Intro to Poisson Regression

Reminders

- Exam 1 ove Friday
- No class on Friday

When, and when not, to use model selection

When to use model selection:

- many variables, want a subset
- we care more about prediction than inference
  - we don't know which variables are useful for predicting

Problems with model selection:

- resulting model might not be interpretable
- model selection doesn't fix violations of assumptions
- should not do inference with the same data you use for model selection

Possible solution: Data splitting

- use part of the data to select model
- use remainder to test hypotheses

Count variables

Data: Data on medical facilities and doctors from a sample of 53 different counties in the US. Variables include:

- `MDs`: the number of medical doctors in the county
- `Hospitals`: the number of hospitals in the county

countvariable
values 0, 1, 2, ...

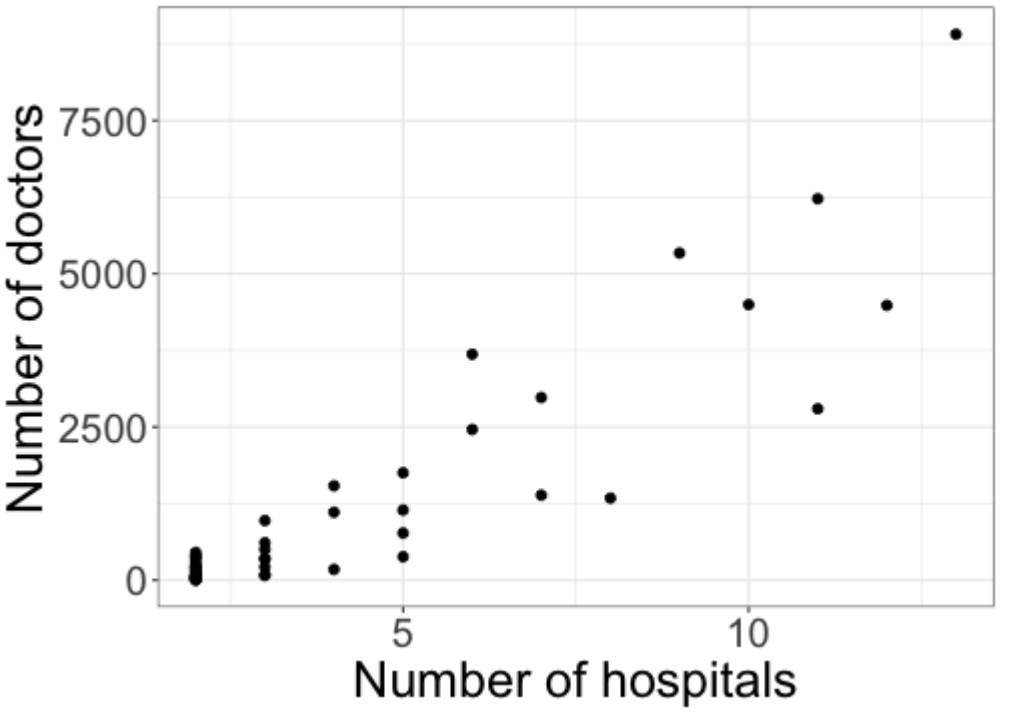
Research question: Can we model the relationship between the number of hospitals and the number of doctors?

y_i = #MDs

$y_i \sim N(\mu_i, \sigma^2)$

$\mu_i = \beta_0 + \beta_1 \text{Hospitals}_i$

Plotting the data



assumes constant variance σ^2!

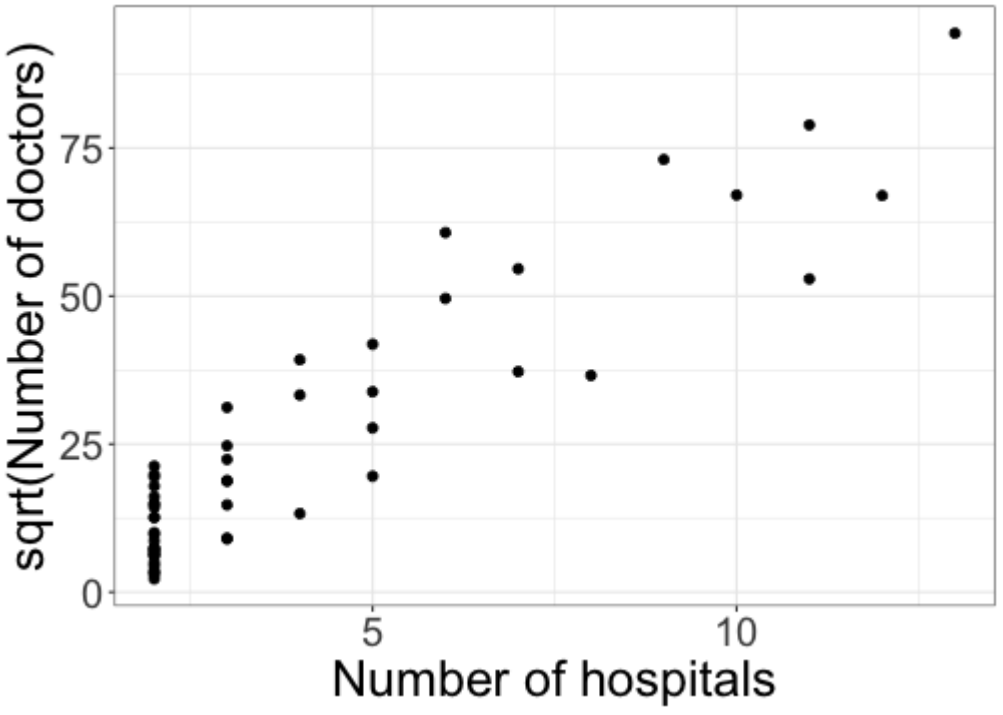
But here, $\text{Var}(y_i)$ increases with μ_i

Does a linear regression model seem appropriate for this relationship?

Trying a transformation

$\sqrt{Y_i} \sim N(\mu_i, \sigma^2)$

$\mu_i = \beta_0 + \beta_1 \text{Hospitals}_i$



Problems:

1) $\sqrt{Y_i}$ is discrete but $N(\mu, \sigma^2)$ is continuous

2) $\sqrt{Y}$ harder to interpret

3) $\sqrt{Y} \geq 0$

> Is a linear regression model appropriate now?

Poisson regression

(random component) $Y_i \sim \text{Poisson}(\lambda_i)$

$\mathbb{E}[Y_i] = \lambda_i$

$\text{Var}(Y_i) = \lambda_i$

(systematic component) $g(\lambda_i) = \beta^T X_i$

g: link function

Canonical link: $g(\lambda_i) = \log(\lambda_i)$

$\log(\lambda_i) = \beta^T X_i$

1) $\log(\lambda_i) \in (-\infty, \infty)$
2) Leads to nice interpretation
3) Nice mathematical properties...

Fitting the Poisson regression model

```
m1 <- glm(MDs ~ Hospitals, data = CountyHealth,
          family = poisson)
summary(m1)
```

```
...
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 5.116896   0.009801   522.1   <2e-16 ***
## Hospitals   0.312442   0.001048   298.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 111627  on 52  degrees of freedom
## Residual deviance:  22799  on 51  degrees of freedom
## AIC: 23197
...
```

$E(\log(Y)) \neq \log E(Y)$

Interpreting the Poisson regression model

```
m1 <- glm(MDs ~ Hospitals, data = CountyHealth,
          family = poisson)
summary(m1)
```

```
...
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 5.116896   0.009801   522.1   <2e-16 ***
## Hospitals   0.312442   0.001048   298.2   <2e-16 ***
...
```

$\log(\hat{\lambda}_i) = 5.12 + 0.31\ \text{Hospitals}_i$

One additional hospital is associated with...

- a 0.31 increase in log average # of doctors
- an increase in the average # of doctors by a factor of $e^{0.31} = 1.37$

Exponential dispersion models

probability function for Poisson:

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \frac{1}{y!}\exp\{y\log\lambda - \lambda\}$$

$$= a(y,\phi)\exp\left\{\frac{y\theta - \kappa(\theta)}{\phi}\right\} \quad \text{(EDM)}$$

$a(y,\phi) = \frac{1}{y!}$ normalizing function

$\theta = \log\lambda$ canonical parameter

$\kappa(\theta) = \lambda$ cumulant function

$\phi = 1$ dispersion parameter