**Slide-1**

**Advanced Regression**

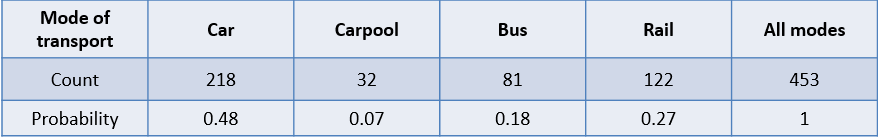
**Agenda**

* Multinomial Regression
* Poisson Regression
* Negative Binomial
* Zero Inflated
* Hurdle

**Slide-2**

**Multinomial Regression**

* Logistic regression (Binomial distribution) is used when output has ‘2’ categories
* Multinomial regression (classification model) is used when output has > ‘2’ categories
* Extension to logistic regression
* No natural ordering of categories



* Response variable has > ‘2’ categories & hence we apply multilogit
* Understand the impact of cost & time on the various modes of transport

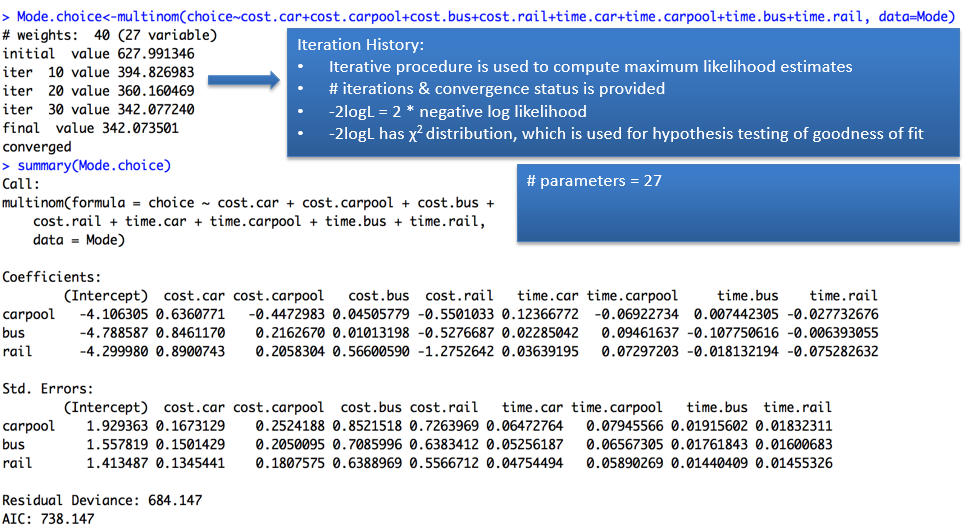
**Slide-3**

**Multinomial Regression**

* Whether we have ‘Y’ (response) or ‘X’ (predictor), which is categorical with ‘s’ categories
  + Lowest in numerical / lexicographical value is chosen as baseline / reference
  + Missing level in output is baseline level
  + We can choose the baseline level of our choice based on ‘relevel’ function in R
  + Model formulates the relationship between transformed (logit) Y & numerical X linearly
  + Modeling quantitative variables linearly might not always be correct



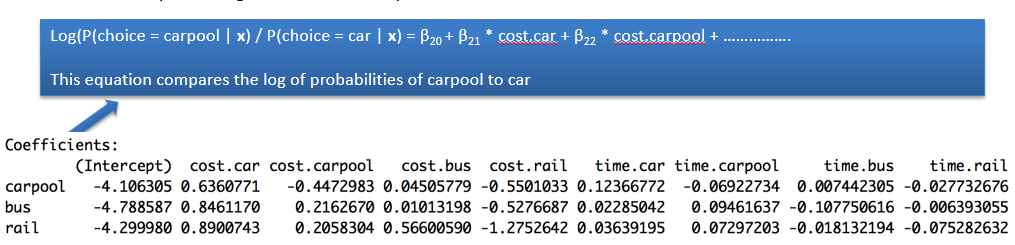
**Slide-4**



**Slide-5**

**Multinomial Regression - Output**

* ‘car’ has been chosen as baseline
* x = vector representing the values of all inputs



* The regression coefficient 0.636 indicates that for a ‘1’ unit increases the ‘cost.car’, the log odds of ‘carpool’ to ‘car’ increases by 0.636
* Intercept value does not mean anything in this context
* If we have a categorical X also, say Gender (female = 0, male = 1), then regression coefficient (say 0.22) indicates that relative to females, males increase the log odds of ‘carpool’ to ‘car’ by 0.22

**Slide-6**

**Probability**

* Let p = p(**x** | A) be the probability of any event (say attrition) under condition A (say gender = female)

**Odds**

* Then p(**x** | A) ÷ (1 - p(**x** | A) is called the odds associated with the event

**Odds Ratio**

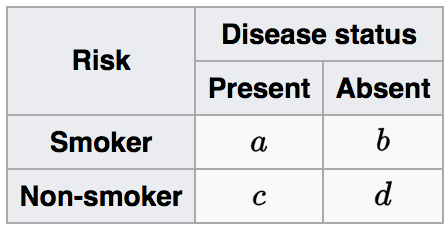
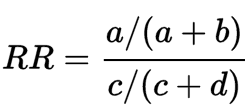
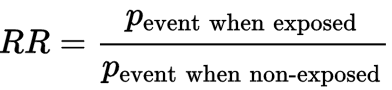
* If there are two conditions A (gender = female) & B (gender = male) then the ratio

P(**x** | A) ÷ (1 - p(**x** | A) / p(**x** | B) ÷ (1 - p(**x** | B) is called as

Odds ratio of A with respect to B

**Relative Risk**

* p(**x** | A) ÷ p(**x** | B) is called as relative risk

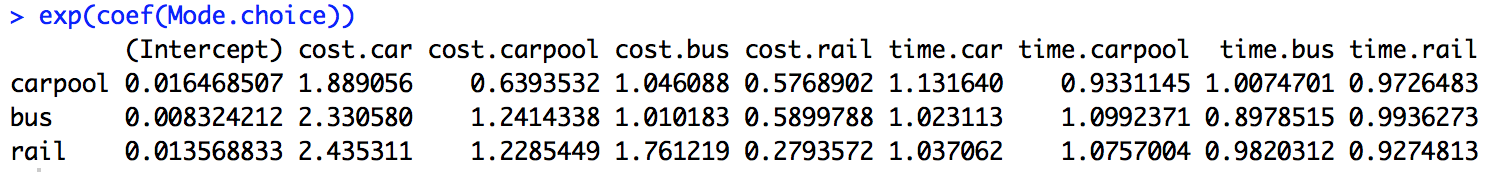


<https://en.wikipedia.org/wiki/Relative_risk>

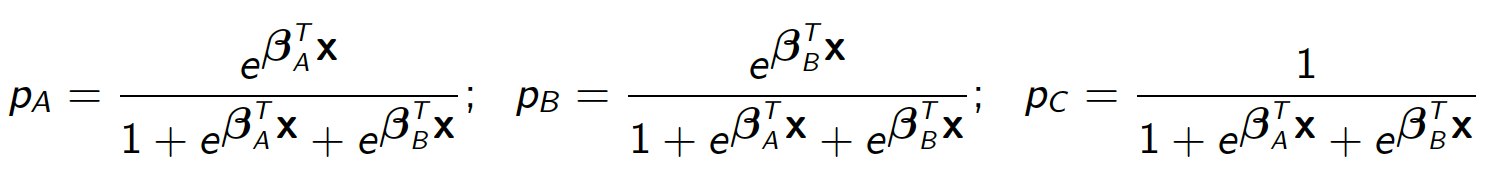
Slide-7

Odds Ratio

* Odds ratio is computed from the coefficients in the linear model equation by simply exponentiating.
* Exponentiated regression coefficients are odds ratio for a unit change in a predictor variable



* The odds ratio for a unit increase in cost.car is 1.88 for choosing carpool vs car



**Slide-8**- **Goodness of fit**

|  |  |
| --- | --- |
| Linear | GLM |
| Analysis of Variance | Analysis of Deviance |
| Residual Deviance | Residual Sum of Squares |
| OLS | Maximum Likelihood |

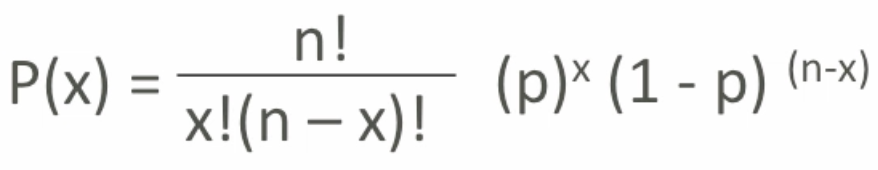
* Residual Deviance is -2 log L
* Adding more parameters to the model will reduce Residual Deviance even if it is not going to be useful for prediction
* In order to control this, penalty of “2 \* number of parameters” is added to to Residual deviance
* This penalized value of -2 log L is called as AIC criterion
* AIC = -2 log L + 2 \* number of parameters

Note: “Multilogit Model with *Interaction*”

**Slide-9**

**Binomial Distribution**

* Two possible outcomes & determine probability of # of successes over given number of trials & not its magnitude
* Used for discrete independent events with replacement



* + x = event/ # defectives / # of successes desired
  + p = probability that ‘x’ will occur on any particular occasion
  + n = # of trials / sample size
* Two ways of calculating binomial probabilities – formula or distribution table
* # of defective or non-defective items in a sample

**Slide-10**

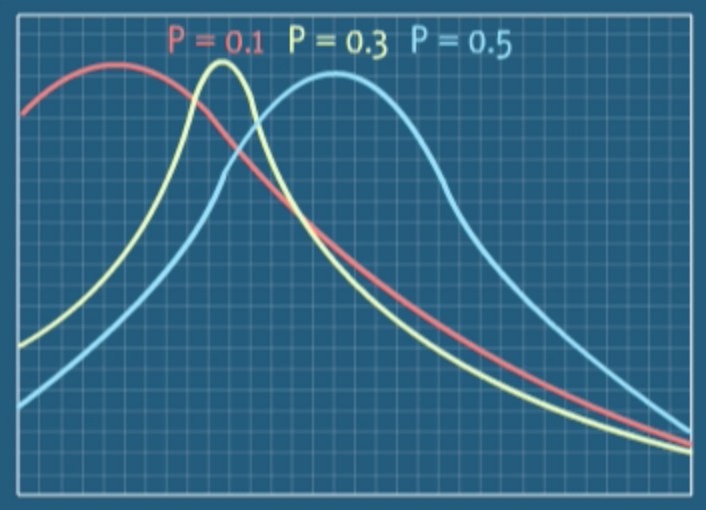
**Binomial Distribution**

* Mean μ = np; Standard Deviation σ = √np(1-p)
* You are a data scientist in an electronics manufacturing plant. Historically, the USB manufacturing process is shown to yield about 6% defectives. Find out if the specification limit (4) is exceeded, for each day producing exactly 50 items
* n = # of trials = 50; p = 6% => Mean = 50 \* 0.06 = 3

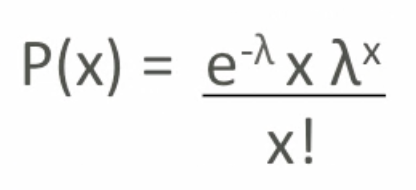
**Slide-11**

**Poisson Distribution**

* Special case of binomial distribution when probability ‘p’ is small
* Occurrences within a unit space or time



* Fixed observation period
* Ideal for modeling rare events, which are independent & occurs at constant rate

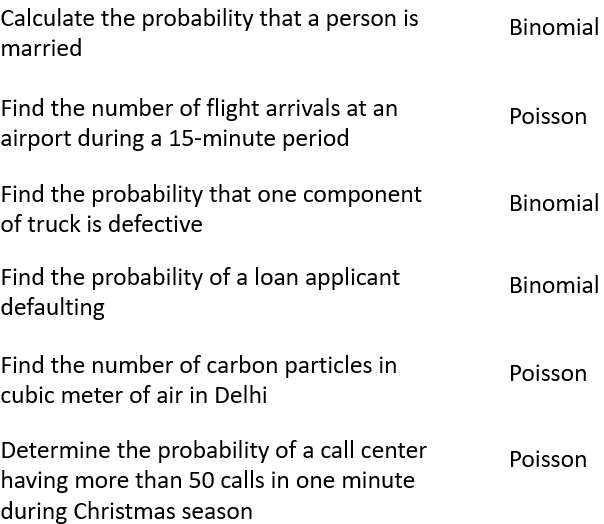


* λ = mean = variance = np; x = # defects; e = base of natural logarithm

# Of defects in fuel pipes produced by a company averages 6 per day. Also daily production is the same. What is the probability of getting exactly 8 defects on a given day?

**Slide-12**

**Quiz**



**Slide-13**

**Negative Binomial Distribution**