Estimating compound expectation in a regression framework with the new cereg command

2015 Nordic and Baltic Stata Users Group meeting

Celia García-Pareja Matteo Bottai

Unit of Biostatistics, IMM, KI

September 4th, 2015

Outline

Introduction

- Statistics is about summarizing information contained in observed data.
- The most informative, representative and precise the summary is, the better.
- Typical summary measures to provide are, for example, the sample mean and the quantiles.

Question

Which summary measure is more "suitable"? How precise is the information it provides?

Motivating example I

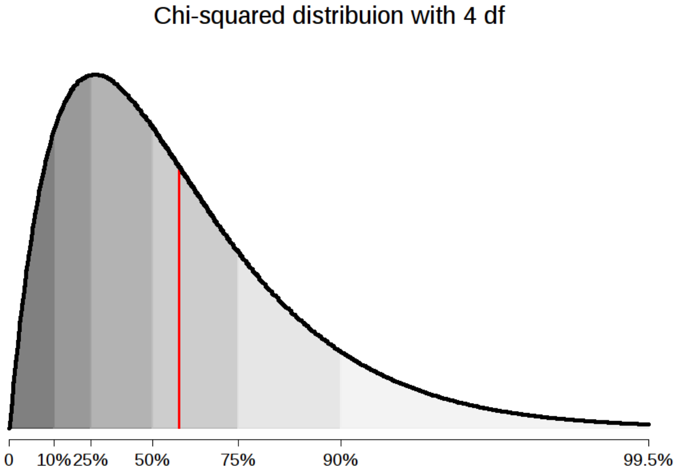
- Simulated data on 450 observations drawn from a chi square with 4 d.f.

```
. sqreg c, q(0.1 0.25 0.5 0.75 0.9) reps(200)
------------------------------------------------------------------------------
           t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q10          |
       _cons |   1.166131   .0805198    14.48   0.000     1.007888    1.324373
-------------+----------------------------------------------------------------
q25          |
       _cons |   2.055943   .1183237    17.38   0.000     1.823406     2.28848
-------------+----------------------------------------------------------------
q50          |
       _cons |   3.603483   .1191959    30.23   0.000     3.369232    3.837734
-------------+----------------------------------------------------------------
q75          |
       _cons |   5.423563   .1963908    27.62   0.000     5.037604    5.809522
-------------+----------------------------------------------------------------
q90          |
       _cons |   7.642251    .343475    22.25   0.000     6.967232    8.317269
------------------------------------------------------------------------------
. regress t
------------------------------------------------------------------------------
           t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   4.077057   .1276203    31.95   0.000     3.826249    4.327864
------------------------------------------------------------------------------
```

Remarks

- Quantiles provide information about the whole distribution whereas the mean just refers to the mass center.
- Estimation of low quantiles is more precise than that of high quantiles.
- Inference on the mean is better than in high quantiles but worse than in low quantiles.

Motivating example II



Mean vs quantiles

The mean...

- Summarizes the data in a single number and it is easy to interpret.
- Its inference is extremely sensitive to the presence of outliers.
- Is informative just in case there is little variability in the data.

Quantiles...

- Provide a detailed picture of the underlying statistical distribution.
- Can be estimated with high precision in regions with high density of data.
- Provide information about single points of the distribution.

Mean vs quantiles

The mean...

- Summarizes the data in a single number and it is easy to interpret.
- Its inference is extremely sensitive to the presence of outliers.
- Is informative just in case there is little variability in the data.

Quantiles...

- Provide a detailed picture of the underlying statistical distribution.
- Can be estimated with high precision in regions with high density of data.
- Provide information about single points of the distribution.

Proposal

- Combine both summary measures providing a bridge between mean and quantiles.

Compound Expectation I

The conditional expectation of Y can be written in terms of its quantile function as

$$\mu(\mathbf{x}) = \mathsf{E}[Y|\mathbf{x}] = \int_{-\infty}^{\infty} y \mathrm{d}F_Y(y|\mathbf{x}) = \int_0^1 Q_Y(p|\mathbf{x})\mathrm{d}p.$$

Given a set of specified proportions $\{0, \lambda_1, \lambda_2, \ldots, \lambda_{K-1}, 1\}$, we split $\mu(\mathbf{x})$ into components

$$\mu(\mathbf{x}) = \int_0^1 Q_Y(p|\mathbf{x})\mathrm{d}p = \int_0^{\lambda_1} Q_Y(p|\mathbf{x})\mathrm{d}p + \int_{\lambda_1}^{\lambda_2} Q_Y(p|\mathbf{x})\mathrm{d}p + \ldots + \int_{\lambda_{K-1}}^{1} Q_Y(p|\mathbf{x})\mathrm{d}p.$$

Each component $\mu_k(\mathbf{x}) = \int_{\lambda_{k-1}}^{\lambda_k} Q_Y(p|\mathbf{x})\mathrm{d}p$ measures the contribution of a fraction of the population to $\mu(\mathbf{x})$.

Compound Expectation II

We might also calculate the expectation of every k-th component

$$\overline{\mu}_k(\mathbf{x}) = \frac{\mu_k(\mathbf{x})}{\lambda_k - \lambda_{k-1}}.$$

$\mu(\mathbf{x})$ can be then expressed as a weighted average of these expectations

$$\mu(\mathbf{x}) = \sum_{k=1}^{K} (\lambda_k - \lambda_{k-1}) \overline{\mu}_k(\mathbf{x}).$$

Special interest application settings

Distributions with large variability:

- The mean is not representative and the quantiles might be insufficient.

Censored data:

- Lack of information in the upper tail: the components can be computed up to the last observed quantile.

Estimation of the CE

Suppose that the conditional quantile function can be estimated as a linear combination of a set of covariates of interest:

$$\widehat{Q}_Y(p|\mathbf{x}) = \hat{\beta}_{0p} + \hat{\beta}_{1p}x_1 + \ldots + \hat{\beta}_{sp}x_s = \sum_{j=0}^{s} \hat{\beta}_{jp}x_j.$$

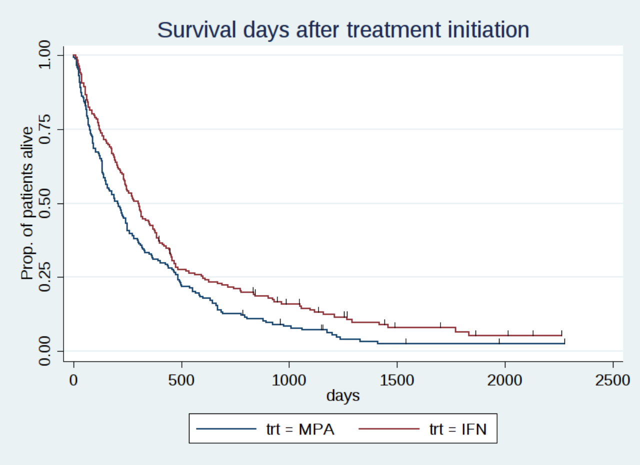
Every component $\hat{\mu}_k(\mathbf{x})$ can be expressed as

$$\hat{\mu}_k(\mathbf{x}) = \int_{\lambda_{k-1}}^{\lambda_k} \widehat{Q}_Y(p|\mathbf{x})\mathrm{d}p = \int_{\lambda_{k-1}}^{\lambda_k} \sum_{j=0}^{s} \hat{\beta}_{jp}x_j\mathrm{d}p = \sum_{j=0}^{s} \left(\int_{\lambda_{k-1}}^{\lambda_k} \hat{\beta}_{jp}\mathrm{d}p \right) x_j = \sum_{j=0}^{s} \hat{B}_{jk}x_j.$$

- Therefore, $\hat{B}_{jk}$ is the effect of the j-th covariate in the k-th component.

Data Example

- 347 patients with metastatic renal carcinoma.
- Patients randomly assigned to either subcutaneous interferon-α (IFN) or oral medroxyprogesterone acetate (MPA).
- After the total follow-up time, 322 patients had died and the censoring rate was 7.2%.



Results I: components vs the overall mean

```
. cereg days trt, f(died) c(0.01 0.25 0.5 0.6 0.7 0.85 0.99) reps(50)

Compound Expectation regression                    No. of subjects  =       347
                                                   No. of failures  =       322
-------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
q1_25        |
         trt |   4.782368   12.34894     0.39   0.699     -19.4211    28.98584
       _cons |   10.14267   21.43635     0.47   0.636     -31.8718    52.15713
-------------+-----------------------------------------------------------------
q25_50       |
         trt |   18.51029   11.62389     1.59   0.111    -4.272117    41.29269
       _cons |   33.18598   16.40143     2.02   0.043     1.039758     65.3322
-------------+-----------------------------------------------------------------
q50_60       |
         trt |   10.10135   4.666368     2.16   0.030     .9554328    19.24726
       _cons |   23.39756   3.992694     5.86   0.000     15.57202    31.22309
-------------+-----------------------------------------------------------------
q60_70       |
         trt |   9.886872   4.607069     2.15   0.032      .857183    18.91656
       _cons |   32.78498   2.640678    12.42   0.000     27.60935    37.96062
-------------+-----------------------------------------------------------------
q70_85       |
         trt |   25.59985    10.4843     2.44   0.015     5.051003    46.14869
       _cons |   79.52381   8.404415     9.46   0.000     63.05146    95.99616
-------------+-----------------------------------------------------------------
q85_99       |
         trt |   56.32177   20.61547     2.73   0.006     15.91619    96.72734
       _cons |   145.1348   31.29068     4.64   0.000     83.80616    206.4634
-------------------------------------------------------------------------------
```

- The overall life expectancy after treatment initiation for those who had MPA was 324.17 days and for those who had IFN was 449.34 days (125.20 days of difference).

Results II: life expectancy in portions of the population

```
. cereg days trt, f(died) c(0.01 0.25 0.5 0.6 0.7 0.85 0.99) reps(50) means

Compound Expectation regression                   No. of subjects  =       347
                                                  No. of faiures   =       322
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
M1_25        |
         trt |   19.92653   49.75396     0.40   0.689    -77.58945    117.4425
       _cons |   42.26111   105.6167     0.40   0.689    -164.7439    249.2661
-------------+----------------------------------------------------------------
M25_50       |
         trt |   74.04115    38.8984     1.90   0.057    -2.198312    150.2806
       _cons |   132.7439   80.56393     1.65   0.099    -25.15848    290.6463
-------------+----------------------------------------------------------------
M50_60       |
         trt |   101.0135   36.90469     2.74   0.006      28.6816    173.3453
       _cons |   233.9756   52.19961     4.48   0.000     131.6662    336.2849
-------------+----------------------------------------------------------------
M60_70       |
         trt |   98.86872   37.31695     2.65   0.008     25.72885    172.0086
       _cons |   327.8498   34.85833     9.41   0.000     259.5288    396.1709
-------------+----------------------------------------------------------------
M70_85       |
         trt |   170.6656   83.89101     2.03   0.042     6.242284     335.089
       _cons |   530.1587   61.65655     8.60   0.000     409.3141    651.0033
-------------+----------------------------------------------------------------
M85_99       |
         trt |   402.2983   153.0327     2.63   0.009     102.3598    702.2368
       _cons |   1036.677   275.0323     3.77   0.000     497.6235     1575.73
------------------------------------------------------------------------------
```

Conclusions

- The compound expectation is a suitable summary measure in any scenario.
- It can be used in a regression framework and thus, it provides information about the effect of a set of covariates of interest.
- It represents a useful tool for groups comparison.
- In the presence of censoring, it can be computed up to the last observed quantile, avoiding extrapolation.

Further work:

- Optimize the components' width for every specific case, in order to achieve better inferences.

Thank you for your attention.