What to do today (Mar 1)?

4. Analysis with Count Response (Chp 4)

4.1 Poisson Model for Counts (Chp 4.1)

4.2 Poisson Regression Analysis (Chp 4.2)

- **4.2.1 Introduction to Poisson regression models**
- **4.2.2 Poisson regression analysis**
- **4.2.3 Categorical explanatory variables**
- **4.2.4 Poisson regression with contingency tables**

4.3 Additional Topics on Count Responses (Chp 4.3-4)

4.2.1 Introduction to Poisson regression models

Goal: To study how a count variable Y depends on another variable X
e.g. the original response in the horseshoe crab study, "number of satellites"

Poisson Regression Model (Loglinear Model): Consider the response $Y|x \sim Poisson\big(\mu(x)\big)$ and assume

$$\log[\mu(x)] = \alpha + \beta x$$

equivalently to $\mu(x) = \exp(\alpha + \beta x)$.

- $\mu(x) \geq 0$ for all x
- β's interpretation: when $x \uparrow x + \Delta x$,
  - $\beta = 0 \Rightarrow \mu(x) = e^{\alpha}$;
  - $\beta > 0 \Rightarrow \mu(x) \uparrow \mu(x)e^{\beta \Delta x}$;
  - $\beta < 0 \Rightarrow \mu(x) \downarrow \mu(x)e^{\beta \Delta x}$

4.2.1 Introduction to Poisson regression models

Goal: To study how a count variable Y depends on other variables $X_1, \ldots, X_K$
e.g. the original response in the horseshoe crab study, "number of satellites"

Poisson Regression Model (Loglinear Model): Consider the response $Y|x_1, \ldots, x_K \sim Poisson\big(\mu(x_1, \ldots, x_K)\big)$ and assume

$$\log[\mu(x_1, \ldots, x_K)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K$$

equivalently to $\mu(x_1, \ldots, x_K) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K)$.

- $\mu(x_1, \ldots, x_K) \geq 0$ for all $x_1, \ldots, x_K$
- β_1's interpretation with fixed $x_2, \ldots, x_K$: when $x_1 \uparrow x_1 + \Delta x_1$,
  - $\beta_1 = 0 \Rightarrow \mu(x_1, \ldots, x_K)$ stays the same $e^{\beta_0 + \beta_2 x_2 + \ldots + \beta_K x_K}$;
  - $\beta_1 > 0 \Rightarrow \mu(x_1, \ldots, x_K) \uparrow \mu(x_1, \ldots, x_K) e^{\beta_1 \Delta x_1}$;
  - $\beta_1 < 0 \Rightarrow \mu(x_1, \ldots, x_K) \downarrow \mu(x_1, \ldots, x_K) e^{\beta_1 \Delta x_1}$

4.2.2 Poisson regression analysis

Statistical Inference with Poisson Regression (Loglinear) Models

For the count response Y and explanatory variables $\mathbf{X}$ with $Y|\mathbf{X} = \mathbf{x} \sim Poisson\big(\mu(\mathbf{x})\big)$, and data $\big\{(y_i, \mathbf{x}_i) : i = 1, \dots\big\}$ from n indpnt individuals: the likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{\mu(\mathbf{x}_i)^{y_i}}{y_i!} e^{-\mu(\mathbf{x}_i)}.$$

- Estimation of α, β: MLE and CI?
- Estimation of $\mu(\mathbf{x})$: MLE and CI?
- Testing on hypotheses about α, β: e.g. $H_0 : \beta = 0$?
- Model Comparison/Checking? [to be studied in Chp5]

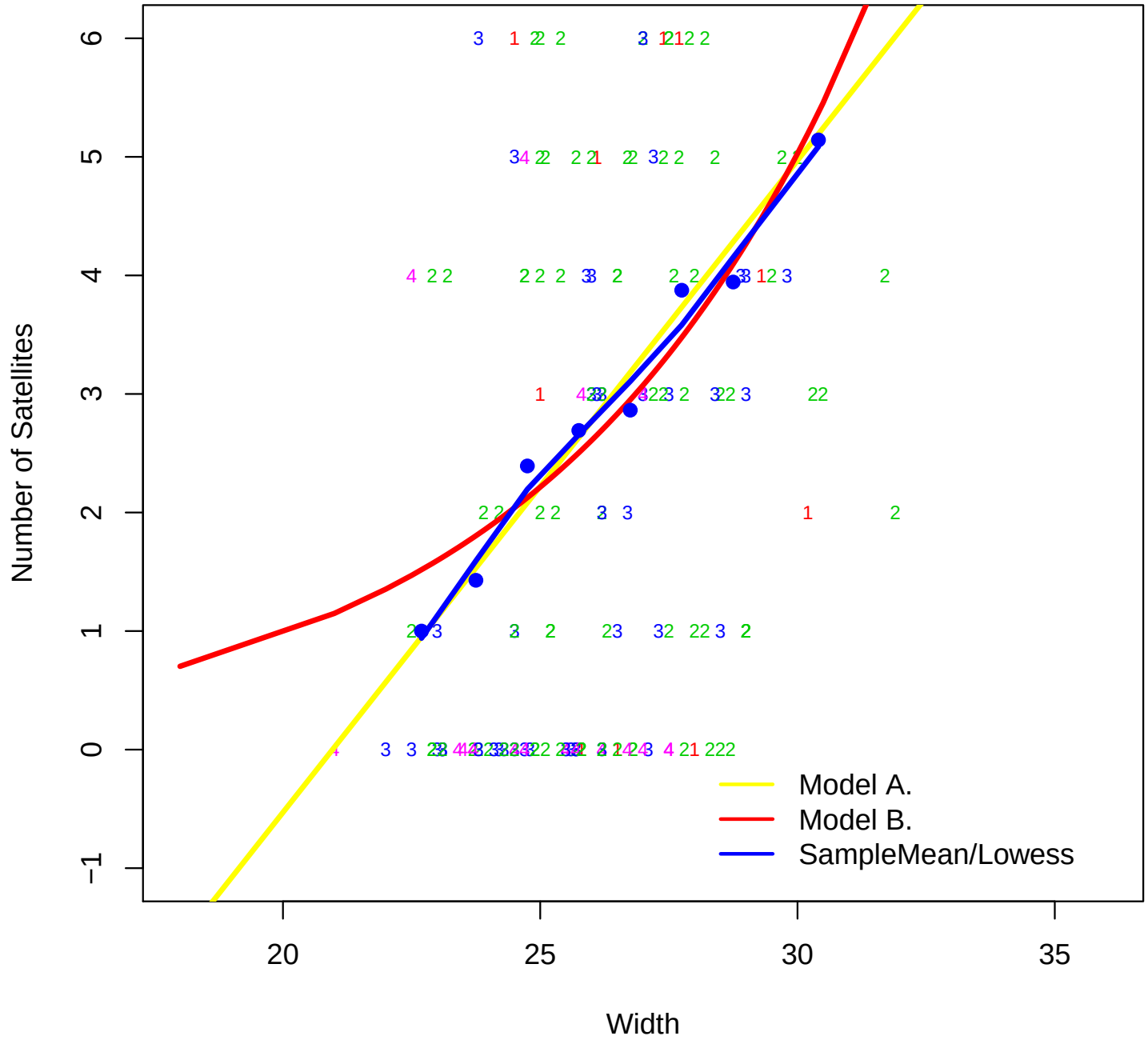
4.2.2 Poisson regression analysis

Example. Revisit 4 to the Horseshoe Crab Study: (Regression with Count Response) Y=number of satellites; X=width

Assume $Y \sim Poisson(\mu(x))$ with

- Model A. Linear mean model: $\mu(x) = \alpha + \beta x$;
  $\hat{\alpha} = -11.53(2.832), \hat{\beta} = .55(.107)$
  with R function $glm(formula, family = gaussian)$

- Model B. Loglinear mean model: $\log\big(\mu(x)\big) = \alpha + \beta x$;
  $\hat{\alpha} = -3.305(.542), \hat{\beta} = .164(.020)$
  with R function $glm(formula, family = poisson)$


Number of Satellites
Width
Model A.
Model B.
SampleMean/Lowess

4.2.3A Categorical explanatory variables

For example, to study how the count Y depends on two categorical variables X and Z: $E(Y|X = x, Z = z) = \mu(x, z)$?

Consider the Poisson (loglinear) regression model

$$\log(\mu_{ik}) = \beta_0 + \beta_i^X + \beta_k^Z$$

for $i = 1, \ldots, I$ and $k = 1, \ldots, K$ with $\beta_1^X = \beta_1^Z = 0$ and $\mu_{ik} = \mu(i, k)$.

Alternatively, consider

$$\log(\mu(x, z)) = \beta_0 + \beta_2^X x_2 + \ldots + \beta_I^X x_I + \beta_2^Z z_2 + \ldots + \beta_K^Z z_K$$

with dummy variables $x_2, \ldots, x_I$ for X and dummy variables $z_2, \ldots, z_K$ for Z.

4.2.3A Categorical explanatory variables

For example, to study how the count Y depends on two categorical variables X and Z: $E(Y|X = x, Z = z) = \mu(x, z)$?

Consider the Poisson (loglinear) regression model

$$\log(\mu_{ik}) = \beta_0 + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$$

for $i = 1, \ldots, I$ and $k = 1, \ldots, K$ with
$\beta_1^X = \beta_1^Z = \beta_{1k}^{XZ} = \beta_{i1}^{XZ} = 0$ and $\mu_{ik} = \mu(i, k)$,

Alternatively, consider

$$\log(\mu(x, z)) = \beta_0 + \sum \beta_i^X x_i + \sum \beta_k^Z z_k + \sum \beta_{ik}^{XZ} x_i z_k$$

with dummy variables $x_2, \ldots, x_I$ for X and dummy variables $z_2, \ldots, z_K$ for Z.

4.2.3B Categorical explanatory variables

For example, in the crab study, the variable color has categories of light medium, medium, dark medium, and dark. **How to handle it?**

- *Ignore the underlying order.* Treat it as nominal?
- *Score the categories.* Treat it as quantitative using the scores ?
  - e.g. assigning 1,2,3,4 to light medium, medium, dark medium, and dark colors
  - e.g. assigning 4,3,2,1 to light medium, medium, dark medium, and dark colors
  - e.g. assigning 1,3,4,6 to light medium, medium, dark medium, and dark colors

The scoring needs to be as non-subjective as possible.

4.2.3C Categorical explanatory variables

Example. Revisit 4 to the Horseshoe Crab Study (cont'd)
Step A. Data Description.

$R : ex.crab[1:3,]$

Obstn	C	S	W	Wt	Sa
1	2	3	28.3	3.05	8
2	3	3	22.5	1.55	0
3	1	1	26.0	2.30	9

- who? $n = 173$ female horseshoe crabs selected by a study
- what?
  - C=color: 1,2,3,4 for light med, medium, dark med and dark
  - S=spine: 1, 2,3 for both good, one or both worn/broken
  - W=width: ranging 21.0 to 33.5cm
  - Wt=weight: ranging 1.2kg to 5.2kg
  - Sa=number of satellites
- why? to explore the association of Sa with other variables
- when and where?

4.2.3C Categorical explanatory variables

Example. Revisit 4 to the Horseshoe Crab Study (cont'd)

Step B. Data Analysis.

Preparation ... ...
$C < -as.factor(ex.crab[, 1]);$
$S < -as.factor(ex.crab[, 2]);$
$W < -ex.crab[, 3]; Wt < -ex.crab[, 4];$
$Sa < -round(ex.crab[, 5]);\ tmpyB < -Sa$

Modeling ... ...
Consider $E(Sa|C = c, S = s, W = w) = \mu(c, s, w)$ as

$$\log(\mu(c, s, w)) = \beta_0 + \beta_i^C + \beta_j^S + \beta^W w$$

for $i = 1, 2, 3, 4$ and $j = 1, 2, 3$

Analyzing ... ...

$R : tmp.outB1a < -glm(tmpyB \sim C + S + W, family = poisson)$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-2.54385	0.62426	-4.075	4.60e-05	***
C2	-0.22158	0.16789	-1.320	0.1869	
C3	-0.46036	0.19554	-2.354	0.0186	*
C4	-0.48544	0.22824	-2.127	0.0334	*
S2	-0.13879	0.21269	-0.653	0.5141	
S3	0.02363	0.11729	0.201	0.8403	
W	0.14596	0.02189	6.669	2.58e-11	***

Null deviance: 632.79 on 172 degrees of freedom
Residual deviance: 558.63 on 166 degrees of freedom

Alternative ways of using the color variable? tmpC=1,2,3,4

$R : tmp.outB1c < -glm(tmpyB \sim tmpC + W, family = poisson)$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-2.51998	0.61063	-4.127	3.68e-05	***
tmpC	-0.16940	0.06184	-2.739	0.00616	**
W	0.14957	0.02068	7.233	4.72e-13	***

Null deviance: 632.79 on 172 degrees of freedom
Residual deviance: 560.20 on 170 degrees of freedom

4.2.4A Poisson regression with contingency tables: Two-Way Contingency Tables

Recall that

- n individuals cross-classified according to (row) X and (column) Y variables $\Longrightarrow$ an $I \times J$ contingency table with cell counts $\{N_{ij} : i = 1, \ldots, I; j = 1, \ldots, J\}$
- the joint cell prob $\pi_{ij} = P(X = i, Y = j)$, and the expected cell count $\mu_{ij} = E(N_{ij}) = n\pi_{ij}$

To establish the association of the cell count (V) with X and Y:

- view $\mu_{ij} = E(N_{ij}) = E(V|X = i, Y = j)$
- assume $Nij \sim Poisson(\mu_{ij})$
- consider a loglinear regression model, where V is the response and X, Y are the explanatory variables

4.2.4A Poisson regression with contingency tables: Two-Way Contingency Tables

Loglinear Model of Independence If X and Y are independent, $\pi_{ij} = \pi_{i+}\pi_{+j}$ and $\log \mu_{ij} = \log n + \log \pi_{i+} + \log \pi_{+j} \Longrightarrow$

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

λ_i^X, λ_j^Y: the main effects of X, Y (row/column effects)

- **parameterization:** one of λ_i^X and one of λ_j^Y are redundant, since $\sum_i \pi_{i+} = 1$ and $\sum_j \pi_{+j} = 1$

the three commonly used coding schemes

4.2.4A Poisson regression with contingency tables: Two-Way Contingency Tables

- **interpretation:**
  e.g. I=2 and X=1,2 for male, female; J=2 and Y=1,2 for success, failure:

$$\log\left[\frac{\mu_{i1}}{\mu_{i2}}\right] = [\lambda + \lambda_i^X + \lambda_1^Y] - [\lambda + \lambda_i^X + \lambda_2^Y] = \lambda_1^Y - \lambda_2^Y$$

  - $\log\left[\frac{\mu_{i1}}{\mu_{i2}}\right] = \log\left[\frac{\pi_{i1}}{\pi_{i2}}\right] = \log\left[\frac{P(Y=1|X=i)}{1-P(Y=1|X=i)}\right]$: the log-odds of sucess with $X = i$ is $\lambda_1^Y - \lambda_2^Y$.
  - the log-odds ratio of success of male comparing with female is $\log\left[\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)}\right] - \log\left[\frac{P(Y=1|X=2)}{1-P(Y=1|X=2)}\right] = 0$, i.e. the OR=1.

4.2.4A Poisson regression with contingency tables: Two-Way Contingency Tables

Saturated Loglinear Model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

- the association terms λ_{ij}^{XY} (two-factor interactions): reflecting deviations from $X \perp Y$
- the number of non-redundant parameters: 1+(I-1)+(J-1)+(I-1)(J-1) = IJ
- with a 2×2 table, $OR = 1 \Leftrightarrow \lambda_{ij}^{XY} = 0$ for all i,j:

$$\log OR = \log\left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\right) = (\lambda_{11}^{XY} + \lambda_{22}^{XY}) - (\lambda_{12}^{XY} + \lambda_{21}^{XY})$$

- perfect fit: num unknowns = d.f. of the table [not practical – non-saturated model: smooth the sample data, simpler interpretation]

4.2.4A Poisson regression with contingency tables: Two-Way Contingency Tables

Statistical Inference

the statistical analysis with the loglinear (Poisson) regression model with two categorical predictors: need be careful with coding X, Y

- estm model parameters: the main effects and two factor interactions $\lambda_i^X, \lambda_j^Y, \lambda_{ij}^{XY}$
- estm μ_{ij}, the expected counts $\Rightarrow$ e.g. estm of OR
- model checking/comparison: Pearson's χ^2-test, LRT-test recall testing for $X \perp Y$ with a two way contingency table?

What will we study next?

4. Analsyis with Count Response (Chp 4)

- *4.1 Poisson Model for Count Data (Chp 4.1)*
- **4.2 Poisson Regression Analysis (Chp 4.2)**
  - *4.2.1 Introduction to Poisson regression models*
  - *4.2.2 Poisson regression analysis*
  - *4.2.3 Categorical explanatory variables*
  - **4.2.4 Poisson regression with contingency tables**
- **4.3 Additional Topics on Count Responses (Chp 4.3-4)**