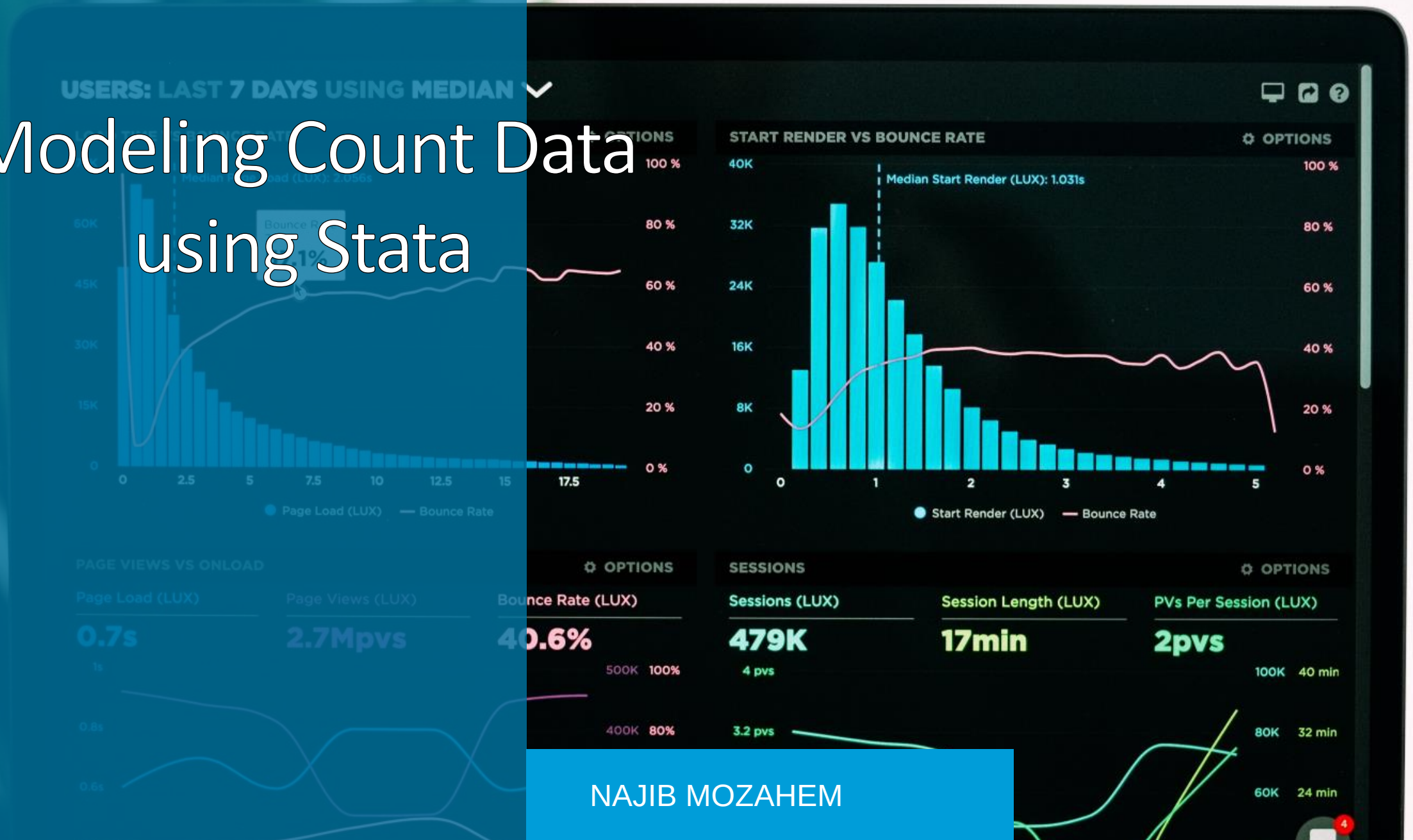
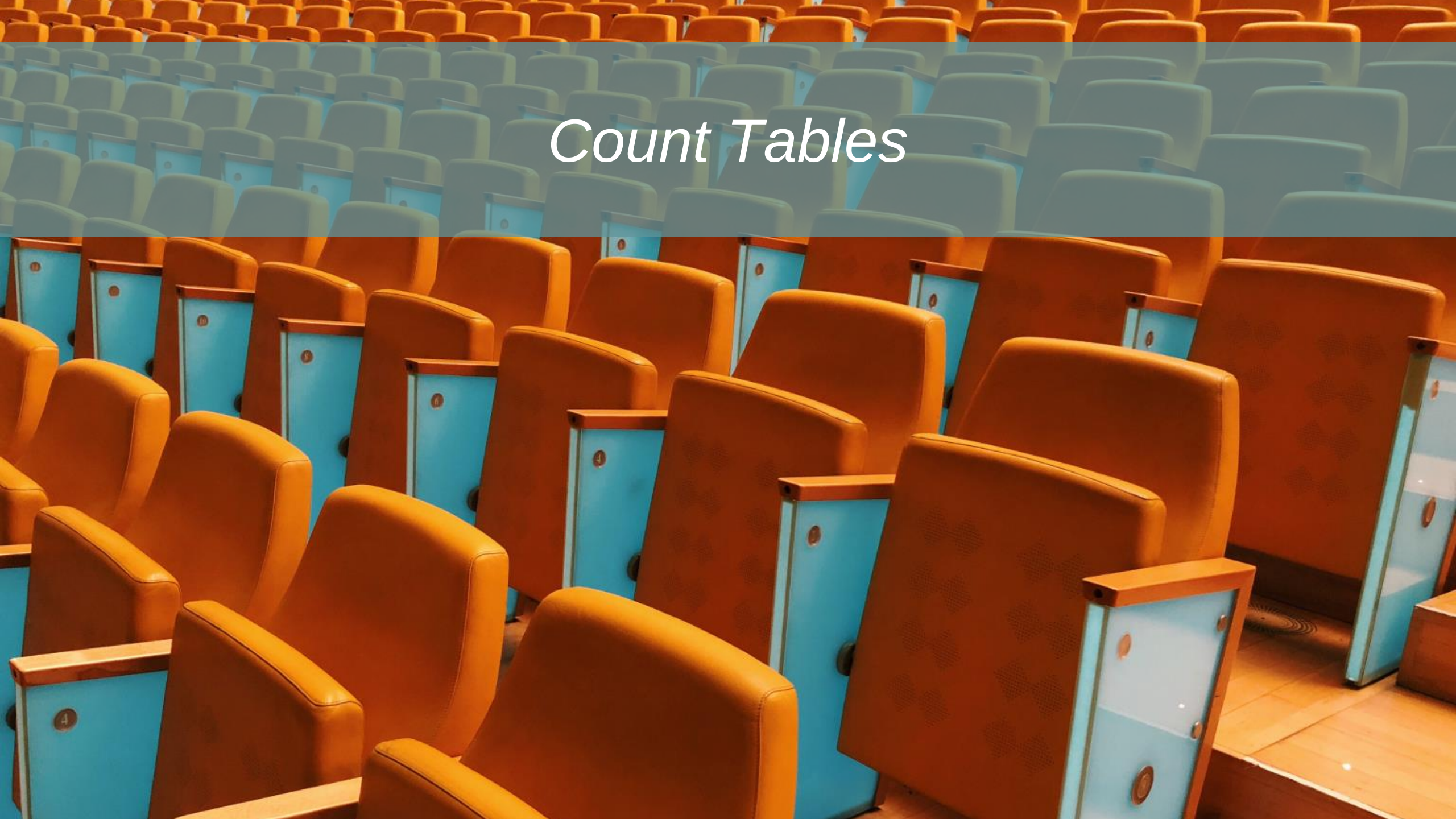
Modeling Count Data using Stata



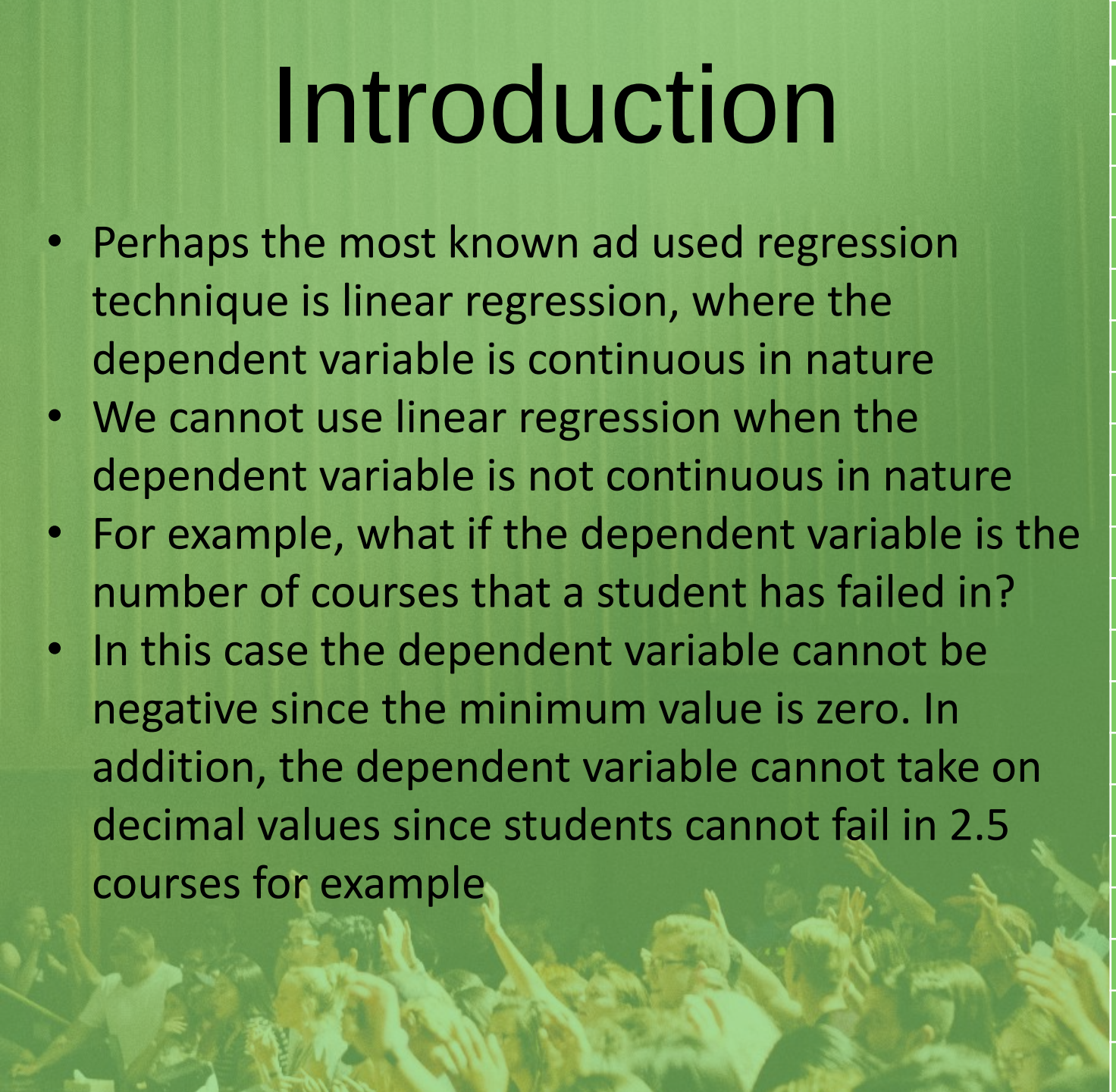
NAJIB MOZAHEM

Count Tables



Introduction

- Perhaps the most known ad used regression technique is linear regression, where the dependent variable is continuous in nature
- We cannot use linear regression when the dependent variable is not continuous in nature
- For example, what if the dependent variable is the number of courses that a student has failed in?
- In this case the dependent variable cannot be negative since the minimum value is zero. In addition, the dependent variable cannot take on decimal values since students cannot fail in 2.5 courses for example



Number of failed courses	Number of courses passed	College
3	20	Business
0	9	Business
5	25	Business
7	21	Business
1	26	Engineering
2	15	Business
2	13	Business
0	18	Business
4	23	Engineering
2	8	Engineering
5	23	Engineering
3	14	Engineering
8	19	Business
0	24	Business
1	5	Engineering
2	13	Engineering
3	17	Business
4	21	Engineering
5	27	Business
9	25	Engineering

Two-by-Two Tables

	College		
Failed	Business	Engineering	Total
No	208	158	366
Yes	35	31	66
Total	243	189	432

- This table sums up the results
- We see that out of a total of 243 courses taken by business students, there are 35 failing grades and 208 passing grade. For engineers, out of a total of 189 courses, there are 31 failing grades and 158 passing grades. Therefore, the count of failed courses for engineers is actually smaller than the count for business (31 for engineers and 35 for business).

Number of failed courses	Number of courses passed	College
3	20	Business
0	9	Business
5	25	Business
7	21	Business
1	26	Engineering
2	15	Business
2	13	Business
0	18	Business
4	23	Engineering
2	8	Engineering
5	23	Engineering
3	14	Engineering
8	19	Business
0	24	Business
1	5	Engineering
2	13	Engineering
3	17	Business
4	21	Engineering
5	27	Business
9	25	Engineering

Risk

- Calculate the risk that a student in business would fail a course and to compare it to the risk that a student in engineering would fail a course
- Risk here indicates the probability of the event happening
- For business students for example, out of a total of 243 courses, 35 resulted in a failed grade. This means that the probability of failure, or the risk of failure, is 35/243 = 0.144.
- For an engineering student, the risk is 31/189 = 0.164.

Incidence-rate Ratio

- To directly compare the two risks we can calculate the incidence-rate ratio, which is simply the ratio of the two numbers:

$$Risk\ ratio = \frac{0.144}{0.164} = 0.878$$

- This means that the likelihood of a business student failing a course is 0.878 times the likelihood of an engineering student failing the course

Two-by-three Tables

	College			
Failed	Business	Engineering	Life Sciences	Total
No	208	158	182	548
Yes	35	31	20	86
Total	243	189	202	634

- The above logic is maintained even when we have more than two groups
- We already know that the risk of failure for business and engineering students are 0.144 and 0.164 respectively.
- The risk for students in the life sciences school is 20/202 = 0.099, which is smaller than the other two risks

Two-by-three Tables (Incidence-Rate Ratio)

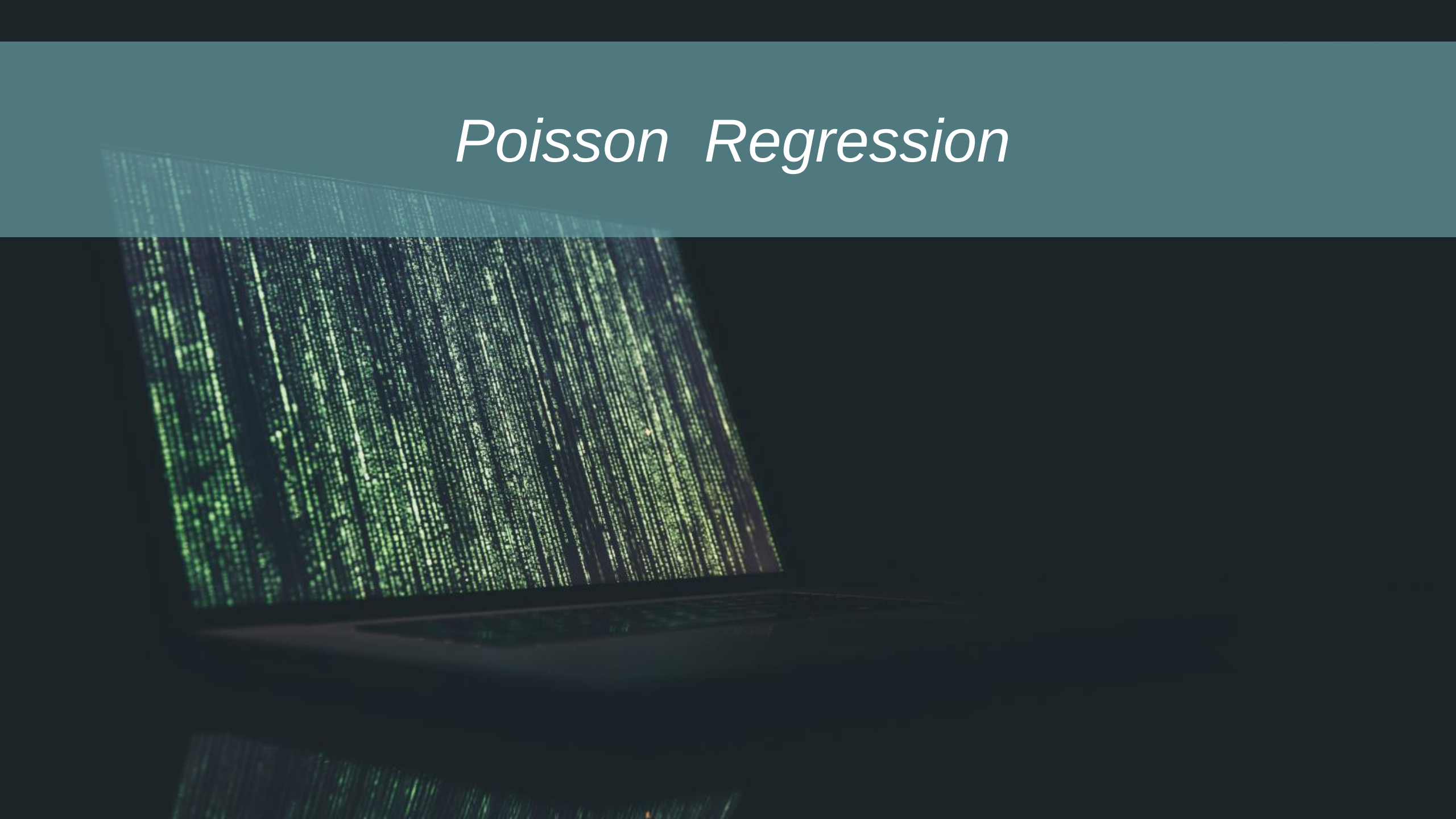
	College			
Failed	Business	Engineering	Life Sciences	Total
No	208	158	182	548
Yes	35	31	20	86
Total	243	189	202	634

$$Risk\ ratio_1 = \frac{0.144}{0.164} = 0.878$$

$$Risk\ ratio_2 = \frac{0.099}{0.164} = 0.604$$

- The above exercise is useful when we want to compare the risk across certain groups.
- This type of analysis however will not take us very far. The reason is that usually, we are interested in studying the effect that several variables have on the probability of the outcome.
- What if we wanted to see whether the risk of failure was affected by the college, gender, and the GPA, all at the same time?

Poisson Regression



Poisson Regression

- In linear regression, the model is represented by the linear equation

$$y = ax + b$$

- In the above equation, *y* is the dependent variable, *x* is the independent variable, *a* is the slope, and *b* is the y-intercept

Poisson Regression

- One of the nice things about linear regression is how easy it is to interpret the relationship between the dependent variable and the independent variable

$$y = 3x + 2$$

- If *x* is equal to 2, *y* will be equal to 8, and if *x* is equal to 3, *y* will be equal to 11. Note that for every one unit increase in *x*, the value of *y* increases by 3, which is the value of the slope

Poisson Regression

- Unfortunately, in Poisson regression things are not that simple. The reason is that the Poisson regression model has the following form:

$$\ln(\mu) = ax + b$$

- In the above equation, μ is the rate of occurrences, which is the dependent variable

Poisson Regression

$$\ln(\mu) = ax + b$$

- The above equation is linear, but instead of having the dependent variable on the left hand side we have the natural logarithm of the dependent variable.
- This means that the slope a represents the amount by which ln(μ) increases when x increases by one unit.
- As you can see, this is not a natural way of interpreting things.

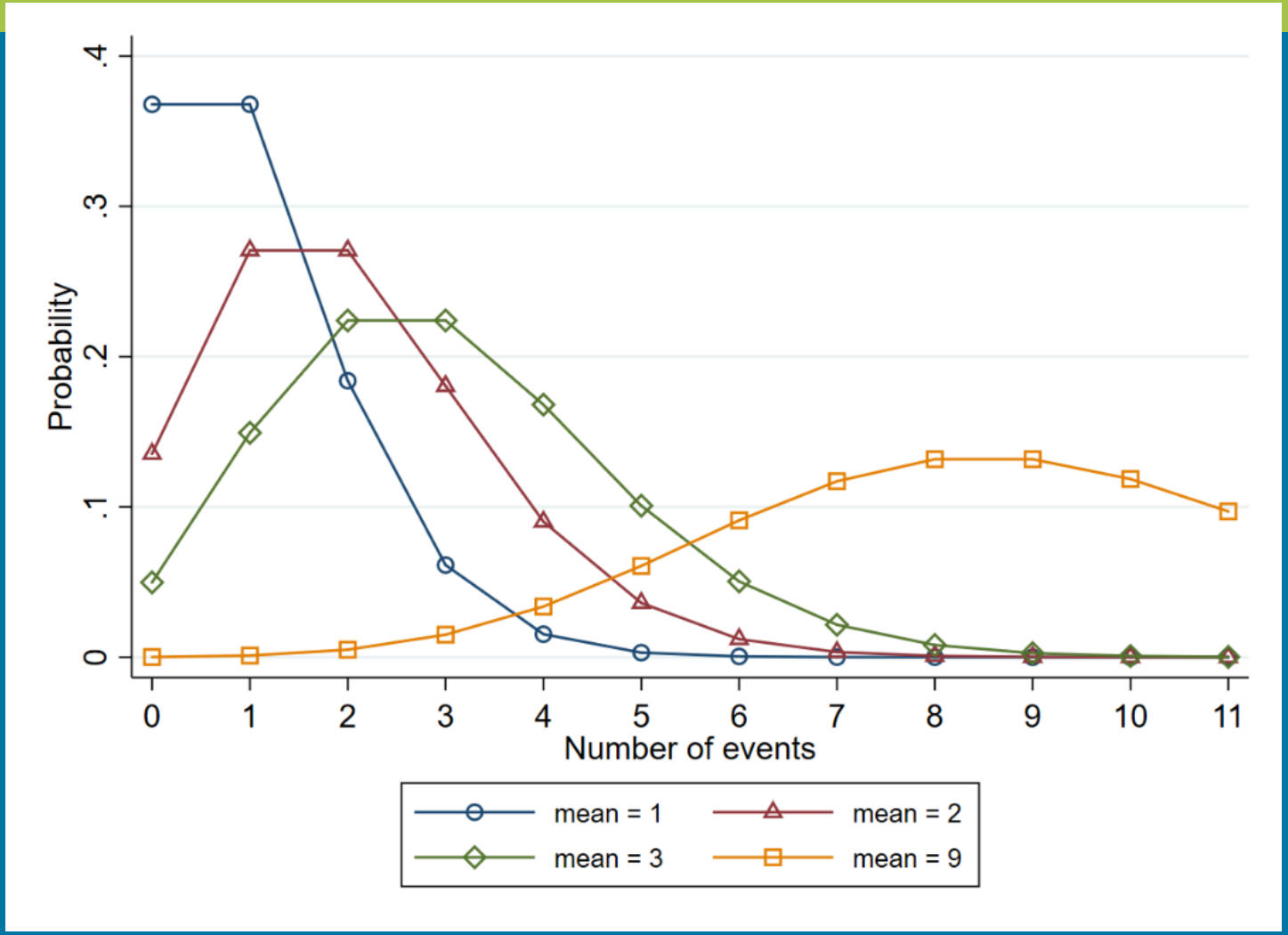
Poisson Regression

$$\ln(\mu) = ax + b$$

$$e^{\ln(\mu)} = e^{ax+b}$$

$$\mu = e^{ax+b}$$

Poisson Regression



Poisson Regression

$$\mu = e^{ax+b}$$

- This new form is better because now the dependent variable, which is the rate μ, is on the left side.
- For example, if a is positive, when x increases the term e^{ax+b} will increase. Since this term is equal to the rate of occurrence of the event, this means that the number of times that the event is expected to occur will also increase. On the other hand, when a is negative, when x increases the expected number of occurrences will decrease.

Illustration

- Assume that we perform Poisson regression where the dependent variable is the number of courses in which a student has failed and the independent variable is the GPA of the student:

$$\ln(\mu) = -0.099(GPA) + 7.091$$

- Let's consider the more intuitive form:

$$\mu = e^{-0.099(GPA)+7.091}$$

Now consider two students, one with a GPA of 77 and the other with a GPA of 78. According to our model, the expected number of withdrawals for each is:

Student with a GPA of 77: $\mu = e^{-0.099(77)+7.091} = 0.587$

Student with a GPA of 78: $\mu = e^{-0.099(78)+7.091} = 0.532$

$$Incidence - rate\ ratio = \frac{0.532}{0.587} = 0.906$$

The great news is that 0.906 is actually $e^{-0.099}$

$$\ln(\mu) = ax + b \qquad \mu = e^{ax+b}$$

- We now have a very intuitive meaning for the slope a.
- When we fit a Poisson model and obtain a value for the coefficient associated with an independent variable, we know that when the independent variable x increases by one unit, the expected number of occurrences is multiplied by e^a.
- When a is positive, e^a>1, which means that the expected number of occurrences increases when x increases.
- When a is negative, e^a<1, which means that the expected number of occurrences decreases when x increases

Example 1

Assume that we fit a Poisson model where the dependent variable is the number of customers that entered the store today, and where the independent variable is the number of advertisements that were ran on radio the preceding day. Once we fit the model we get the following results:

$$\ln(\mu) = 0.447(ads) + 0.241$$

Here, μ is the expected number of customers that will enter the shop. What does this output mean? Since the value of the coefficient associated with the independent variable, which is *ads*, is 0.447, this means that when *ads* increases by one, the expected number of customers is multiplied by $e^{0.447} = 1.564$. This means if the shop runs five radio ads the expected number of customers that will come is 1.564 times the expected number of customers if it runs four ads.

Example 2

Consider that we fit a Poisson regression model where the dependent variable is the number of times that a student goes out with his or her friends during the week and the independent variable is the student's grades. The output of the model is the following:

$$\ln(\mu) = -0.073(grades) + 6.629$$

Here, the coefficient is negative. Since $e^{-0.073} = 0.930$, the output indicates that the expected number of times that a student goes out during the week are multiplied by 0.93 (so they decrease) when grades increase by a single unit. This means that students with higher grades go out fewer times during the week.

Binary Variables

Number of failed courses	Gender	Binary
2	male	0
0	male	0
3	male	0
0	female	1
2	female	1
3	female	1
8	female	1
0	male	0
5	female	1
7	male	0
5	female	1
3	male	0
1	male	0
1	male	0
2	female	1
9	female	1
5	female	1
4	male	0
4	male	0
2	female	1

Gender

- What if we wanted to investigate whether the count of failed courses could be explained by the gender of the students? Here, the variable gender is not numeric
- In such a case, we can create a binary variable to represent the two categories. A binary number takes on the values of zero or one. We next assign each of these values to a category. Let us assign a zero to males and a one to females

Poisson Regression with a Binary Variable

- Now that the variable gender has been quantified, it is possible to include it in a regression model:

$$\ln(\mu) = ax + b$$

- If we use a statistical software to run the model, we will get the following output:

$$\ln(\mu) = 0.495(gender) + 0.916$$

Poisson Regression with a Binary Variable

- What does it mean that the coefficient of gender is 0.495?

$$\text{Male: } \mu = e^{0.495(0)+0.916} = 2.499$$

$$\text{Female: } \mu = e^{0.495(1)+0.916} = 4.1$$

- From these odds, we can calculate the incidence-rate ratio:

$$Incidence - rate\ ratio = \frac{4.1}{2.499} = 1.64$$

Poisson Regression with a Binary Variable

- This means that females have higher expected count than males.
- The nice thing is that the number 1.64 happens to be $e^{0.495}$.
- This means that when we are dealing with binary variables, the exponent of the coefficient is the incidence-rate ratio when we compare an individual who belongs to the group that is assigned a value of one and an individual who belongs to the group that is assigned the value zero

Number of failed courses	GPA	Gender	Binary
2	80	male	0
0	95	male	0
3	77	male	0
0	90	female	1
2	75	female	1
3	72	female	1
8	60	female	1
0	82	male	0
5	74	female	1
7	69	male	0
5	69	female	1
3	79	male	0
1	81	male	0
1	78	male	0
2	83	female	1
9	62	female	1
5	72	female	1
4	70	male	0
4	71	male	0
2	87	female	1

Multiple Independent Variables

- The table includes the dependent variable which is the number of courses in which the student has failed and the independent variables gender and GPA
- The equation of this model is:

$$\ln(\mu) = a_1 x_1 + a_2 x_2 + b$$

Poisson Regression with Multiple Variables

- If we run the model, the output will be:

$$\ln(\mu) = -0.086(GPA) + 0.033(gender) + 7.464$$

- The more intuitive form of the equation:

$$\mu = e^{-0.086(GPA)+0.033(gender)+7.464}$$

- Let us now calculate the expected count for two students where both of them have a GPA of 74, but one is male and the other is female:

$$\text{Male: } \mu = e^{-0.086(74)+0.033(0)+7.464} = 3.004$$

$$\text{Female: } \mu = e^{-0.086(74)+0.033(1)+7.464} = 3.105$$

Poisson Regression with Multiple Variables

- This means that the incidence-rate ratio is:

$$\text{Incidence–rate ratio: } \frac{3.105}{3.004} = 1.034$$

- A simpler way to get this value is just to calculate the exponent of the coefficient, $e^{0.033} = 1.034$
- This shows that even when there are several independent variables, the coefficients retain their meanings
- Therefore, to find the difference between two groups of students, just calculate e^{a_1}

Poisson Regression with Multiple Variables

- Let us now calculate the expected count for two female students, one of whom has a GPA of 79 and another who has a GPA of 80:

$$\text{GPA of 79: } \mu = e^{-0.086(79)+0.033(1)+7.464} = 2.02$$

$$\text{GPA of 80: } \mu = e^{-0.086(80)+0.033(1)+7.464} = 1.853$$

- This means that the incidence-rate ratio is:

$$\text{Incidence-rate ratio: } \frac{1.853}{2.02} = 0.917$$

- This is also obtained by finding the exponent of the coefficient, $e^{-0.086} = 0.917$

Categorical Variables

College (Three Groups)

- What if we had a categorical variable that divided the observations into more than two groups?

	College			
Failed	Business	Engineering	Life Sciences	Total
No	208	158	182	548
Yes	35	31	20	86
Total	243	189	202	634

- In this case, we cannot use a binary variable because there are three groups instead of two

College (Three Groups)

	College			
Failed	Business	Engineering	Life Sciences	Total
No	208	158	182	548
Yes	35	31	20	86
Total	243	189	202	634

The number of binary variables needed is the number of categories minus one. In our case, we have four categories, so it is 3 – 1 = 2

	x_1	x_2
Engineering	0	0
Business	1	0
Life sciences	0	1

$$\ln(\mu) = a_1 x_1 + a_2 x_2 + b$$

Poisson Regression with a Categorical Variable

- If we fit this model to the data, the output will be:

$$\ln(\mu) = -0.079x_1 - 0.438x_2 + 1.237$$

Let us now calculate the expected number of occurrences for the three types of students. As usual, we use the more intuitive form of the equation:

$$\mu = e^{-0.079(x_1)-0.438(x_2)+1.237}$$

Poisson Regression with a Categorical Variable

Engineering:

$$\mu = e^{-0.079(0)-0.438(0)+1.237} = 3.445$$

Business:

$$\mu = e^{-0.079(1)-0.438(0)+1.237} = 3.184$$

Life Sciences:

$$\mu = e^{-0.079(0)-0.438(1)+1.237} = 2.223$$

Poisson Regression with a Categorical Variable

We can now calculate the incidence-rate ratios in order to be able to compare different groups:

$$\frac{Business}{Engineering} = \frac{3.184}{3.445} = 0.924$$

$$\frac{Life\ sciences}{Engineering} = \frac{2.223}{3.445} = 0.645$$

We can get the same values by calculating the exponents of the coefficients: $e^{-0.079} = 0.924$ and $e^{-0.438} = 0.645$

Poisson Regression with a Categorical Variable

- We see that the exponent of the coefficient for each variable produces the incidence-rate ratio when we compare the group associated with the variable to the base group, which is the group that is assigned the values of zero.
- In other words, in our example, engineering students are the base, or referent group, since they have a value of zero for both x_1 and x_2. Business students have a value of one for x_1, which means that the exponent of the coefficient of x_1 is the incidence-rate ratio of business students to engineering students.
- Life sciences students have a value of one for x_2, which means that the exponent of the coefficient of x_2 is the incidence-rate ratio of life sciences students to engineering students.

Exposure

Exposure

- There is an issue here which you have probably not noticed
- Earlier, we had calculated the incidence rate ratios for the exact same data using the count table, and we got different answers

	Count tables	Poisson regression
Business / Engineering	0.878	0.924
Life sciences / Engineering	0.604	0.645

- Why are the results different?
- The answer is actually simple. When we calculated the incidence-rate ratios using the count tables, we took into consideration the total number of courses taken by each group of students. In the count tables section, we did not compare the total number of failed courses for business students with the total number of failed courses for engineering students. We compared the proportion of failed courses for business students with the proportion of failed courses for engineering students (this is why we were dividing the number of failed courses by the total number of courses).
- This is an important point because the larger the number of courses taken by a group, the larger the expected number of failures

- Because the concept of exposure is important, we need to tell the statistical package to take it into account when calculating the regression equation. So far, we have not been doing that
- Let us now tell the statistical software to take into account the exposure in each observation. In this case, the exposure is the total number of courses taken by each student

Number of failed courses	College	Total number of courses
3	Business	23
0	Business	24
5	Business	30
7	Business	28
1	Engineering	6
2	Business	17
2	Business	15
0	Business	18
4	Engineering	27
2	Engineering	15
5	Engineering	28
3	Engineering	17
8	Business	27
0	Business	9
1	Engineering	27
2	Engineering	10
3	Business	20
4	Engineering	25
5	Business	32
9	Engineering	34
0	Life sciences	17
1	Life sciences	16
1	Life sciences	20
3	Life sciences	28
5	Life sciences	32
4	Life sciences	26
3	Life sciences	17
1	Life sciences	18
2	Life sciences	28

Exposure

Now we need to tell the statistical software to run a Poisson regression while taking into account that different students have gone through a different number of courses. The following is the output of this model:

$$\ln(\mu) = -0.130x_1 - 0.505x_2 - 1.808$$

As usual, we use the more intuitive form of the equation:

$$\mu = e^{-0.130(x_1)-0.505(x_2)-1.808}$$

Engineering:

$$\mu = e^{-0.130(0)-0.505(0)-1.808} = 0.164$$

Business:

$$\mu = e^{-0.130(1)-0.505(0)-1.808} = 0.144$$

Life Sciences:

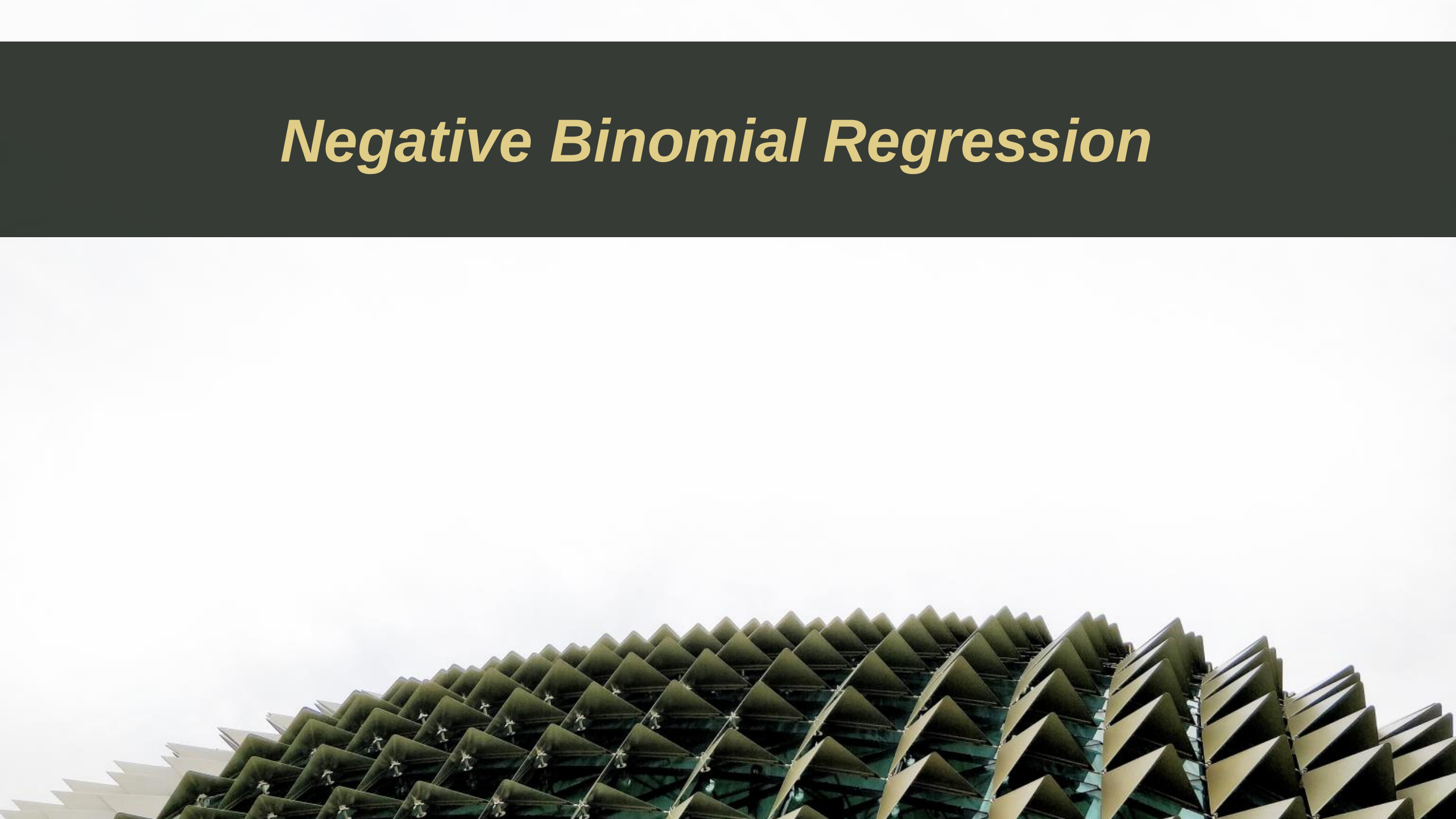
$$\mu = e^{-0.130(0)-0.505(1)-1.808} = 0.099$$

We can now calculate the incidence-rate ratios in order to be able to compare different groups:

$$\frac{Business}{Engineering} = \frac{0.144}{0.164} = 0.878$$

$$\frac{Life\ sciences}{Engineering} = \frac{0.099}{0.164} = 0.604$$

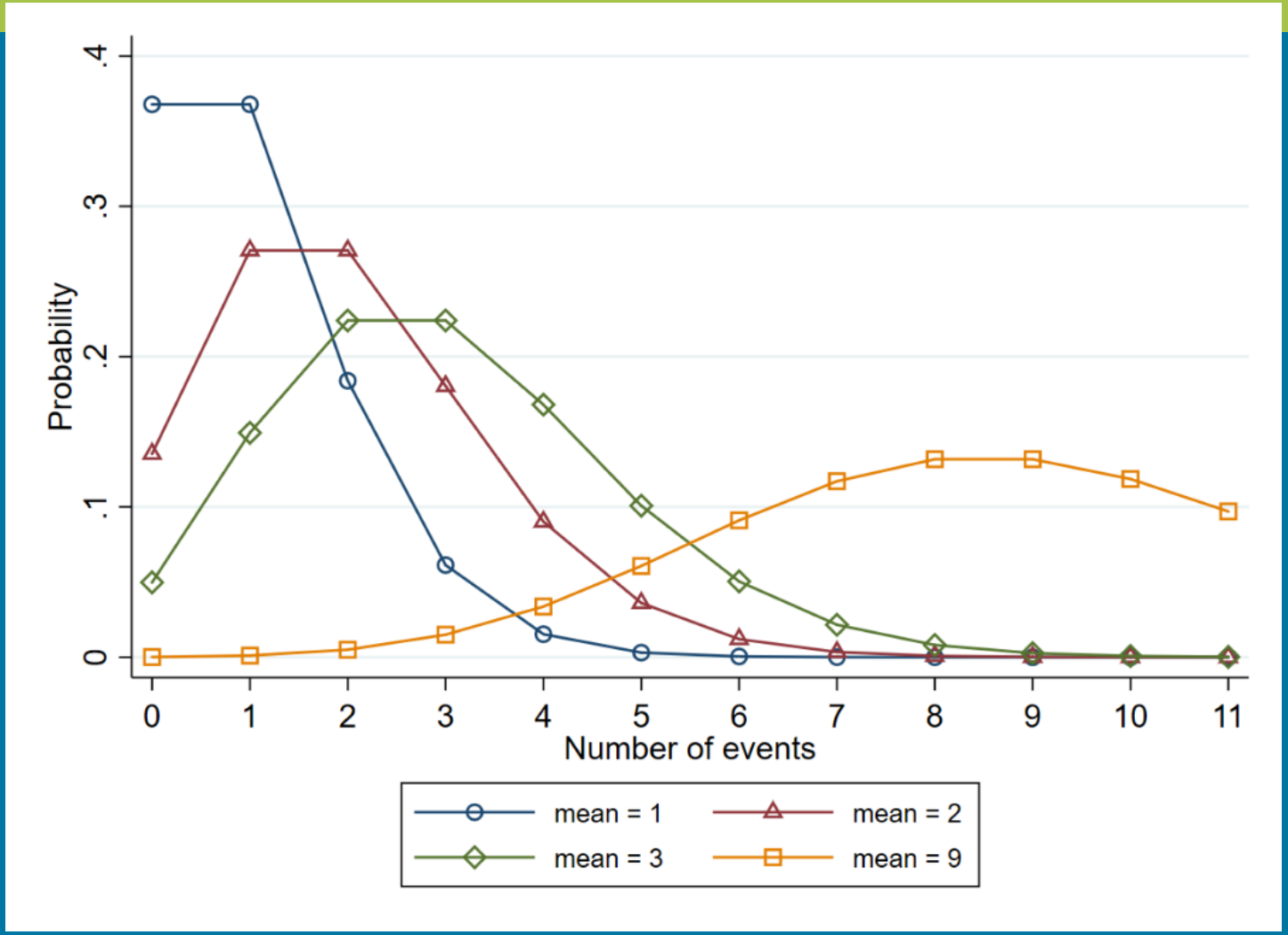
Negative Binomial Regression

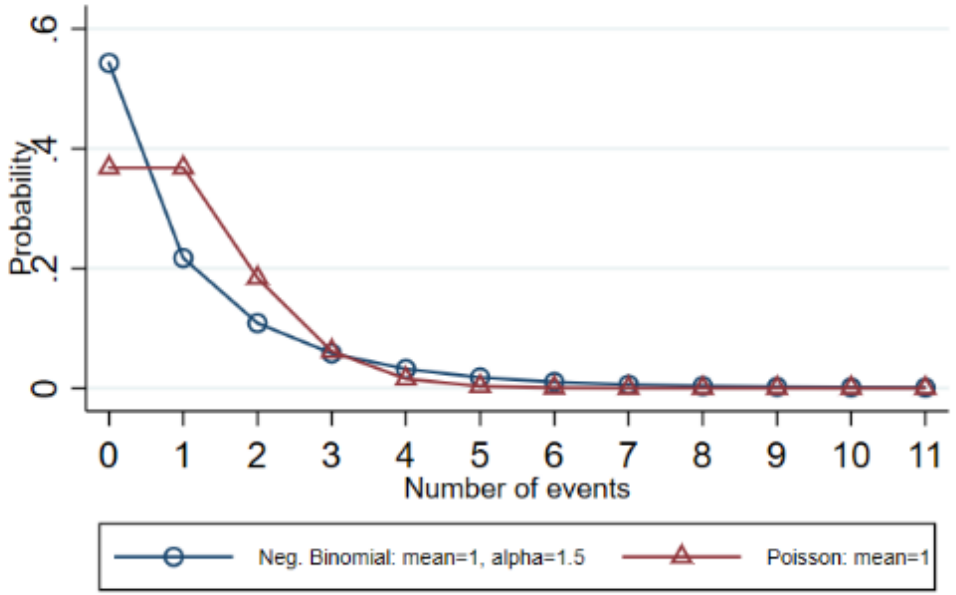


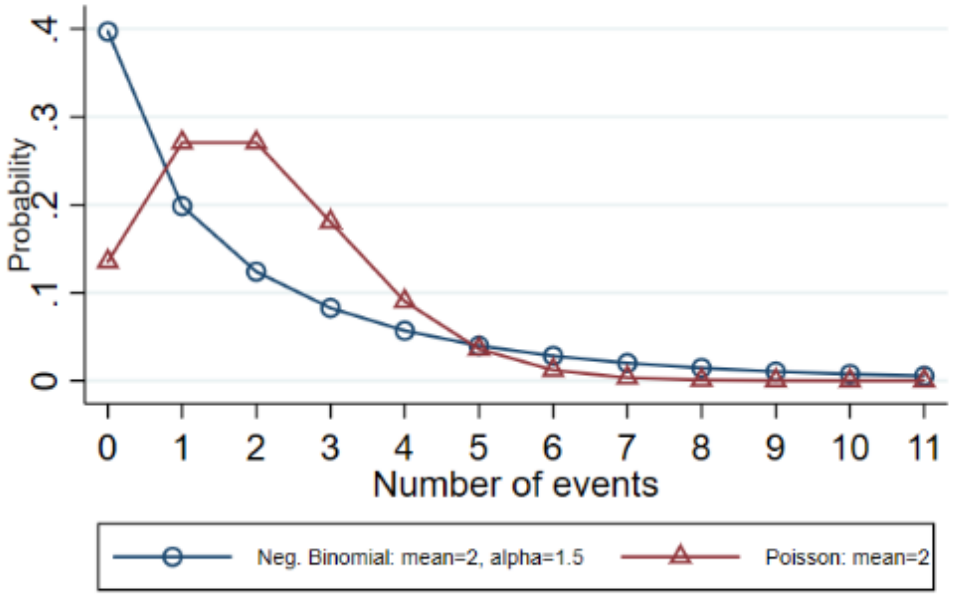
Overdispersion

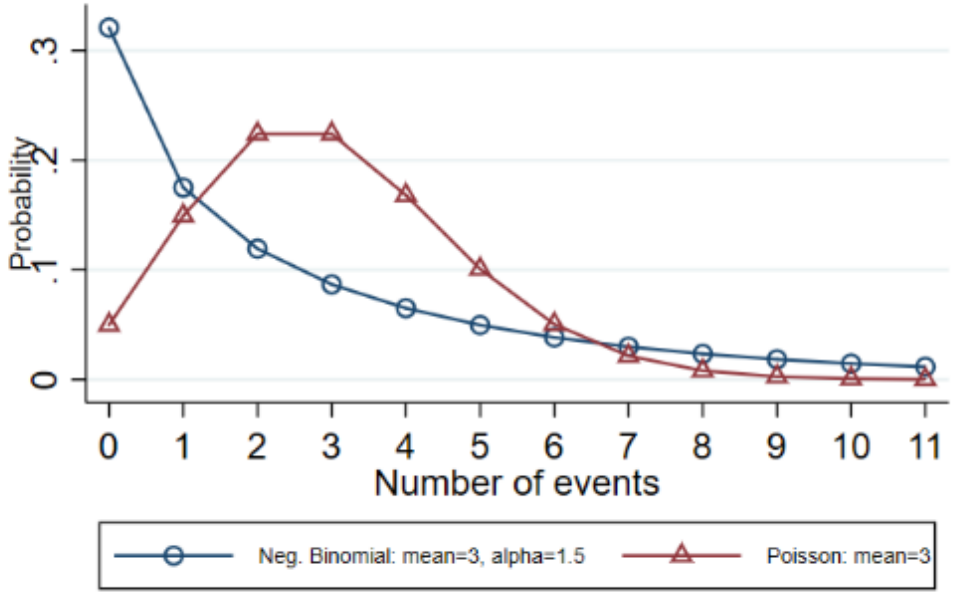
- Although the Poisson regression model is the basic count model, it actually rarely fits the data because of what is referred to as overdispersion
- An important characteristic of the Poisson probability density function is that the mean and the variance are equal. This means that as the mean increases, so does the variability in the data, a characteristic that is called equidispersion.
- When this assumption is violated, we say that the data displays overdispersion
- In order to address this issue, a negative binomial regression model is used

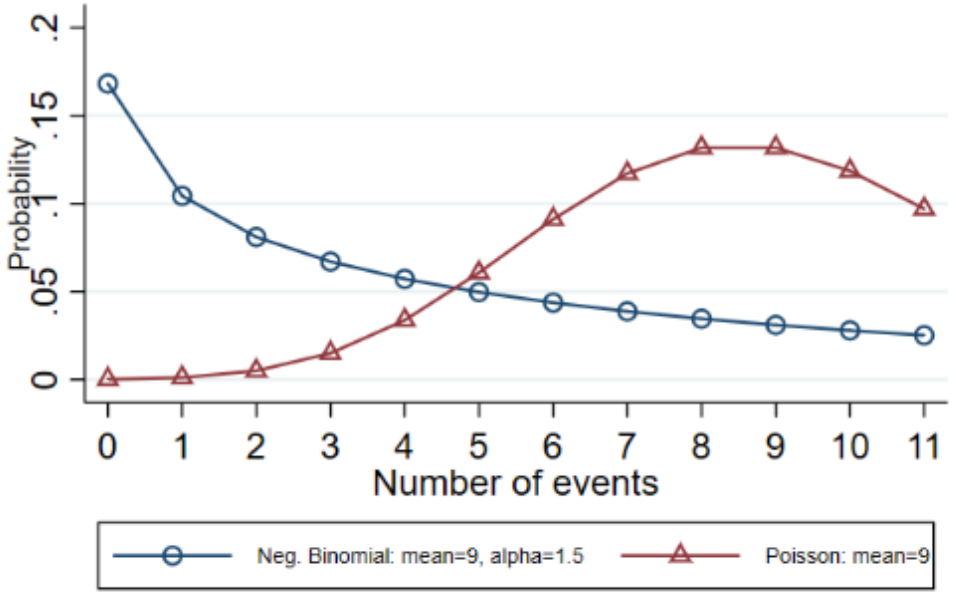
Poisson Regression




Probability
0
.2
.4
.6
0 1 2 3 4 5 6 7 8 9 10 11
Number of events
Neg. Binomial: mean=1, alpha=1.5
Poisson: mean=1


Probability
0
.1
.2
.3
.4
0 1 2 3 4 5 6 7 8 9 10 11
Number of events
Neg. Binomial: mean=2, alpha=1.5
Poisson: mean=2


Probability
0
.1
.2
.3
0 1 2 3 4 5 6 7 8 9 10 11
Number of events
Neg. Binomial: mean=3, alpha=1.5
Poisson: mean=3

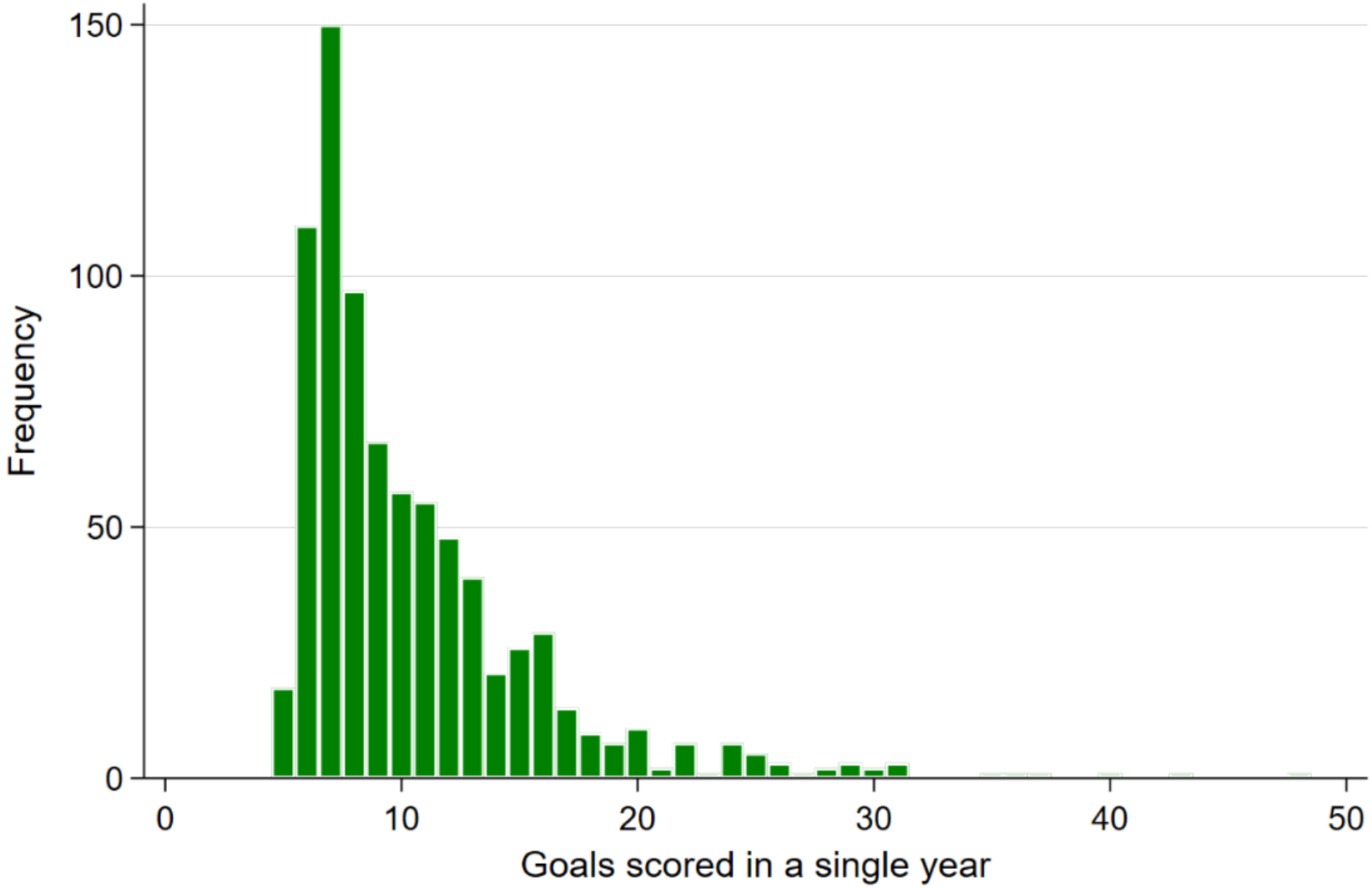

Probability
0
.05
.1
.15
.2
0 1 2 3 4 5 6 7 8 9 10 11
Number of events
Neg. Binomial: mean=9, alpha=1.5
Poisson: mean=9

Negative Binomial Regression

- What implication does this have for us?
- Fortunately, very little.
- Since we are not interested in the math that goes on behind the scenes, all you need to know is that when we fit a Poisson regression model, we should always follow it up with a negative binomial regression model in order to test whether overdispersion exists. The beauty of it all is that everything we have covered with regards to the meaning of the coefficients when adding the independent variables still applies the exact way.

Truncated Models

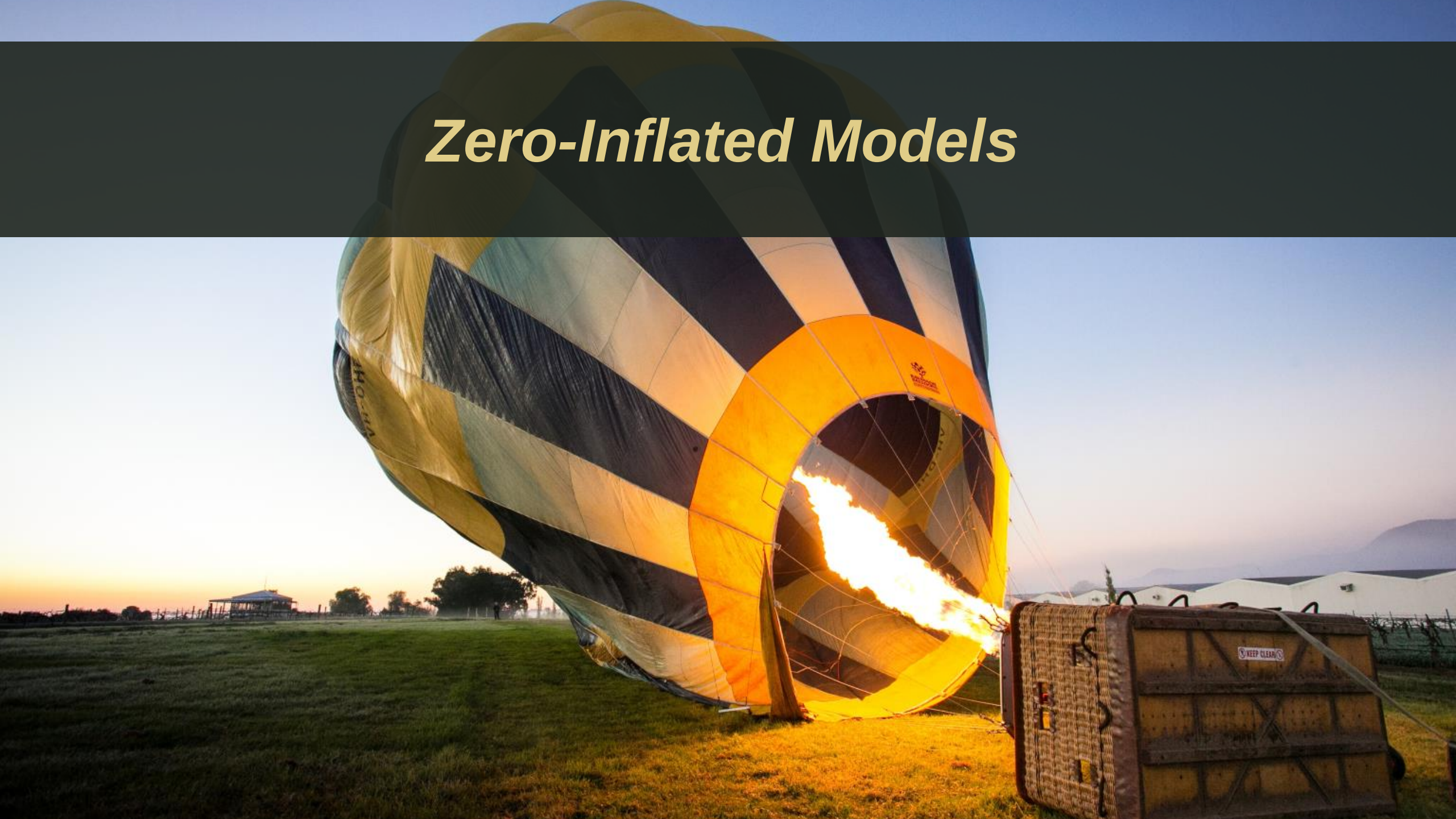
Example


150
100
50
0
Frequency
0
10
20
30
40
50
Goals scored in a single year

Truncated Models

- In such cases, whether the data is zero-truncated or truncated at any other point, we use what is referred to as truncated models.
- These models take into account that a count that is less than a certain value is not possible
- There is a truncated Poisson model and a truncated negative binomial model
- Everything that we have said previously about the Poisson and the negative binomial models applies to the truncated models: the meaning of the coefficients, the incidence-rate ratios, and testing for overdispersion

Zero-Inflated Models



Zero-Inflated Models

- When we discussed the negative binomial model, it was noted that the model corrects for the underprediction of zero counts
- Sometimes the number of zeros in the dataset is much larger than what both the Poisson and the negative binomial model assume. In such a case, we say that the number of zeros is inflated, i.e. it is greater than usual
- Why would the number of zeros be inflated? This might be due to an underlying mechanism that is acting like a hurdle

Zero-Inflated Models

- As an example, assume that we want to model the number of heart attacks that men under 45 have suffered
- In such a case, we would expect that most men at such a young age would not have suffered from a heart attack. A normal male under 45 years of age should not have suffered from a heart attack
- This means that the dataset would contain a disproportionately large number of zeros
- In this case, we can think of the dataset as containing two different types of men: healthy men who have a zero count, and men with health issues who have a count that is greater than zero

Zero-Inflated Models

- When we have this type of situation, we use a zero-inflated model to account for the large number of zero counts
- Zero-inflated models are thus made up of two parts:
  - The first part is a logistic model that predicts whether someone has never experienced the event or has experienced it at least once
  - The second part models the number of times that the event has been experienced by those who have experienced the event at least once

Step 1

- If a male is healthy and living under normal conditions, we would expect that he has had no heart attacks, i.e. the count is zero. For males that are unhealthy for their age, or who lead a very stressful life, we would expect that the count would be greater than zero
- This means that we can divide the observations into two groups. The members of the first group have a count of zero and the members of the second group have a count that is greater than zero
- To model this situation, we can think of a dependent variable that is binary: an individual is either in the first group or in the second group. This is done by using a binary model that predicts the probability that the event has never occurred as opposed to it having occurred at least once

Step 2

- After predicting whether an individual is in the first group (where the count is zero) or the second group (where the count is greater than zero), the analysis moves on to predicting the counts for those in the second group
- This is done by using a count model, such as a Poisson model or a negative binomial model

Inflation Variables

- The output of zero-inflated models is divided into two parts, one part for each step.
- The output helps us understand what are the independent variables that lead to someone being in either of the two groups, and what are the independent variables that increase the frequency of the count for those who have experienced the event.
- The two sets of the variables need not be the same. This means that the variables that determine to which group an individual belong can be different than the variables that lead to a higher count

Example

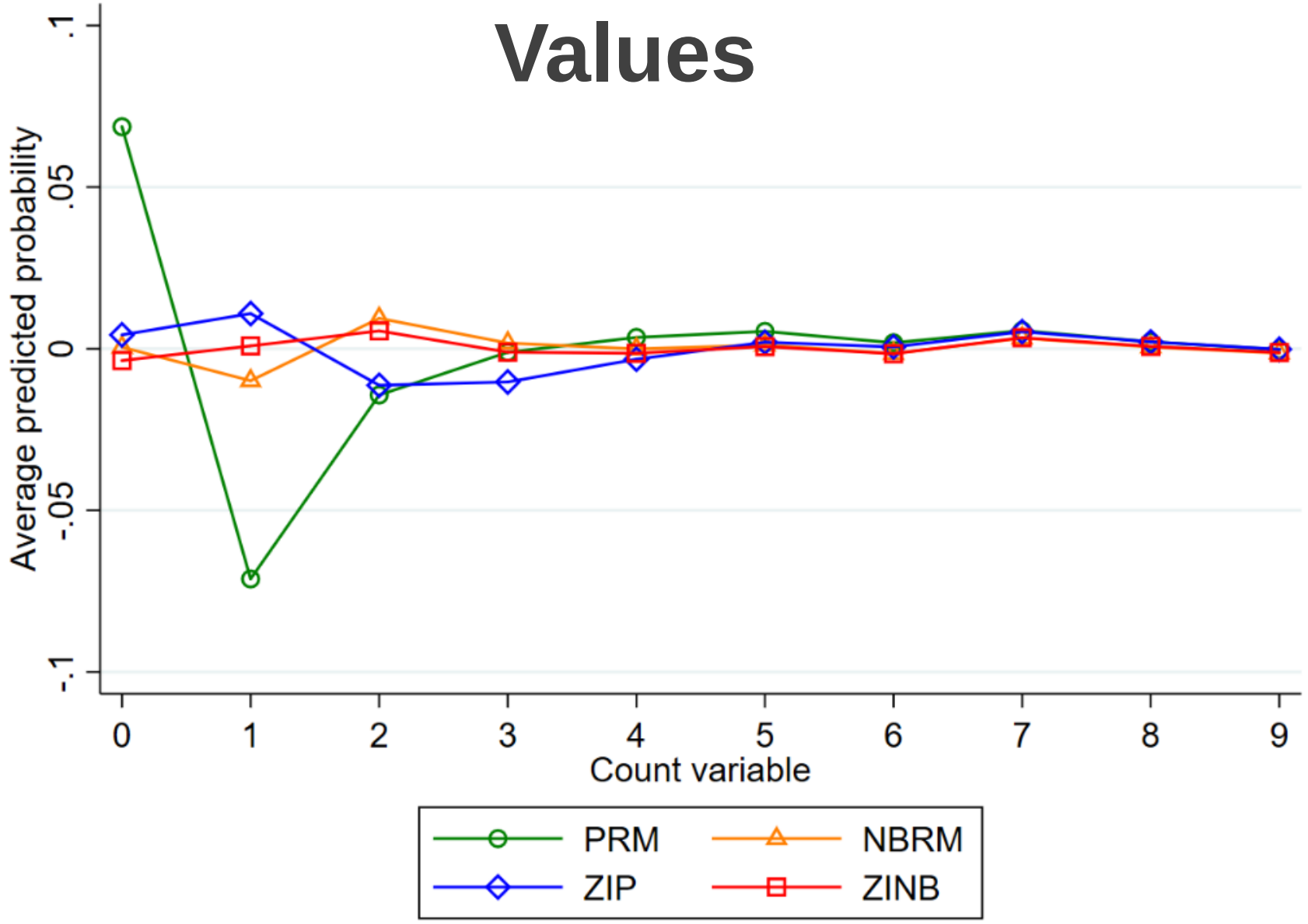
- Consider a dataset that contains the count variable visits which records the number of times that a patient has visited the doctor in the past year
- Assume that this variable has an unusually large number of zeros. This means, that many of the patients have had no need to visit the doctor during the past year
- Also, assume that recently the hospital has modified its internal policies in order to increase the efficiency of their patient care. The purpose of these reforms is to make sure that the hospital staff are able to respond quickly and efficiently to the needs of the patients thereby reducing the number of subsequent visits from the same patient
- In this case, we would expect that the new reforms would reduce the frequency of the counts, but they would not have an effect on whether a patient initially visits or not.

Example

	Coefficient	$e^{Coefficient}$	P-value
Count			
gender	0.20	1.22	0.07
age	0.30	1.35	0.00
reform	-0.19	0.83	0.00
Inflate			
gender	0.64	1.90	0.01
age	-0.55	0.58	0.00
reform	0.20	1.22	0.12

Comparison of Models

Comparing Predicted Values with Observed Values



Other Tests

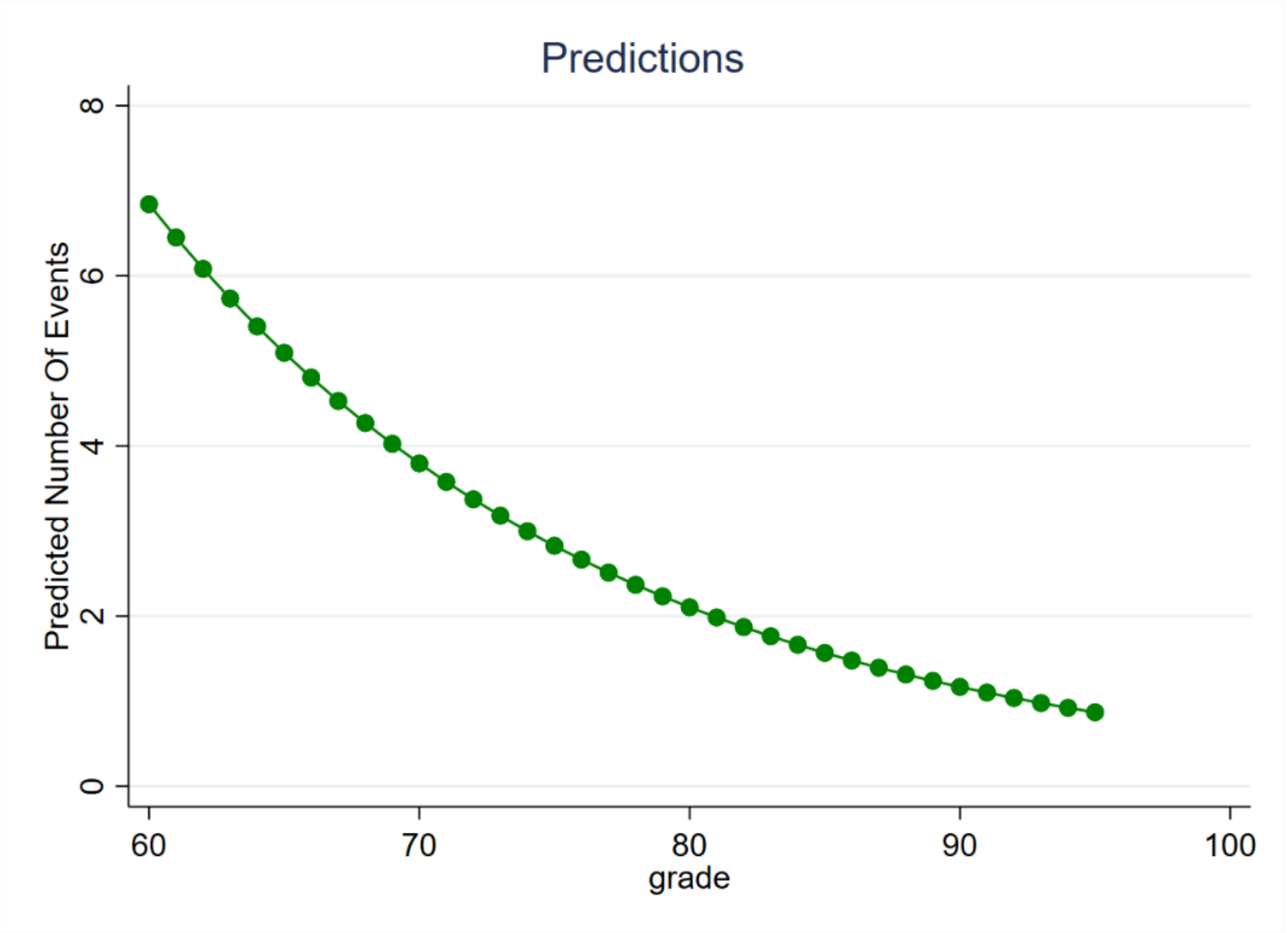
- Likelihood-Ratio Test of Alpha
  - This test can be used to compare the Poisson model to the negative binomial model, and it can also be used to compare the zero-inflated Poisson model to the zero-inflated negative binomial model
- Vuong Test
  - Used to compare regular model with the zero-inflated model
- AIC and BIC Statistics
  - Favor the one with smaller values of both

Prediction

Prediction

- In linear regression we use the model in order to predict the dependent variable. In count models, there are two different predictions that we can make
- First, we can predict the value of the dependent variable by predicting the number of events for certain values of the independent variables
- Second, we can use count models in order to predict the probabilities for several values of the count variable

Predicting the Number of Events



Predict the Probabilities for Certain Values