# Regression Models for Count Data with R

By

Olga Korosteleva, CSULB

# Linear Regression: Review

- *Linear Regression* model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

where $\varepsilon$ is a $N(0, \sigma^2\ )$ *random error*.

- Equivalently, $y$ is a normally distribution random variable with mean

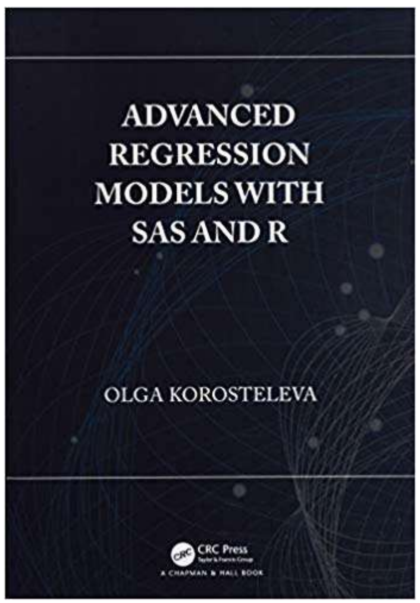
$$Ey = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$ and variance $\sigma^2$.

- Parameters are $\beta_0, \beta_1, \ldots \beta_k$, and $\sigma^2$.
- Fitted model is $\hat{E}y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$.

# We use Linear regression for:

- Prediction $y^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0$.
- Interpretation of fitted coefficients:
  - If $x_1$ is continuous, since $\hat{\beta}_1 = \hat{E}y|_{x_1+1} - \hat{E}y|_{x_1}$, as $x_1$ increases by one unit, the estimated mean of $y$ changes by $\hat{\beta}_1$.
  - If $x_1$ is 0 -1 variable, since $\hat{\beta}_1 = \hat{E}y|_{x_1=1} - \hat{E}y|_{x_1=0}$, the difference of the estimated means of $y$ for $x_1 = 1$ and $x_1 = 0$ is $\hat{\beta}_1$.

# Other regression models




## Idea:

- Model y as having certain distribution defined by the setting.

- Model mean Ey as a certain function of linear regression $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$: $g(Ey) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ where $g(.)$ is called a *link* function.

- Predict as $y^0 = g^{-1}(\hat{\beta}^0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0)$.

- Interpret as
  - If $x_1$ is continuous, $\hat{\beta}_1 = g^{-1}(\hat{E}y)|_{x_1+1} - g^{-1}(\hat{E}y)|_{x_1}$.
  - If $x_1$ is 0 -1 variable, $\hat{\beta}_1 = g^{-1}(\hat{E}y)|_{x_1=1} - g^{-1}(\hat{E}y)|_{x_1=0}$.

- My recently published book “Advanced Regression Models with SAS and R” discusses 60 different regressions.

# QUICK EXAMPLE: BINARY LOGISTIC REGRESSION

- Suppose $y = 1$ with probability $\pi = P(y = 1)$, and 0, otherwise. Then $y$ has a *Bernoulli* (or *binary*) distribution with mean $Ey = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi = P(y = 1)$.

- This mean lies between 0 and 1, so we can relate it to the linear regression via the *logistic* function $\frac{\exp(x)}{1+\exp(x)}$:

$$\pi = P(y = 1) = \frac{Exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + Exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

- *Binary logistic regression* models the mean of $y$ through the *logit* link function $g(x) = ln \frac{x}{1-x}$:

$$ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

## BINARY LOGISTIC REGRESSION EXAMPLE CONTINUES

- Fitted model is $\hat{\pi} = \hat{P}(y=1) = \frac{Exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k)}{1+Exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k)}$. Equivalently, the fitted *odds in favor of* $y=1$ can be written as

$$\frac{\hat{\pi}}{1-\hat{\pi}} = Exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k).$$

- Interpretation:
  - If $x_1$ is continuous, as $x_1$ increases by one unit, the estimated odds change by $\frac{\widehat{odds}_{x_1+1}-\widehat{odd}\ _{x_1}}{\widehat{odds}_{x_1}} \cdot 100\% = \left(Exp\left(\hat{\beta}_1\right)-1\right) \cdot 100\%$.
  - If $x_1$ is 0 -1 variable, the ratio of estimated odds for $x_1 = 1$ and $x_1$ =0 is $\frac{\widehat{odds}_{x_1=1}}{\widehat{odds}_{x_1=0}} \cdot 100\% = Exp\left(\hat{\beta}_1\right) \cdot 100\%$.

# POISSON MODEL for count data

- *Count data* means that $y$ assumes values 0, 1, 2, etc.

Suppose 0 is quite a common value and so is 1; 2 is more rare; 3, 4, 5 are even less frequent; 6, 7, 8 are very infrequent. Overall, we can model $y$ as having a Poisson distribution with mean $\lambda$ and probability mass function

$$P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}, y = 0, 1, 2, \dots .$$

- We know that $\lambda$ must be positive, thus we can model

$$\lambda = Ey = Exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).$$

- *Poisson regression* models $y$ as having Poisson distribution, and the mean relating to the linear regression through the *log* link function

$$\ln(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

# POISSON MODEL CONTINUES

- Fitted model is $\hat{\lambda} = \hat{E}y = Exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\right)$.
- Prediction: $y^0 = Exp\ (\hat{\beta}^0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0)$.
- Interpretation of fitted coefficients:
  - If $x_1$ is continuous, as $x_1$ increases by one unit, the estimated mean changes by $\frac{\widehat{\lambda}_{x_1+1} - \widehat{\lambda}_{x_1}}{\widehat{\lambda}_{x_1}} \cdot 100\% = \left(Exp\left(\hat{\beta}_1\right) - 1\right) \cdot 100\%$.
  - If $x_1$ is 0 -1 variable, the ratio of estimated means for $x_1 = 1$ and $x_1$ =0 is $\frac{\widehat{\lambda}_{x_1=1}}{\widehat{\lambda}_{x_1=0}} \cdot 100\% = Exp\left(\hat{\beta}_1\right) \cdot 100\%$.

## EXAMPLE: POISSON REGRESSION

- Number of days of hospital stay was recorded for 45 patients along with their gender, age, and history of chronical cardiac illness.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | F | 31 | yes | 0 | F | 28 | no | 0 | M | 52 | yes |
| 1 | M | 72 | yes | 0 | F | 29 | no | 0 | F | 30 | no |
| 1 | M | 74 | no | 2 | M | 30 | yes | 2 | F | 72 | no |
| 1 | M | 58 | no | 2 | F | 28 | no | 2 | F | 65 | no |
| 2 | M | 65 | no | 1 | M | 52 | no | 4 | M | 51 | no |
| 2 | F | 63 | no | 0 | F | 31 | no | 1 | F | 47 | yes |
| 1 | M | 49 | no | 2 | M | 71 | yes | 2 | M | 48 | no |
| 2 | F | 47 | no | 0 | F | 31 | no | 3 | M | 44 | yes |
| 3 | M | 44 | no | 3 | M | 54 | yes | 4 | F | 72 | yes |
| 4 | M | 56 | yes | 3 | F | 73 | yes | 1 | F | 46 | no |
| 3 | M | 58 | no | 4 | M | 70 | yes | 2 | M | 36 | no |
| 1 | M | 50 | no | 1 | M | 59 | no | 0 | M | 52 | no |
| 6 | M | 68 | yes | 2 | F | 41 | no | 1 | M | 31 | yes |
| 1 | M | 69 | no | 3 | M | 73 | no | 3 | F | 77 | yes |
| 2 | F | 54 | no | 4 | M | 69 | yes | 5 | M | 68 | yes |

POISSON REGRESSION EXAMPLE CONTINUES

- We fit Poisson regression model using R:

```
hospitalstay.data<-read.csv(file= "./data.csv" , header=TRUE, sep= ",")
summary(fitted.model<- glm(days ~ gender + age + illness,
data=hospitalstay.data, family=poisson(link=log)))
```

- The fitted model is

$\hat{\lambda} = Exp(-0.8263+0.2264*\text{male}+0.0205*\text{age}+0.4477*\text{illness})$.

- Prediction: The predicted length of stay for a 55-year old male with no chronic cardiac illness is computed as

$$y^0 = Exp(-0.8263+0.2264+0.0205*55) = 1.6949.$$

- Interpretation of estimated regression coefficients:

- (gender) Estimated average length of hospital stay for males is exp{0.2264} · 100% = 125.41% of that for females.
- (age) For a one-year increase in patient's age, the estimated average number of days of hospital stay increases by (exp{0.0205}–1)·100% = 2.07%.
- (illness) The estimated average number of days of hospital stay for patients with a chronic cardiac illness is exp{0.4477}·100% = 156.47% of that for patients without it.

# ZERO-TRUNCATED POISSON MODEL FOR COUNT DATA

- Suppose $y$ follows a Poisson distribution but no zeros are observed.
- Then $y$ can be modeled via a *zero-truncated Poisson regression* where

  the distribution of $y$ is $P(Y = y) = \frac{\lambda^y}{y!} \cdot \frac{e^{-\lambda}}{1-e^{-\lambda}}, y = 1, 2, \dots,$

  with $\lambda = Exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$.
- Fitted *model is* $\hat{\lambda} = Exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\right)$.
- Prediction $y^0 = \hat{E}y = \frac{\hat{\lambda}}{1-E \quad (-\hat{\lambda})} = \frac{Exp\left(\hat{\beta}^0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0\right)}{1-Exp\left(-Exp\left(\hat{\beta}^0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0\right)\right)}$.
- Interpretation of estimated regression coefficients is the same as in Poisson model.

# EXAMPLE: ZERO-TRUNCATED POISSON REGRESSION

- Suppose in the previous example, the data were reduced to the 38 patients who spent at least one day in the hospital. We run a zero-truncated Poisson model using R:

```
hospitalstay.data<-read.csv(file= “./data.csv” , header=TRUE, sep= “,” )
#eliminating zeros from the original data set
hospitaldays.data<-hospitalstay.data[which(hospitalstay.data$days!=0),]
install.packages("VGAM")
library(VGAM)
summary(fitted.model<- vglm(days ~ gender + age + illness,
data=hospitaldays.data, family=pospoisson()))
```

- The fitted model is
  $\hat{\lambda} = Exp$(−0.7041+0.2146∗male+0.01604∗age+0.5903∗illness).

- Prediction:

To predict the number of days of hospital stay for a 55-year old male without a chronic cardiac illness, we calculate

$$y^0 = \frac{\exp\left\{-0.7041 + 0.2146 + 0.01604 * 55\right\}}{1 - \exp\left\{-\exp\left\{-0.7041 + 0.2146 + 0.01604 * 55\right\}\right\}} = 1.9169.$$

ZERO-TRUNCATED POISSON EXAMPLE CONTINUES

## ❑ Interpretation of estimated regression coefficients:

- (gender) Estimated average length of hospital stay for males is $\exp\{0.2146\} \cdot 100\% = 123.94\%$ of that for females.
- (age) For a one-year increase in patient's age, the estimated average number of days of hospital stay increases by $(\exp\{0.01604\}-1)\cdot 100\% = 1.62\%$.
- (illness) The estimated average number of days of hospital stay for patients with a chronic cardiac illness is $\exp\{0.5903\}\cdot 100\% = 180.45\%$ of that for patients without it.

# ZERO-INFLATED POISSON MODEL FOR COUNT DATA

- Suppose $y$ follows a Poisson distribution but too many zeros are observed. For example, suppose that one of the variables recorded during a health survey is the number of cigarettes the respondent smoked yesterday. Some respondents may have reported zero number of cigarettes smoked because they either do not smoke at all (*structural zero*), or they happened not to smoke a single cigarette that day (*chance zero*).
- Then $y$ can be modeled via a *zero-inflated Poisson (ZIP) regression* where the distribution of $y$ is

$$\mathbb{P}(Y = y) = \begin{cases} \pi + (1-\pi)\exp\{-\lambda\}, & \text{if } y = 0, \\ (1-\pi)\,\dfrac{\lambda^y \exp\{-\lambda\}}{y!}, & \text{if } y = 1, 2, \ldots, \end{cases}$$

where

$$\pi = \frac{\exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m\}},$$

and

$$\lambda = \exp\left\{\gamma_0 + \gamma_1 x_{m+1} + \cdots + \gamma_{k-m} x_k\right\}.$$

# ZIP MODEL (CONTINUED)

- The fitted model is

$$\widehat{\pi} = \frac{\exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_m x_m\}}{1 + \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_m x_m\}},$$

$$\widehat{\lambda} = \exp\left\{\widehat{\gamma}_0 + \widehat{\gamma}_1 x_{m+1} + \cdots + \widehat{\gamma}_{k-m} x_k\right\}.$$

- The fitted mean is

$$\hat{E}y = (1 - \hat{\pi}) \cdot \hat{\lambda} = \frac{\exp(\widehat{\gamma}_0 + \widehat{\gamma}_1 x_{m+1} + \cdots + \widehat{\gamma}_{k-m} x_k)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_m x_m)}.$$

- Prediction:

$$y^0 = \frac{\exp(\widehat{\gamma}_0 + \widehat{\gamma}_1 x_{m+1}^0 + \widehat{\gamma}_{k-m} x_k^0)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1^0 + \cdots + \widehat{\beta}_m x_m^0)}.$$

# ZIP MODEL (CONTINUED)

- Interpretation of estimated regression coefficients:
  - Probability of structural zero $\pi$ is modeled as in the binary logistic regression, thus, estimated beta coefficients are interpreted in terms of estimated odds.
  - The mean of $y$ is $Ey = (1-\pi)\cdot\lambda,$ and since we assume $x$ variables are non-overlapping in $\pi$ and $\lambda$, interpretation of gamma coefficients in $\lambda$ is the same as in Poisson regression model.
- Note that it is possible to use the same $x$ variables in the regression parts of $\pi$ and $\lambda$, but the estimates of the regression coefficients won't be easily interpretable. Can be useful for prediction.

# EXAMPLE: ZERO-INFLATED POISSON REGRESSION

- A health survey was been administered to a random sample of 40 people aged between 25 and 50. Their gender, self-reported health condition (excellent or good), age, and the number of cigarettes they smoked yesterday were recorded. The data are:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M | good | 34 | 3 | F | exclnt | 48 | 1 | M | exclnt | 26 | 0 | M | good | 39 | 0 |
| F | good | 27 | 1 | M | good | 28 | 5 | F | good | 44 | 1 | M | exclnt | 30 | 0 |
| F | exclnt | 26 | 0 | F | good | 38 | 2 | F | good | 40 | 1 | F | exclnt | 31 | 0 |
| M | good | 27 | 3 | F | exclnt | 34 | 1 | F | good | 36 | 2 | F | exclnt | 34 | 2 |
| F | exclnt | 39 | 0 | F | good | 42 | 1 | F | good | 48 | 4 | M | good | 32 | 5 |
| M | good | 47 | 2 | M | good | 29 | 3 | M | exclnt | 38 | 0 | F | good | 50 | 4 |
| M | good | 30 | 3 | M | good | 38 | 2 | M | good | 31 | 6 | F | exclnt | 33 | 0 |
| F | good | 28 | 0 | F | good | 42 | 3 | M | exclnt | 28 | 0 | M | good | 31 | 2 |
| F | exclnt | 31 | 0 | F | exclnt | 42 | 0 | F | good | 44 | 4 | F | good | 39 | 1 |

# ZERO-INFLATED POISSON REGRESSION EXAMPLE CONTINUES

- We fit a ZIP model with health condition modeling structural zeros and gender and age predicting the Poisson part:

```
smoking.data<-read.csv(file="./data.csv", header=TRUE, sep=",")
install.packages("pscl")
library(pscl)
#specifying reference category
health.rel<- relevel(smoking.data$health, ref="good")
#fitting zero-inflated Poisson model
summary(fitted.model<- zeroinfl(cigarettes ~ gender +
age|health.rel, data=smoking.data))
```

- The fitted model is

$$\widehat{\pi} = \frac{\exp\{-3.7950 + 4.9195 * excellent\_health\}}{1 + \exp\{-3.7950 + 4.9195 * excellent\_health\}},$$

$$\widehat{\lambda} = \exp\{-0.1381 + 0.0186 * age + 0.7268 * male\}.$$

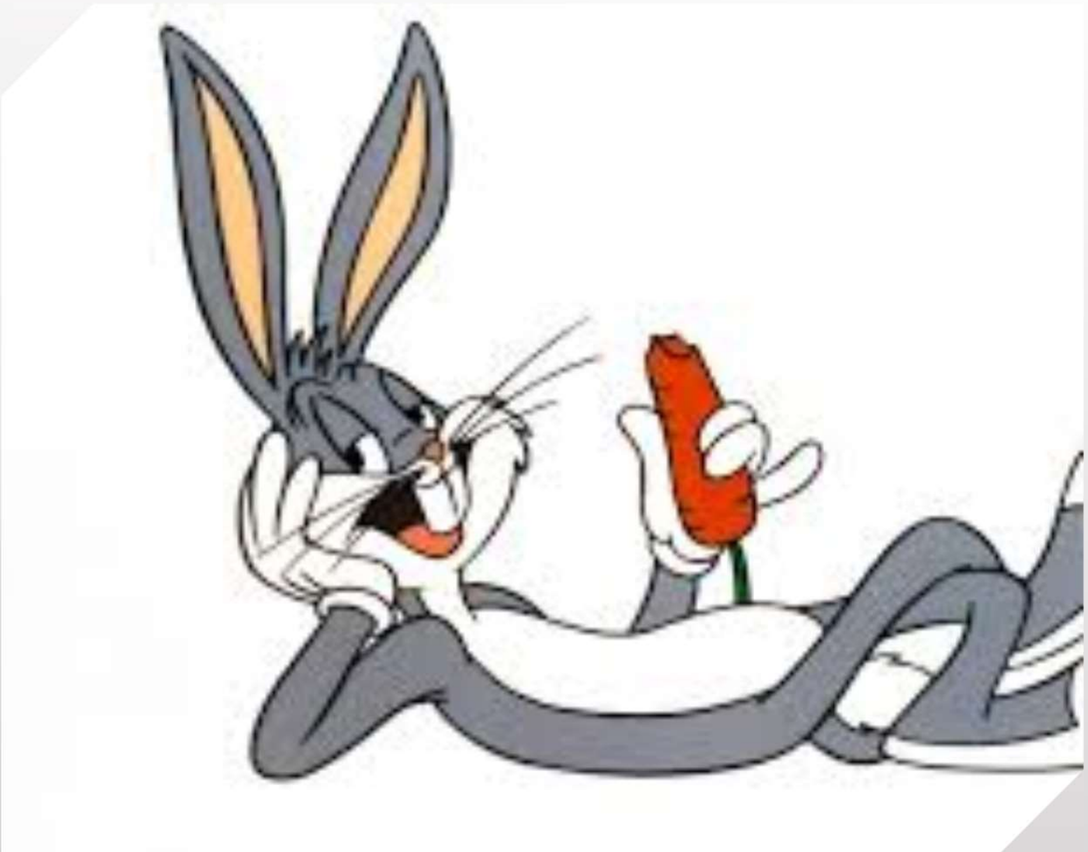
# ZIP REGRESSION EXAMPLE CONTINUES

- Prediction: The predicted number of cigarettes smoked per day by a 50-year old male who is in good health is found as

$$y^0 = \frac{\exp(-0.1381 + 0.0186 * 50 + 0.7268)}{1 + \exp(-3.7950)} = 4.4659.$$

- Interpretation of estimated regression coefficients:

- (health condition) The estimated odds of not smoking for people in excellent health is exp{4.9195}· 100% = 13,694.26% of those for people in good health.
- (age) As age increases by one year, the estimated average number of cigarettes smoked in a day increases by (exp{0.0186}-1) · 100% =1.88%.
- (gender) The estimated average number of cigarettes smoked in a day by men is exp{0.7268}·100% = 206.85% of that by women.



THANK YOU

PLEASE ATTEND MY PRESENTATION ON OCTOBER 5