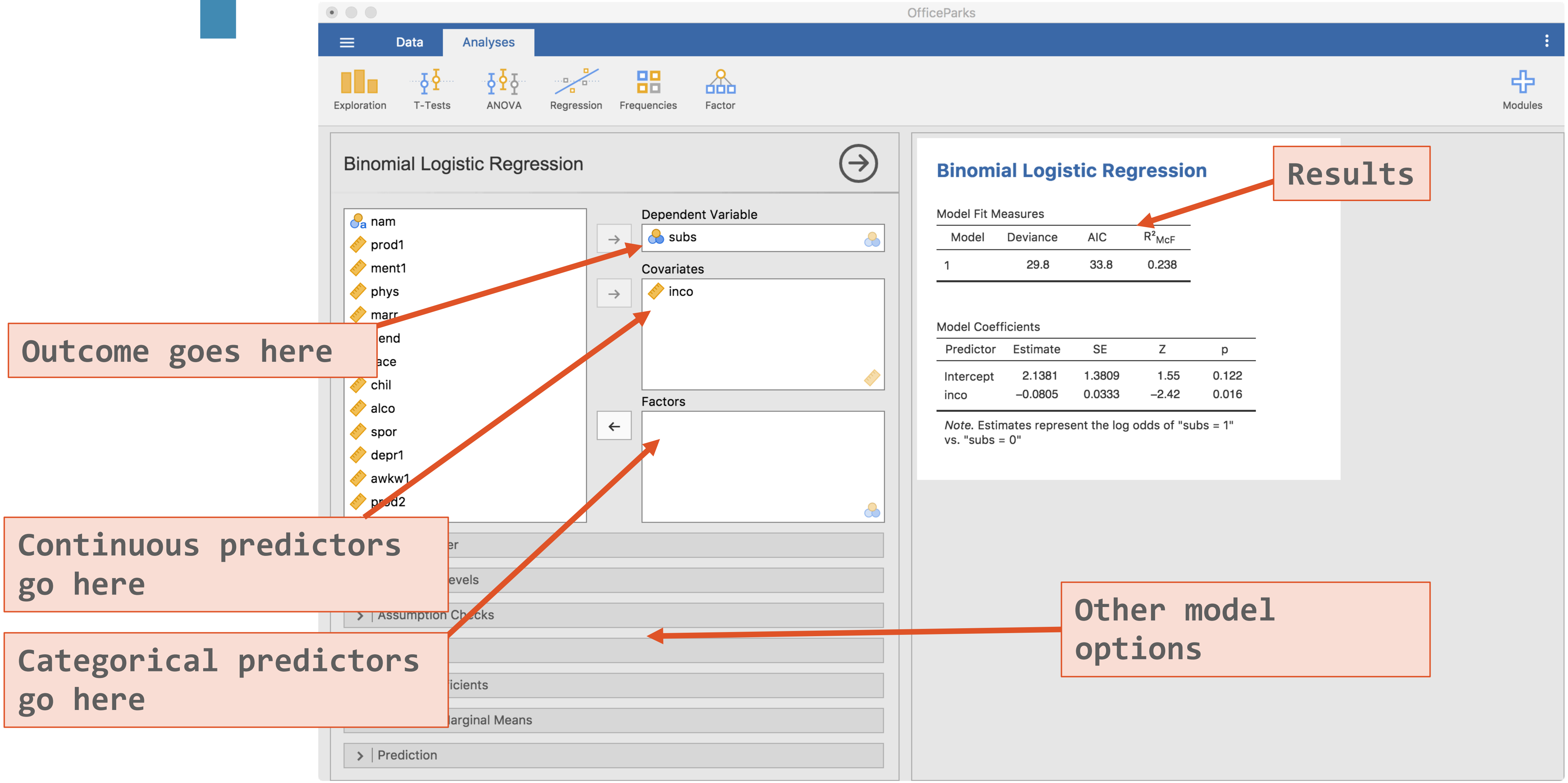
Use software’s p-value to judge if it is below .05

4 Compute the Test Statistic



Click on
"2 Outcomes
Binomial"

4 Compute the Test Statistic



4 Continuous Predictor

Model Coefficients

Predictor	Estimate	SE	Z	p	Odds ratio	95% Confidence Interval	
						Lower	Upper
Intercept	2.1381	1.3809	1.55	0.122	8.483	0.566	127.060
Income	$\boxed{-0.0805}$	0.0333	-2.42	$\boxed{0.016}$	$\boxed{0.923}$	0.864	0.985

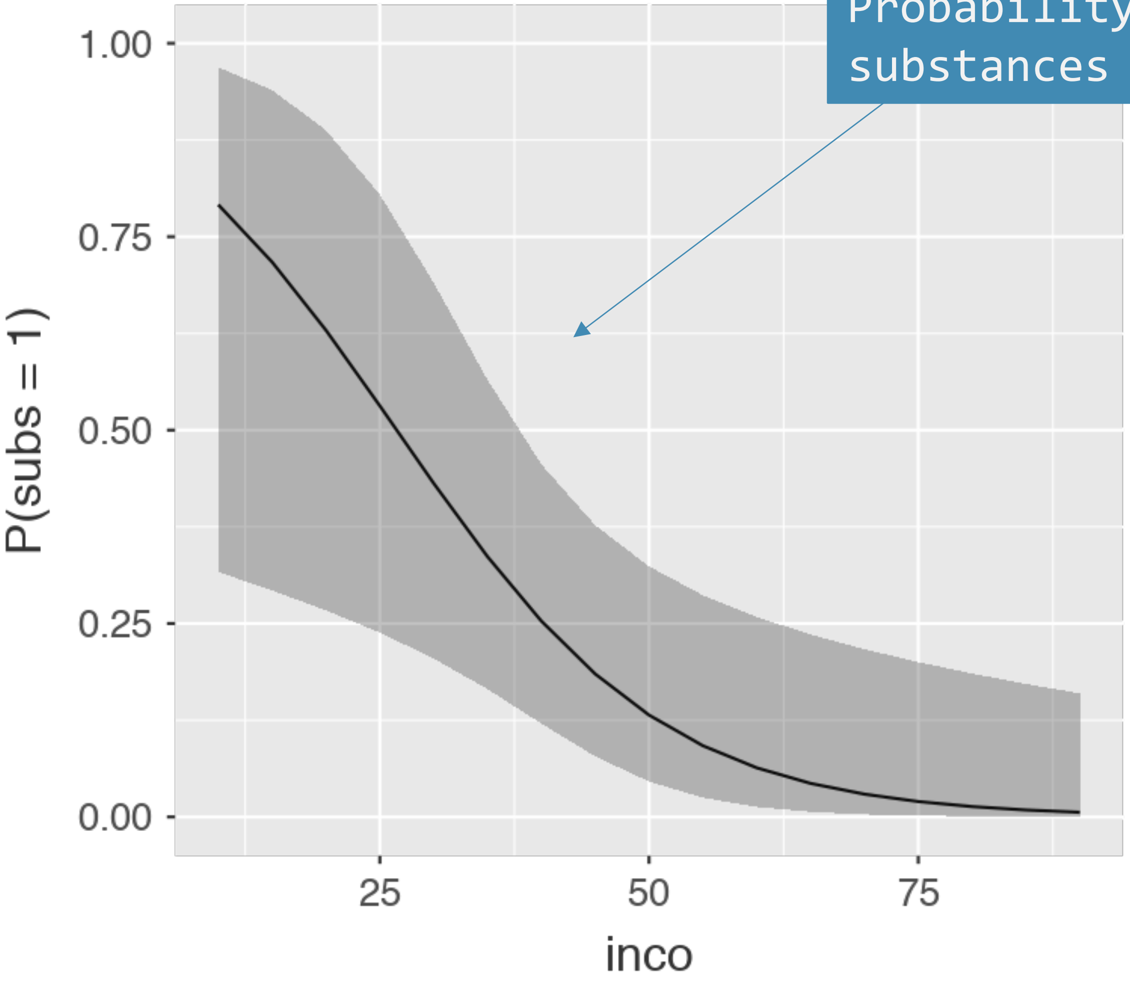
Note. Estimates represent the log odds of "subs = 1" vs. "subs = 0"

Estimate in “log-odds” units

Significant

The odds ratio is below 1 so as income increases, the odds of using substances decreases by ~1 - .923 = .077 (7.7% decrease)

4 Continuous Predictor



Probability of using substances by income level

How well can we predict substance use with just income?

Classification Table – subs

	Predicted		
Observed	0	1	% Correct
0	29	1	96.7
1	5	3	37.5

Note. The cut-off value is set to 0.5

4 Categorical Predictor

Model Coefficients

						95% Confidence Interval	
Predictor	Estimate	SE	Z	p	Odds ratio	Lower	Upper
Intercept	-1.504	0.553	-2.721	0.007	0.222	0.0752	0.657
Show:							
The Office – Parks and Rec	0.405	0.799	0.507	0.612	1.500	0.3131	7.186

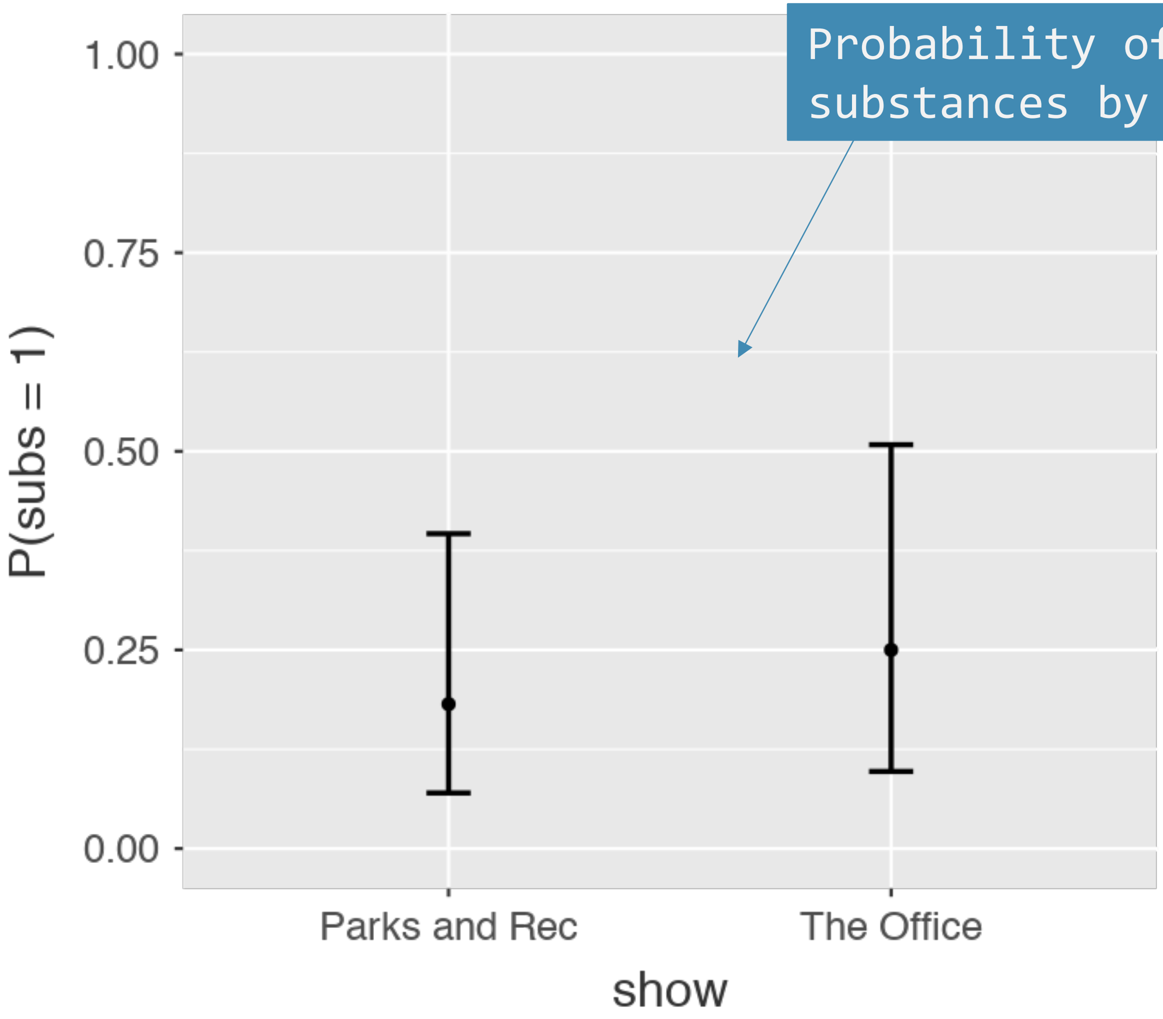
Note. Estimates represent the log odds of "subs = 1" vs. "subs = 0"

Estimate in "log-odds" units

Not Significant

The odds ratio is above 1 so individuals on The Office have an odds of using substances 50% (1.5 – 1 = .5 = 50%) higher than PR

4 Categorical Predictor



Probability of using substances by show

How well can we predict substance use with just income?

Classification Table – subs

	Predicted		
Observed	0	1	% Correct
0	30	0	100
1	8	0	0.00

Note. The cut-off value is set to 0.5

4 Compute the Test Statistic

Intercept = Odds of Y when X is zero

Slope = The change in the odds of Y = 1 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

$$Odds\ Ratio = \frac{Odds\ of\ Y\ when\ X\ is\ one\ unit\ higher}{\text{Odds of Y when X is } not\ one\ unit\ higher}$$

OR	Estimated Size of the Effect
Close to .01	Small
Close to .09	Moderate
Close to .25	Large

6 Interpreting the results

The logistic regression analysis showed that income significantly predicted the odds of substance use (OR = -.923, p = .016). As income increased by $1000, the odds of using substances decreased by 7.7%.

Multiple Logistic Regression

Multiple Logistic Regression

More than one predictor in the same model

This change the interpretation just a little:

- Slope is now the change in the odds of Y = 1 for a one unit change in X, **while holding the other predictors constant.**

Multiple Regression

Provides us with a few more things to think about

1. **Variable Selection**
2. **Assumption Checks (much more difficult in logistic regression)**
3. **Multi-collinearity**
4. **Interactions**

Variable Selection

Several Approaches

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

Variable Selection when theory isn't clear

Several Approaches

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

I'd recommend these two

Assumption Checks

Difficult (we won’t cover it in this class)

Jamovi doesn’t provide many checks (only collinearity)

Multi-Collinearity

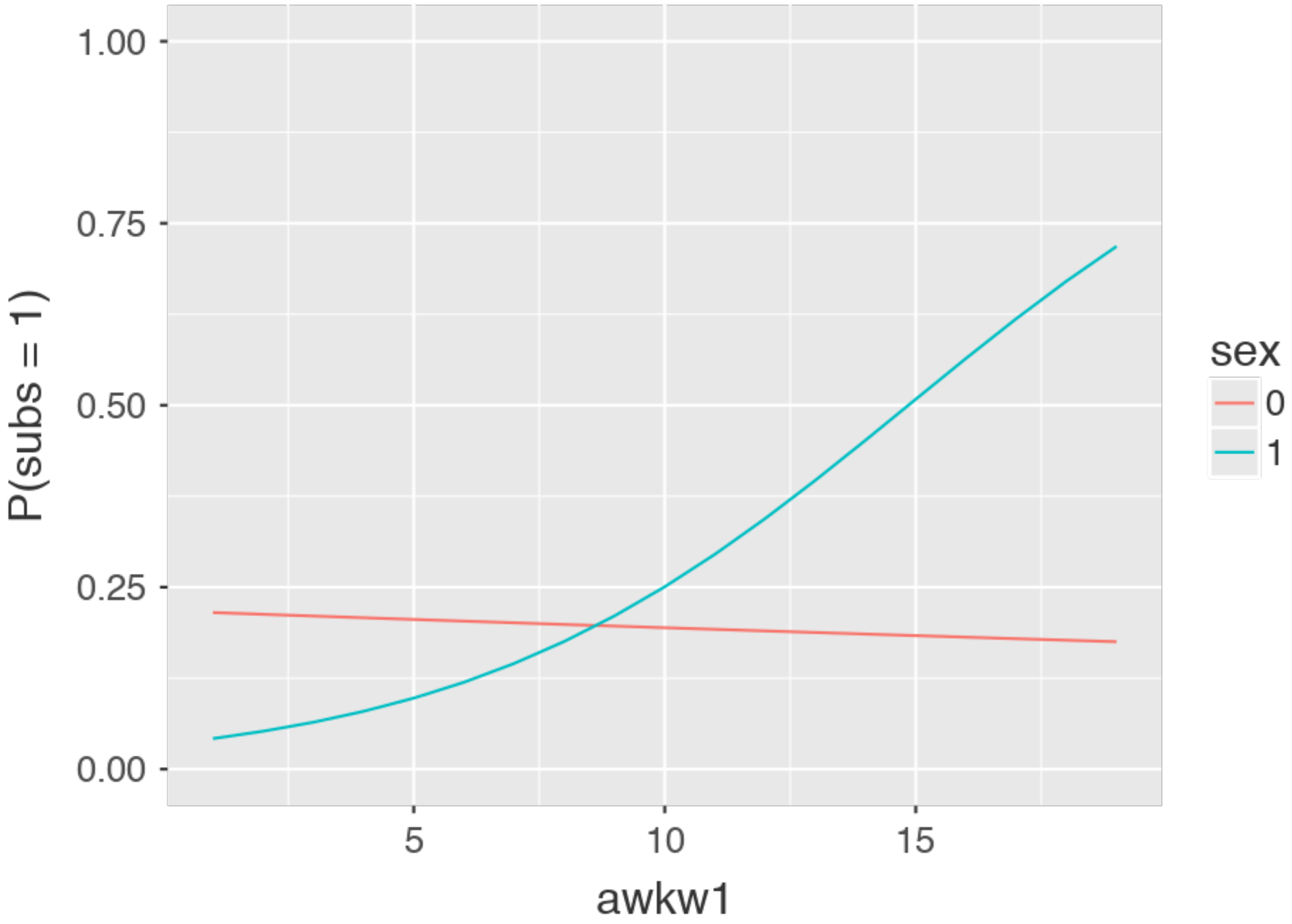
When two or more predictors are very related to each other or are linear combinations of each other

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

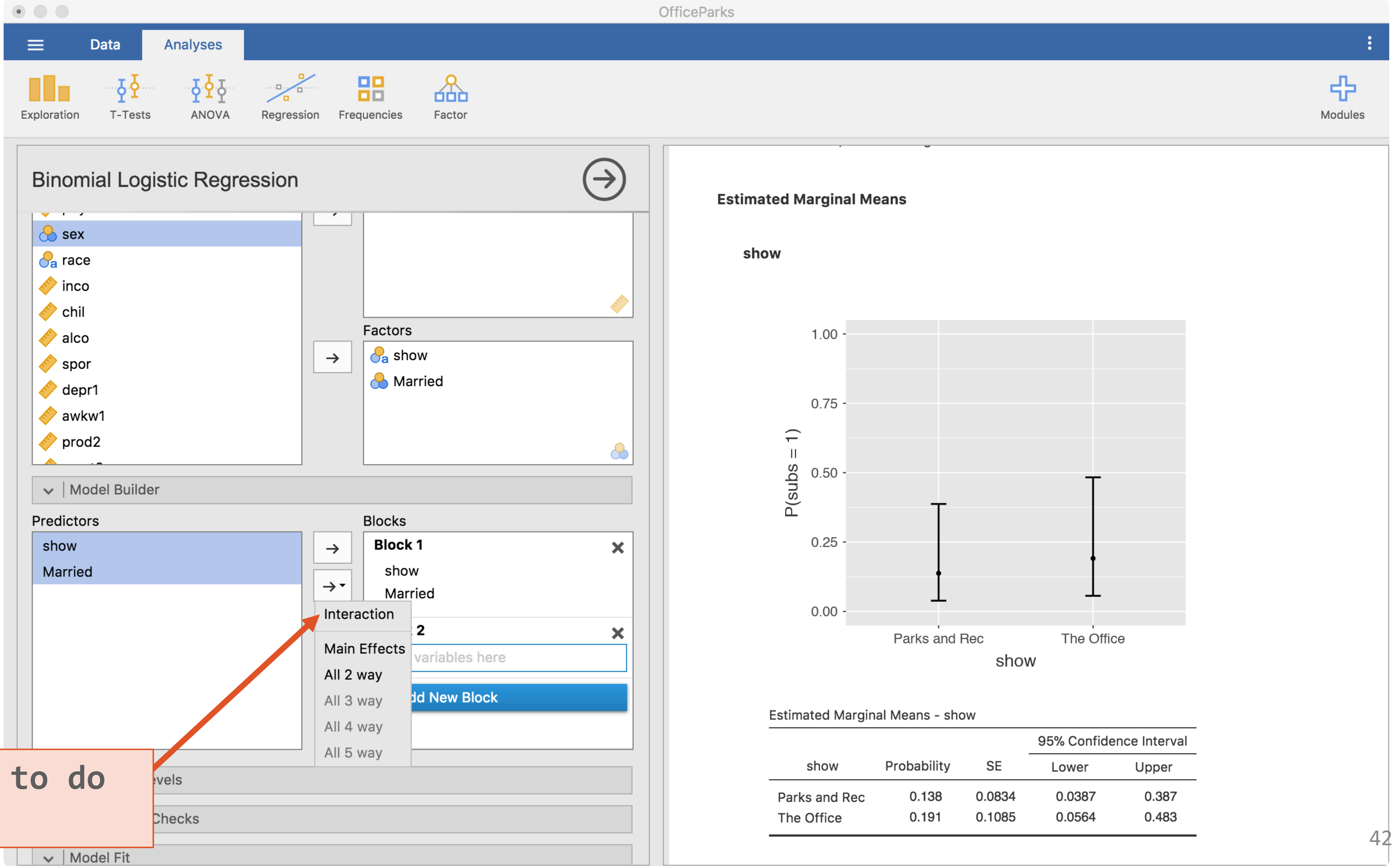
Interactions

Just as we do in linear models

Can have 2+ variables in the interaction



Interactions



Can tell Jamovi to do an interaction

Mediation vs. Moderation

Mediation vs. Moderation

Mediation

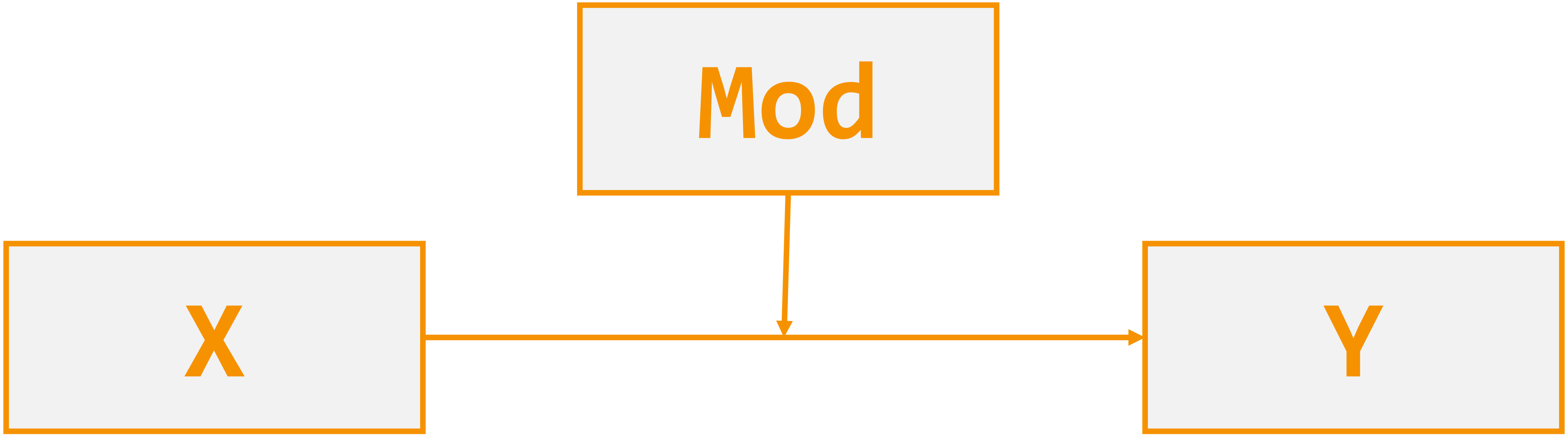
- Tells us about the path of an effect (from one variable to an intermediate one then to the outcome)
- Jamovi performs this
- "The effect of X goes through M to Y"

Moderation

- Synonym for interaction
- Jamovi performs this
- "The effect of X on Y depends on M"

Mediation

Moderation



Mediation

Use specialized software for this (Jamovi can’t do it yet)

Be aware of it but you do not need to know how to do it for this class

Questions?

Please post them to the discussion board before class starts

End of Pre-Recorded Lecture Slides

In-class discussion slides

Application

Example Using
The Office/Parks and Rec Data Set

Hypothesis Test with
Logistic Regression