Lecture 7

Count variables

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

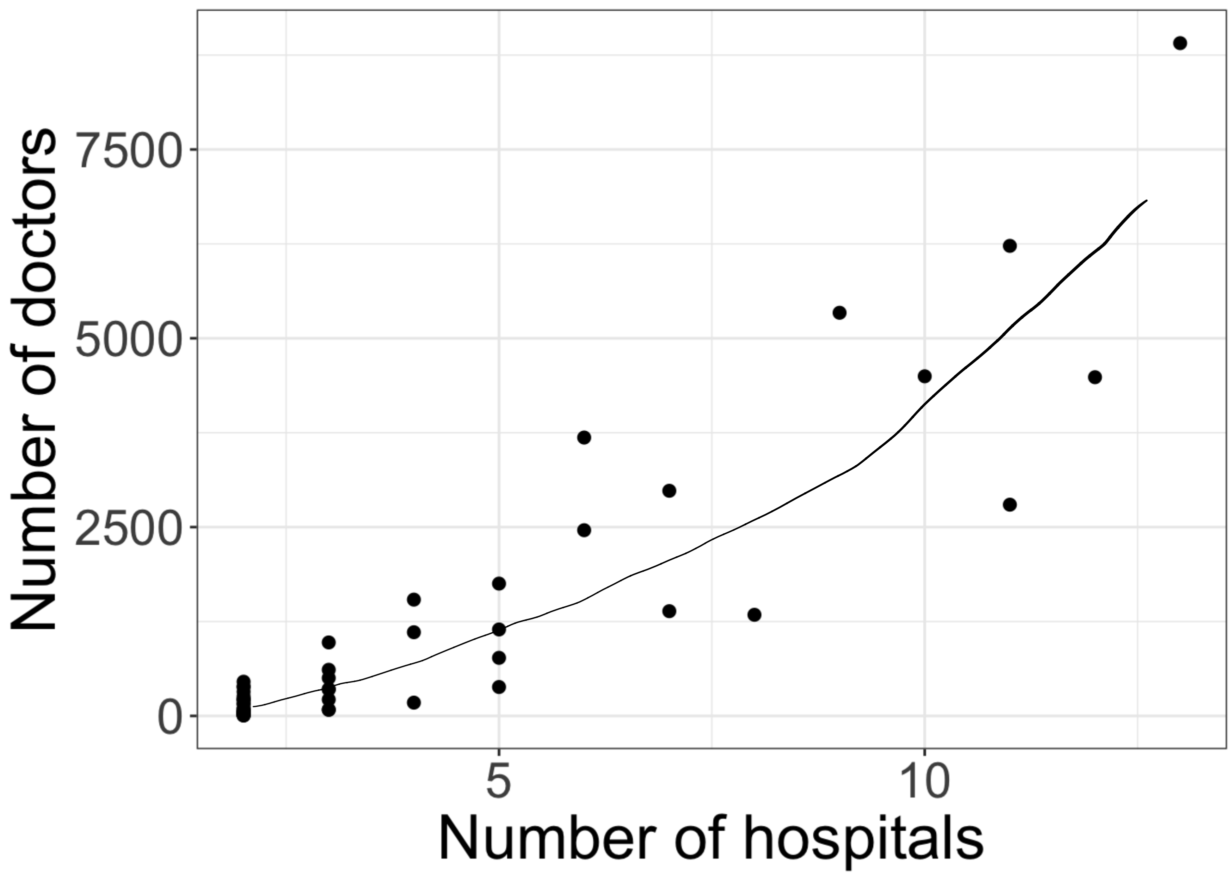
- `MDs`: the number of medical doctors in the county
- `Hospitals`: the number of hospitals in the county

count variable

values 0, 1, 2, ...

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

Plotting the data



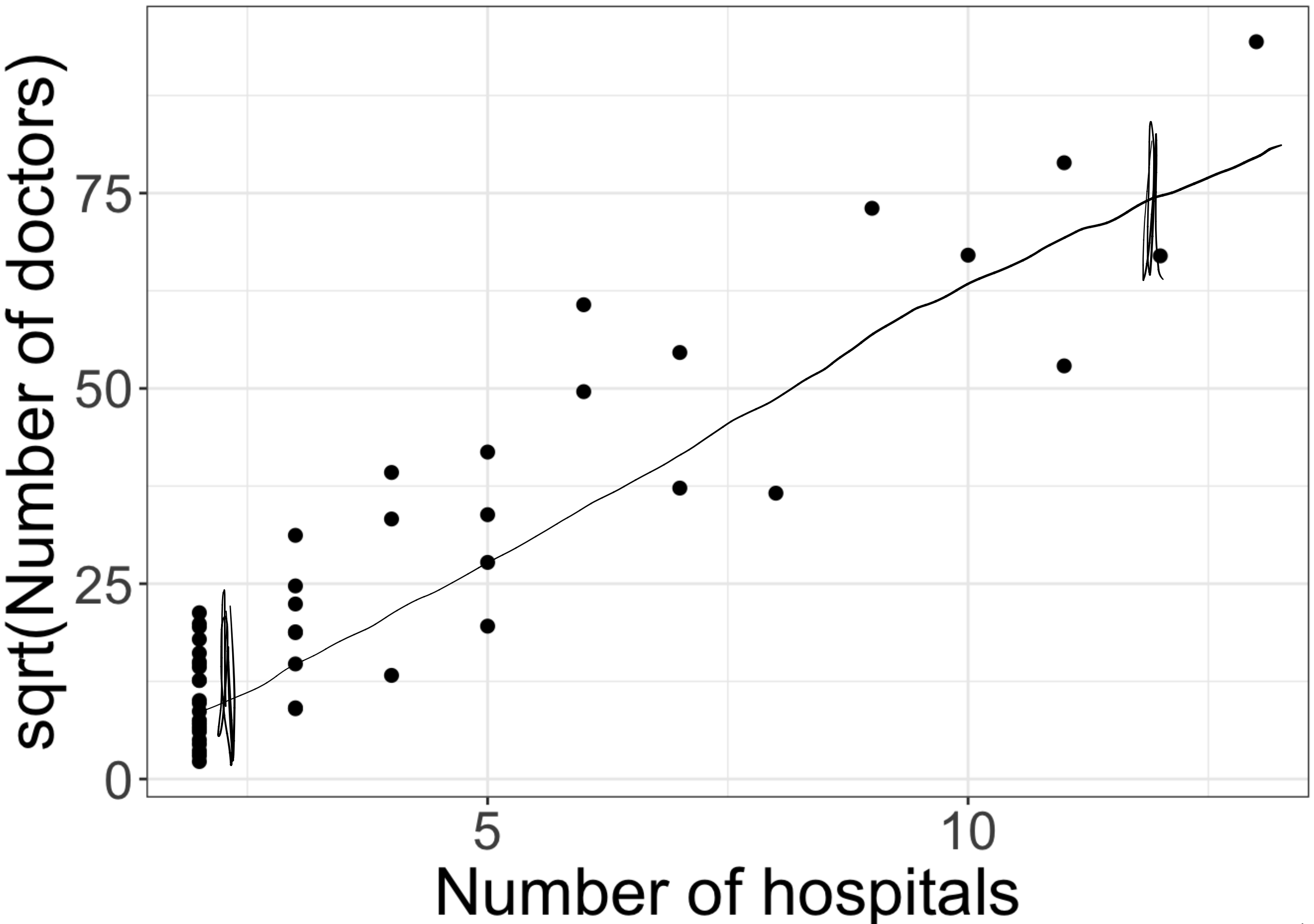
Question: Does a linear regression model seem appropriate for this relationship?

- may have issues predicting negative # of MDs
- maybe some non-linear relationship? (maybe transform)
- non-constant variance (transformation?)

Trying a transformation

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

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



· $\sqrt{Y_i}$ is discrete, but $N(\mu, \sigma^2)$ is continuous (⇒ linear regression would not be MLE)

· Interpret in terms of $\sqrt{Y_i}$ (harder to interpret than Y_i)

Is a linear regression model appropriate now?

· still might have issues with negative predictions

· shape & constant variance look better

Poisson regression

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

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

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

logistic:

(systematic component) $\underbrace{g(\lambda_i)}_{\text{link function}} = \beta^T X_i$

$\log\left(\frac{p_i}{1-p_i}\right) = \beta^T X_i$

linear:

$\mu_i = \beta^T X_i$

Canonical link: $g(\lambda_i) = \log(\lambda_i)$ (for Poisson regression)

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

Fitting the Poisson regression model

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

```
...
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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

Number of Fisher Scoring iterations: 5
...
```

Interpreting the Poisson regression model

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

```
...
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
...
```

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

One additional hospital is associated with...

- an increase of 0.31 in log average # of doctors

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

Exponential dispersion models

probability function for Poisson:

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

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

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

$\phi = 1$ (dispersion parameter)

$\theta = \log\lambda$ (canonical parameter)

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

$$f(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta - \kappa(\theta)}{\phi}\right\}$$

<u>Normal</u>: $Y \sim N(\mu, \sigma^2)$ $\qquad \sigma^2$ is <u>known</u>

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left[y^2 - 2\mu y + \mu^2\right]\right\}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2}\right\}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-y^2}{2\sigma^2}\right\}\exp\left\{\frac{y\mu - (\mu^2/2)}{\sigma^2}\right\}$$

$$a(y,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-y^2}{2\sigma^2}\right\}$$

$$\phi = \sigma^2$$

$$\theta = \mu$$

$$\kappa(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$$