Lecture 29: Poisson Regression

Pratheepa Jeganathan

12/04/2019

Recap

- What is a regression model?
- Descriptive statistics – graphical
- Descriptive statistics – numerical
- Inference about a population mean
- Difference between two population means
- Some tips on R
- Simple linear regression (covariance, correlation, estimation, geometry of least squares)
  - Inference on simple linear regression model
  - Goodness of fit of regression: analysis of variance.
  - F-statistics.
  - Residuals.
  - Diagnostic plots for simple linear regression (graphical methods).

Recap

- Multiple linear regression
  - Specifying the model.
  - Fitting the model: least squares.
  - Interpretation of the coefficients.
  - Matrix formulation of multiple linear regression
  - Inference for multiple linear regression
    - T-statistics revisited.
    - More F statistics.
    - Tests involving more than one β.
- Diagnostics – more on graphical methods and numerical methods
  - Different types of residuals
  - Influence
  - Outlier detection
  - Multiple comparison (Bonferroni correction)
  - Residual plots:
    - partial regression (added variable) plot,
    - partial residual (residual plus component) plot.

Recap

- Adding qualitative predictors
  - Qualitative variables as predictors to the regression model.
  - Adding interactions to the linear regression model.
  - Testing for equality of regression relationship in various subsets of a population
- ANOVA
  - All qualitative predictors.
  - One-way layout
  - Two-way layout
- Transformation
  - Achieving linearity
  - Stabilize variance
  - Weighted least squares
- Correlated Errors
  - Generalized least squares
- Bootstrapping linear regression
- Selection

Recap

- Colliniarity
  - Bias-variance tradeoff
  - Penalized Regression
    - Ridge
    - LASSO
    - Elastic net
- Generalized linear regression
  - Logistic regression
  - Probit model

Outline (Poisson regression)

- Contingency tables.
- Log-linear regression.
- Log-linear regression as a generalized linear model.

Count data

- Example: Afterlife

Men and women were asked whether they believed in the after life (1991 General Social Survey).

Y	N or U	Total	
M	435	147	582
F	375	134	509
Total	810	281	1091

-

Question: is belief in the afterlife independent of gender?

Poisson counts

- Definition
  - A random variable Y is a Poisson random variable with parameter λ if

$$P(Y = j) = e^{-\lambda} \frac{\lambda^j}{j!}, \qquad \forall j \geq 0.$$

  - Some simple calculations show that $E(Y) = \text{Var}(Y) = \lambda$.
  - Poisson models for counts are analogous to Gaussian for continuous outcomes – they appear in many common models.

Contingency table

- Model: $Y_{ij} \sim Poisson(\lambda_{ij})$.
- Null (independence):
  $H_0 : \lambda_{ij} = \delta \cdot \alpha_i \cdot \beta_j, \sum_i \alpha_i = 1, \sum_j \beta_j = 1.$
- Alternative: $H_a : \lambda_{ij} \in \mathbb{R}^+$
- Test statistic: Pearson's X^2 : $X^2 = \sum_{ij} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0}{\approx} \chi^2_1$
- Here E_{ij} is the estimated expected value under independence.
- Why 1 df ? Independence model has 5 parameters, two constraints = 3 df. Unrestricted has 4 parameters.
- This is actually a *regression model* for the count data.

```r
Y = c(435,147,375,134)
S = factor(c('M','M','F','F'))
B = factor(c('Y','N','Y','N'))

N = sum(Y)
piS = c((435+147)/N,(375+134)/N)
piB = c((435+375)/N,(147+134)/N)

E = N*c(piS[1]*piB[1], piS[1]*piB[2], piS[2]*piB[1], piS[2
# Pearson's X^2
X2 = sum((Y - E)^2/E)
c(X2, 1-pchisq(X2,1))
```

```
## [1] 0.1620840 0.6872451
```

- The independence test is called `chisq.test` in R. Depending on whether one corrects or not, we get the X^2 or a corrected version.

```r
chisq.test(matrix(Y,2,2), correct=FALSE)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  matrix(Y, 2, 2)
## X-squared = 0.16208, df = 1, p-value = 0.6872
```

```r
chisq.test(matrix(Y,2,2))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity corre
##
## data:  matrix(Y, 2, 2)
## X-squared = 0.11103, df = 1, p-value = 0.739
```

Contingency table as regression model

- Under independence $\log(E(Y_{ij})) = \log \lambda_{ij} = \log \delta + \log \alpha_i + \log \beta_j$
- OR, the model has a *log link*.
- What about the variance? Because of Poisson assumption $Var(Y_{ij}) = E(Y_{ij})$
- OR, the *variance function* is $V(\mu) = \mu$.

The goodness of fit test can also be found using a `glm`.

```r
summary(glm(Y ~ S + B, family=poisson()))
```

```
##
## Call:
## glm(formula = Y ~ S + B, family = poisson())
##
## Deviance Residuals:
##       1        2        3        4
##  0.1394  -0.2377  -0.1494   0.2524
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.87595    0.06787  71.839   <2e-16 ***
## SM           0.13402    0.06069   2.208   0.0272 *
## BY           1.05868    0.06923  15.291   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## (Dispersion parameter for poisson family taken to be 1)
```

- This model has the same fitted values as we had computed by hand above.

```
fitted(glm(Y ~ S+B, family=poisson()))
```

```
##       1       2       3       4
## 432.099 149.901 377.901 131.099
```

```
E
```

```
## [1] 432.099 149.901 377.901 131.099
```

- Here is the deviance test statistic.
- It is numerically close, but not identical to Pearson's X^2 for this data.

```
DEV = sum(2*(Y*log(Y/E)+Y-E))
c(X2, DEV)
```

```
## [1] 0.1620840 0.1619951
```

Contingency table ($k \times m$)

- Suppose we had k categories on one axis, m on the other (i.e. previous example $k = m = 2$). We call this as $k \times m$ contingency table.
- Independence model (H_0): $\log(E(Y_{ij})) = \log \lambda_{ij} = \log \delta + \log \alpha_i + \log \beta_j$
- Test for independence: Pearson's

$$X^2 = \sum_{ij} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0}{\approx} \chi^2_{(k-1)(m-1)}$$

- Alternative test statistic $G = 2\sum_{ij} Y_{ij} \log\left(\frac{Y_{ij}}{E_{ij}}\right)$

Independence tests

- Unlike in other cases, in this case the *full model* has as many parameters as observations (i.e. it's saturated).
- This test is known as a *goodness of fit* test.
- It tests: "how well does the independence model fit this data"?
- Unlike other tests we've seen, the deviance is the test statistic, not a difference of deviance.

Lumber company example

- Y : number of customers visting store from region;
- X_1 : number of housing units in region;
- X_2 : average household income;
- X_3 : average housing unit age in region;
- X_4 : distance to nearest competitor;
- X_5 : distance to store in miles.

Poisson (log-linear) regression model

- Given observations and covariates $Y_i, X_{ij}, 1 \leq i \leq n, 1 \leq j \leq p$.
- Model:

$$Y_i \sim Poisson\left(\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)\right)$$

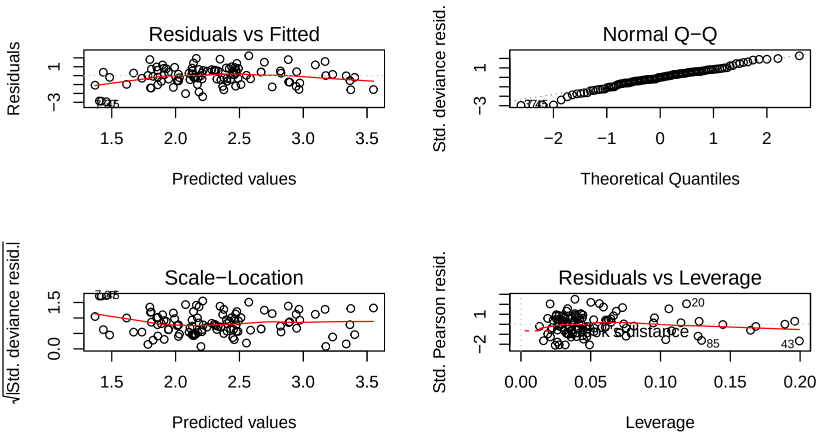
- Poisson assumption implies the variance function is $V(\mu) = \mu$.

```r
url = 'http://stats191.stanford.edu/data/lumber.table'
lumber.table = read.table(url, header=T)
lumber.glm = glm(Customers ~ Housing + Income +
                   Age + Competitor + Store,
                 family=poisson(), data=lumber.table)
```

```r
summary(lumber.glm)
```

```
##
## Call:
## glm(formula = Customers ~ Housing + Income + Age + Compe
##     Store, family = poisson(), data = lumber.table)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.93195  -0.58868  -0.00009   0.59269   2.23441
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.942e+00  2.072e-01  14.198  < 2e-16 ***
## Housing      6.058e-04  1.421e-04   4.262 2.02e-05 ***
## Income      -1.169e-05  2.112e-06  -5.534 3.13e-08 ***
## Age         -3.726e-03  1.782e-03  -2.091   0.0365 *
## Competitor   1.684e-01  2.577e-02   6.534 6.39e-11 ***
## Store       -1.288e-01  1.620e-02  -7.948 1.89e-15 ***
## ---
```

```r
par(mfrow=c(2,2))
plot(lumber.glm)
```



Interpretation of coefficients

- The log-linear model means covariates have *multiplicative* effect.
- Log-linear model model: $\frac{E(Y|\dots,X_j=x_j+h,\dots)}{E(Y|\dots,X_j=x_j,\dots)} = e^{h\cdot\beta_j}$
- So, one unit increase in variable j results in e^{β_j} (multiplicative) increase the expected count, all other parameters being equal.

Generalized linear models

- Logistic model:
  $\text{logit}(\pi(X)) = \beta_0 + \sum_j \beta_j X_j \qquad V(\pi) = \pi(1-\pi)$
- Poisson log-linear model:
  $\log(\mu(X)) = \beta_0 + \sum_j \beta_j X_j, \qquad V(\mu) = \mu$
- These are the ingredients to a GLM

Deviance tests

- To test $H_0 : \mathcal{M} = \mathcal{M}_R$ vs. $H_a : \mathcal{M} = \mathcal{M}_F$, we use

$$DEV(\mathcal{M}_R) - DEV(\mathcal{M}_F) \sim \chi^2_{df_R - df_F}$$

- In contingency example $\mathcal{M}_R$ is the independence model

$$\log(E(Y_{ij})) = \log \delta + \log \alpha_i + \log \beta_j$$

with $\mathcal{M}_F$ being the *saturated model*: no constraints on $E(Y_{ij})$.

```r
lumber.R.glm = glm(Customers ~ Housing + Income + Age,
                   family=poisson, data=lumber.table)
anova(lumber.R.glm, lumber.glm)
```

```
## Analysis of Deviance Table
##
## Model 1: Customers ~ Housing + Income + Age
## Model 2: Customers ~ Housing + Income + Age + Competito
##   Resid. Df Resid. Dev Df Deviance
## 1       106     378.43
## 2       104     114.99  2   263.45
```

```r
pchisq(263.45, 2, lower=FALSE, log=TRUE)
```

```
## [1] -131.725
```

```r
1 - pchisq(263.45, 2)
```

```
## [1] 0
```

Model selection

- As it is a likelihood model, `step` can also be used for model selection.

```
step(lumber.glm)
```

```
## Start:  AIC=571.02
## Customers ~ Housing + Income + Age + Competitor + Store
##
##              Df Deviance    AIC
## <none>            114.98 571.02
## - Age         1   119.36 573.40
## - Housing     1   133.19 587.23
## - Income      1   146.78 600.82
## - Competitor  1   156.65 610.68
## - Store       1   182.49 636.52

##
## Call:  glm(formula = Customers ~ Housing + Income + Age
##     Store, family = poisson(), data = lumber.table)
```

```r
step(glm(Customers ~ 1, data=lumber.table, family=poisson()
```

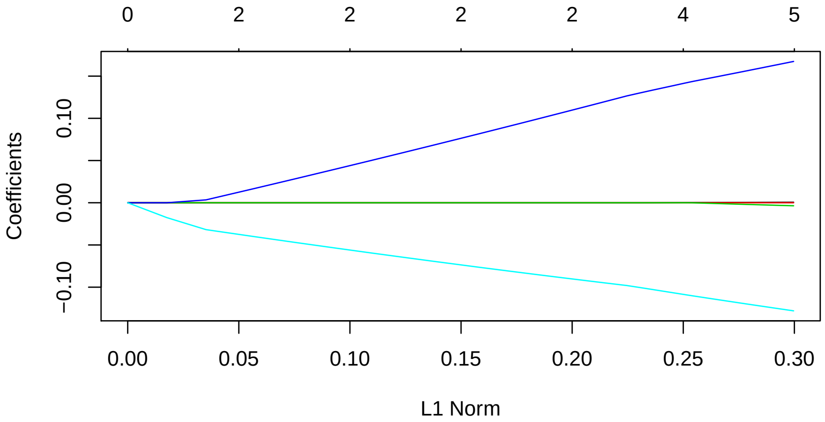
```
## Start:  AIC=868.26
## Customers ~ 1
##
##                Df Deviance    AIC
## + Store         1   184.41 632.45
## + Competitor    1   201.90 649.93
## + Housing       1   379.56 827.60
## + Income        1   399.15 847.19
## <none>              422.22 868.26
## + Age           1   422.20 870.24
##
## Step:  AIC=632.45
## Customers ~ Store
##
##                Df Deviance    AIC
## + Competitor    1   149.33 599.37
## + Income        1   177.45 627.49
## + Housing       1   181.29 631.33
```

LASSO

- LASSO also applicable

```r
library(glmnet)
X = model.matrix(lumber.glm)[,-1]
Y = lumber.table$Customers
G = glmnet(X, Y, family='poisson')
```

```r
plot(G)
```


0 2 2 2 2 4 5
Coefficients
0.10
0.00
−0.10
0.00 0.05 0.10 0.15 0.20 0.25 0.30
L1 Norm

Reference

- **CH** Chapter
- Lecture notes of Jonathan Taylor .