**Count-Related Models**

1. **Poisson distribution**

 the Poisson is a [discrete probability distribution](https://en.wikipedia.org/wiki/Discrete_probability_distribution) that expresses the probability of a given number of events occurring.

(1)

The meaning of this probability (density) function is:

Take the meteorite for an example. Assume that the expected number of meteorite we can observe during a quarter is 20 times i.e., =20. Then, if we wanna know the probability of observing 10 times meteorite, it can be computed by introducing =20 and y=10 into above function.

Therefore, in using Poisson distribution to describe a random variable, you should know or calculate **the expected value of the variable**, i.e**., mean or ** first.

Another important property of the probability function of Poisson distribution is that

the positive [real number](https://en.wikipedia.org/wiki/Real_number)   is equal to the [expected value](https://en.wikipedia.org/wiki/Expected_value) of *Y* and also to its variance.

(Just keep in mind it. Do not ask why. It is unnecessary to know it for now.)

1. **Why need Poisson regression model?**

Often there exists over-dispersion in Y of your dataset. This is because the distribution of Y could be affected by some other independent variables.

For instance, the weather condition in different quarters leads the expected value of meteorite observations to vary across quarters, thus the expected value  should be dependent on some independent variables describing the weather.

Therefore, we use  to model the expected value of Y under different values of X, which depict the weather condition in our example.

Namely, **** = .

Then, after you get the values of X’s coefficients in the Poisson regression model, given specific values of independent variable X, you should:

1. Calculate the expected value (or mean) using **** = .
2. Based on the ****, use function (1) above to calculate the probability of Y
3. **Comparison of Count models**
4. Negative binomial regression: Allows for unobserved heterogeneity, which means that it is not necessary for the mean and variance to be equal. It is a more generic regression model for count data.
5. Zero inflated models: if there are more zeros than expected, assume another process, which produces the zeros in the dataset. In this case, use zero inflated model to regress your data.
6. Zero truncated Poisson: for the dataset having more zeros than expected, you could use zero truncate Poisson as well. Just truncate the data with zero first and then run Poisson regression or negative binomial regression on the truncated data.

**Limited dependent variable models**

1. **Understand what is censored and truncated data**
2. **How to model censored and truncated data**
3. if using OLS:

problem: compared with the true relation between dependent and independent variables, the slope of independent variable in the regression results of OLS is biased towards zero and the intercept is biased positively.

1. Therefore, use **Tobit regression method** for censored data:

**Question 1 :** given the values of dependent and independent variables, calculate the probability of such a data instance?

**Answer**:

Based on the coefficients in the regression result, the probability for the censored data and normal data(e.g., Yi >0 if data is censored at zero) are as follows:

For the data on the lower censored boundary ( if censored on lower bound):

Here C\_lower is the boundary,



e.g.,If C\_lower is zero, above probability formula is written as: **same as the slides**



For the data on the upper censored boundary:



e.g.: If C\_upper bound is 2000( the example in the slide), above formula is written as



For the normal data: the probability is the same for any censor boundary C:



**Question 2:** give the likelihood function for top censoring?

**Answer**:

For top censored data:



2000 is the top censor bound. Change it according to the problem description.

For double censoring data:



0 and 2000 are the upper and lower censor bounds. Change them according to the problem description.

**Question 3:** interpreting the regression result of tobit given some values of independent variables?

**Answer:**

First, as usual, compute .

Second, say:



**Question 4:** provide the marginal effect of Xk, given the values of the independent variables?

**Answer**:

For normal data:

, the marginal effect of Xk is the coefficient 

For the data on the censored boundary:

here is an example.

Note that 0 and 2000 are the lower and upper censored bound. Chang them according to the problem.



1. **truncate regression**

**Question1:** provide the likelihood function for truncate regression?

**Answer:**

**For the lower truncated data:**



Replace the C\_{lower} with the value in the problem. If the lower truncate bound is zero, above formula is the same as the slide one:



Could answer with this one.

**For upper truncated data:**



If upper bound is 2000, as the example in the slide, above formula becomes:



**For both upper and lower truncate data:**

****

For example, if lower bound is 0 and upper bound is 2000, as the example in the slide:

Above formula becomes:



**Question2: given the values of independent variable, calculate the probability of this data instance?**

**Answer:**

Based on the coefficients of variables in the regression result:

For upper truncated data:

****

For lower truncated data:

****

For both upper and lower truncated data:

****

**Note that: change the upper and lower bound according to the problem.**

For example, refer the slides with the example having lower bound 0 and upper bound 2000.

**Question3: given the values of independent variable, calculate the marginal effect of a certain independent variable?**

**Answer:**

The same as OLS. Just the coefficient of that variable.

**Event history analysis**

**Question1: how to calculate the hazard ratio given a dataset?**

**Answer:** refer the problem set 6.

**Question 2: understand the 5 problem in test exercise and problem set6.**