Regression Review

Nambari Short Course

15 July 2019

Overview

- Review basics of regression model specification
- Define standard notation
- Illustrate with some examples

Goals of regression analysis

Notation

$$\begin{aligned} Y &= \text{dependent variable} \\ \boldsymbol{X} &= (X_1, \ldots, X_p) \\ &= \text{vector of covariates} \end{aligned}$$

Sample of data

$$(Y_1, \boldsymbol{X}_1), (Y_2, \boldsymbol{X}_2), \ldots, (Y_n, \boldsymbol{X}_n)$$

Objective

- Want to learn something about the relationship between $E(Y)$ and $\boldsymbol{X}$
- This class will be exclusively concerned with regression involving the *mean* of Y as a function of $\boldsymbol{X}$

Regression model specification

Some familiar regression models

$$E(Y \mid \boldsymbol{X}_i) = \boldsymbol{X}_i\beta$$

$$E(Y \mid \boldsymbol{X}_i) = \exp(\boldsymbol{X}_i\beta)$$

$$E(Y \mid \boldsymbol{X}_i) = \frac{\exp(\boldsymbol{X}_i\beta)}{1 + \exp(\boldsymbol{X}_i\beta)}$$

- What do each of these correspond to?
- What do they have in common?

Generalized linear model

Outcome Y having a specific distribution

Linear predictor
$\boldsymbol{X}_i\boldsymbol{\beta} = X_{1i}\beta_1 + X_{2i}\beta_2 + \cdots + X_{pi}\beta_p$

- This is an additive function of the covariates
- It is linear in the regression coefficients $\boldsymbol{\beta}$

Link function
$g\{E(Y)\}$ that transforms the mean of Y to an appropriate scale

Some example specifications

Linear model with normal errors

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \\
\mu_i &= \boldsymbol{X}_i\beta
\end{aligned}
$$

Poisson regression for count data

$$
\begin{aligned}
Y_i &\sim \mathcal{P}(\theta_i) \\
\log(\theta_i) &= \boldsymbol{X}_i\beta
\end{aligned}
$$

Poisson regression for count data with varying exposure

$$
\begin{aligned}
Y_i &\sim \mathcal{P}(u_i\theta_i) \\
\log(\theta_i) &= \boldsymbol{X}_i\beta + \log(u_i)
\end{aligned}
$$

Some example specifications

Logistic regression for binomial data

$$\begin{aligned} Y_i &\sim \text{Bin}(n_i, \pi_i) \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{X}_i\beta \end{aligned}$$

Probit regression for binomial data

$$\begin{aligned} Y_i &\sim \text{Bin}(n_i, \pi_i) \\ \Phi^{-1}(\pi_i) &= \boldsymbol{X}_i\beta \end{aligned}$$

Each of these specifications is written in a hierarchical format. What are the implied means and variances $E(Y \mid \boldsymbol{X})$ and $\text{var}(Y \mid \boldsymbol{X})$?

Example: Birthweight data

- Data were collected on mothers giving birth in the state of Georgia in the US.
- Dependent variable Y_i is weight in grams of baby
- Independent variable X_i is mother's age
- There are 5 births per woman; we will focus on first one for now

Model specification

We will write the model like this

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i,\, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 X_i \end{aligned}$$

Another equivalent way to write it is like this

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + e_i, \\ e_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

Models we will fit

Model 1	intercept only
Model 2	intercept and maternal age

```
> head(bwt)
   mid order weight age cid age_c group1 group2 weight0 age0 ageDiff
1   80     1   3175  18   1    -3      1      0    3175   -3       0
2   80     2   3572  21   2     0      1      0    3175   -3       3
3   80     3   3317  24   3     3      1      0    3175   -3       6
4   80     4   4281  26   4     5      1      0    3175   -3       8
5   80     5   3827  28   5     7      1      0    3175   -3      10
6   84     1   2892  14   6    -7      1      0    2892   -7       0
7   84     2   3204  16   7    -5      1      0    2892   -7       2
8   84     3   4253  20   8    -1      1      0    2892   -7       6
9   84     4   2948  22   9     1      1      0    2892   -7       8
10  84     5   3402  23  10     2      1      0    2892   -7       9
```

```
> head(bwt[bwt$order==1,], n=10)
   mid order weight age cid age_c group1 group2 weight0 age0 ageDiff
1   80     1   3175  18   1    -3      1      0    3175   -3       0
6   84     1   2892  14   6    -7      1      0    2892   -7       0
11  92     1   3260  18  11    -3      1      0    3260   -3       0
16 113     1   2900  24  16     3      1      0    2900    3       0
21 199     1   3118  15  21    -6      1      0    3118   -6       0
26 200     1   2892  19  26    -2      1      0    2892   -2       0
31 221     1   3260  19  31    -2      0      0    3260   -2       0
36 247     1   3090  17  36    -4      0      0    3090   -4       0
41 336     1   3714  16  41    -5      0      0    3714   -5       0
46 547     1   3969  19  46    -2      0      0    3969   -2       0
```

Model 0

```
> M0 = lm( weight ~ 1, data=bwt.first )
> display(M0)
lm(formula = weight ~ 1, data = bwt.first)
            coef.est coef.se
(Intercept) 3099.03    17.89
---
n = 878, k = 1
residual sd = 530.08, R-Squared = 0.00
```

Model 1

```
> M1 = lm( weight ~ 1 + age, data=bwt.first )
> display(M1)
lm(formula = weight ~ 1 + age, data = bwt.first)
            coef.est coef.se
(Intercept) 2518.52    92.57
age           32.48     5.09
---
n = 878, k = 2
residual sd = 518.45, R-Squared = 0.04
```

Fitted regression line