What to do today (Mar 27)?

1. *Introduction and Preparation*
2. *Analysis with Binary Variables (Chp 1-2)*
3. *Analysis with Multicategory Variables (Chp 3)*
4. *Analysis with Count Variables (Chp 4)*
5. *Model Selection and Evaluation (Chp 5)*

6. **Additional Topics (Chp 6)**
   - *6.1 Exact inference (Chp 6.2)*
   - *6.2 Revisit to Loglinear and Logistic Models for Contingency Tables: the Loglinear-Logit Connection* (Supplementary)
   - **6.3 Revisit III to GLM and Some Advanced Topics (Chp 5.3, Chp 6.5)**
     - **6.3.1 Revisit III to GLM**
     - *6.3.2 Marginal Modeling*
     - *6.3.3 Mixed Ect Models for Correlated Data*

6.3.1 Revisit III to GLM

GOAL: to study how $Y \leftarrow X_1, \ldots, X_K$?

Generalized Linear Models:

- **Random Component.** response r.v. Y follows a distn with $\mu(x_1, \ldots, x_k) = E(Y|x_1, \ldots, x_k)$ to be examined

- **Systematic Component.** $\alpha + \beta_1 x_1 + \ldots + \beta_K x_K$
  Some x_k can be based on others: e.g. $x_3 = x_1 x_2$.

- **Link Function.** $g(\mu) = \alpha + \beta_1 x_1 + \ldots + \beta_K x_K$
  The link function $g(\cdot)$ links the *random componet* through its mean and the *systematic component*.

Recall the *glm* function in R to conduct a GLM analysis:

*R: tmp.out< −glm(Y∼X*Z,family)*

family(object,...) in *R* for function *glm*, for example

- binomial(link = "logit")
- poisson(link = "log")
- gaussian(link = "identity") $\Longrightarrow$ *R: e.g. lm(Y∼X*Z)*
- and some others, such as quasipoisson(link = "log") to be studied

6.3.1B Revisit III to GLM: Additional Examples

To study $Y \leftarrow X, Z$? with binary response $Y = 1$, or 0 and explanatory variables X, Z:

- Recall the *Logistic Regression Model (Logit)*:
  - *Randome Component.* r.v. $Y \sim Bernoulli(\pi)$ with $\mu(x,z) = E(Y|X = x, Z = z) = P(Y = 1|X = x, Z = z) = \pi(x,z)$ and $V(Y|X = x, Z = z) = \pi(x,z)[1 - \pi(x,z)]$
  - *Systematic Component.* $h(x,z) = \alpha + \beta x + \gamma z + \eta xz$, a linear function of x, z, xz
  - *Link Function.* $g : \mu \to logit(\mu)$:
    $logit\left[\mu(x,z)\right] = logit\left[\pi(x,z)\right] = h(x,z)$
    $\Leftrightarrow \pi(x,z) = \frac{\exp(h(x,z))}{1+\exp(h(x,z))}$.

Any alternative model?

Probit Regression Model.

To study $Y \leftarrow X, Z?$ with binary response $Y = 1$, or 0 and explanatory variables X, Z:

- the *Probit Regression Model (Probit)*:
  - *Randome Component.* r.v. $Y \sim Bernoulli(\pi)$ with $\mu(x, z) = E(Y|X = x, Z = z) = P(Y = 1|X = x, Z = z) = \pi(x, z)$ and $V(Y|X = x, Z = z) = \pi(x, z)[1 - \pi(x, z)]$
  - *Systematic Component.* $h(x, z) = \alpha + \beta x + \gamma z + \eta xz$, a linear function of x, z, xz
  - *Link Function.* $g : \mu \to probit(\mu)$: $probit\big[\mu(x, z)\big] = probit\big[\pi(x, z)\big] = h(x, z)$ $\Leftrightarrow \pi(x, z) = \Phi(h(x, z))$

$\Phi(\cdot)$ the cumulative distn of $N(0, 1)$: e.g. $\Phi(-1.645) = 0.05$ and $\Phi(1.96) = 1 - 0.025$

6.3.1B Revisit III to GLM: Additional Examples

Poisson Regression Model To study $Y \leftarrow X, Z$? with count response Y and predictors X, Z:

- Recall *Loglinear Regression Models (Poisson Regression)*:
  - *Randome Component.* r.v. $Y \sim Poisson(\mu)$ with $\mu(x, z) = E(Y|X = x, Z = z)$ and $V(Y|X = x, Z = z) = \mu(x, z)$
  - *Systematic Component.* $h(x, z) = \alpha + \beta x + \gamma z + \eta xz$, a linear function of x, z, xz
  - *Link Function.* $g : \mu \to \log(\mu)$: $\log\left[\mu(x, z)\right] = h(x, z)$ $\Leftrightarrow \mu(x, z) = \exp(h(x, z))$

What if $\mu(x, z) = E(Y|X = x, Z = z)$ but $V(Y|X = x, Z = z) > \mu(x, z)$: greater variability than expected $\leftarrow$ overdispersion? $\Longrightarrow$ to study the following ...

6.3.1B Revisit III to GLM: Additional Examples

To study $Y \leftarrow X, Z?$ with count response Y and predictors X, Z:

Quasi-Poisson Regression:

- *Randome Component.* r.v. Y with $\mu(x,z) = E(Y|X=x, Z=z)$ and $V(Y|X=x, Z=z) = \rho\mu(x,z)$
- *Systematic Component.* $h(x,z) = \alpha + \beta x + \gamma z + \eta xz$, a linear function of x, z, xz
- *Link Function.* $g : \mu \to \log(\mu)$: $\log\left[\mu(x,z)\right] = h(x,z)$ $\Leftrightarrow \mu(x,z) = \exp(h(x,z))$

6.3.1C Revisit III to GLM: Final visit to the Horseshoe Crab Study

Data Description.

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 (with the distn: 12, 95, 44, 22)
  - S=spine: 1, 2,3 for both good, one or both worn/broken (with the distn: 37, 15, 121)
  - W=width: ranging 21.0 to 33.5cm (with mean, sd: 26.4, 2.1)
  - Wt=weight: ranging 1.2kg to 5.2kg (with mean, sd: 2.44, 0.58)
  - Sa=number of satellites (ranging from 0 to 19)
- why? to explore the association of Sa with other variables
- when and where?

Conduct Regression Analyses

A. Regression with Binary Response

Preparation ... ...
$C < -as.factor(ex.crab[,1]); S < -as.factor(ex.crab[,2]);$
$W < -ex.crab[,3]; Wt < -ex.crab[,4];$
$tmpyA < -ifelse(Sa > 0, 1, 0)$

- **A.1 Logistic Regression**
- **A.2 Probit Regression**
- **A.3 Comparisons**

$R : tmp.outA1a <- glm(tmpyA \sim C + S + W + Wt, family = binomial)$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-8.06501	3.92855	-2.053	0.0401	*
C2	-0.10290	0.78259	-0.131	0.8954	
C3	-0.48886	0.85312	-0.573	0.5666	
C4	-1.60867	0.93553	-1.720	0.0855	.
S2	-0.09598	0.70337	-0.136	0.8915	
S3	0.40029	0.50270	0.796	0.4259	
W	0.26313	0.19530	1.347	0.1779	
Wt	0.82578	0.70383	1.173	0.2407	

Null deviance: 225.76 on 172 degrees of freedom

Residual deviance: 185.20 on 165 degrees of freedom

AIC: 201.2

surprising analysis results about the effects of the predictors!

$\Longrightarrow$ the investigation on the possible collinearity ...

Are W and Wt closely correlated?

$\Longrightarrow$ removing Wt from the list of predictors ...

$R : tmp.outA1b < -glm(tmpyA \sim C + S + W, family = binomial)$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-11.09953	2.97706	-3.728	0.000193	***
C2	-0.14340	0.77838	-0.184	0.853830	
C3	-0.52405	0.84685	-0.619	0.536030	
C4	-1.66833	0.93285	-1.788	0.073706	.
S2	-0.05782	0.70308	-0.082	0.934453	
S3	0.37703	0.50191	0.751	0.452540	
W	0.45624	0.10779	4.233	2.31e-05	***

Null deviance: 225.76 on 172 degrees of freedom

Residual deviance: 186.61 on 166 degrees of freedom

AIC: 200.61

Is it the model to use?

Model Selection (Variable Selection):

$tmp.outA1c <- glm(tmpyA \sim C * S * W, family = binomial)$
$step(tmp.outA1c)$
Start: AIC=212.44
$tmpyA \sim C * S * W$

	Df	Deviance	AIC
- C:S:W	3	173.67	209.67
$< none >$		170.44	212.44

Step: AIC=209.67
$tmpyA \sim C + S + W + C : S + C : W + S : W$

⋮ ⋮ ⋮ ⋮

$Call : glm(formula = tmpyA \sim C + W, family = binomial(link = \text{``logit''}))$
Coefficients:

(Intercept)	C2	C3	C4	W
-11.38519	0.07242	-0.22380	-1.32992	0.46796

Degrees of Freedom: 172 Total (i.e. Null); 168 Residual
Null Deviance: 225.8
Residual Deviance: 187.5 AIC: 197.5

Alternative ways of using the color variable?

- C=1,2,3,4 as an ordinal variable?

$glm(formula = tmpyA \sim tmpC + W, family = binomial)$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-10.0708	2.8068	-3.588	0.000333	***
tmpC	-0.5090	0.2237	-2.276	0.022860	*
W	0.4583	0.1040	4.406	1.05e-05	***

Null deviance: 225.76 on 172 degrees of freedom
Residual deviance: 189.12 on 170 degrees of freedom
AIC: 195.12

- Group the categories of color into two: dark vs lighter color?

$glm(formula = tmpyA \sim tmpCb + W, family = binomial)$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-11.6790	2.6925	-4.338	1.44e-05	***
tmpCb2	-1.3005	0.5259	-2.473	0.0134	*
W	0.4782	0.1041	4.592	4.39e-06	***

Null deviance: 225.76 on 172 degrees of freedom
Residual deviance: 187.96 on 170 degrees of freedom
AIC: 193.96

Report the Regression with $logit\left[\pi(i,w)\right] = \alpha + \beta_i^C + \gamma w$: $i = 1,2$ for lighter, dark color.

- The fitted model:

$$logit\left[\hat{\pi}(i,w)\right] = \begin{cases} -11.68 + 0.48w & \text{for i=1 (lighter color)} \\ -11.68 - 1.30 + 0.48w & \text{for i=2 (dark color)} \end{cases}$$

  or $logit\left[\hat{\pi}(x,w)\right] = -11.68 - 1.30x + 0.48w$ if using the dummy variable $x = 0,1$ for lighter, dark color.

- Is YesSa positively associated with W in the presence of C?

  To conduct a test on $H_0 : \gamma = 0$ vs $H_1 : \gamma > 0$:
  $Z = \frac{\hat{\gamma}}{SE_{\hat{\gamma}}}$; $Z_{obs} = 4.59$; $p = 4.39e - 06/2$

An alternative: to compare $M_0 : tmpA \sim Logit(tmpCb)$ vs $M_1 : tmpA \sim Logit(tmpCb, W)$

This can only test on $H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$: (i) fit both M_0 and M_1, (ii) obtain their $G(M_0|M_s) = 214.79$ with df=171, $G(M_1|M_s) = 187.96$ with df=170 $\Rightarrow$ $G(M0|M1) = 214.79 - 187.96$; $df = 1$; $p = 1 - pchisq(26.83, 1) < 0.001$

Report the Regression with $logit\left[\pi(i,w)\right] = \alpha + \beta_i^C + \gamma w$: $i = 1, 2$ for lighter, dark color.

- What is the OR of YesSa comparing ligther vs dark color crab adjusting for W? Give its MLE and an 95% CI.
  $\log OR = \beta_1^C - \beta_2^C$: its MLE is $0 - \hat{\beta}_2^C = 1.30$ with $SE_{\hat{\beta}_2^C} = 0.526$
  $\Longrightarrow$ OR's MLE 3.67 and 95% CI $(1.31, 10.29)$
- Give estimates of the probability of YesSa with lighter and dark colored crabs if their width$=$ 26.3cm (the mean width of the observed crabs') and width=35cm: $\pi(i,w) = \frac{exp(\alpha+\beta_i^C+\gamma w)}{1+exp(\alpha+\beta_i^C+\gamma w)}$

	width=26.3cm		width=35.0cm	
Estimates	lighter (i=1)	dark (i=2)	lighter (i=1)	dark (i=2)
$\hat{\alpha} + \hat{\beta}_i^C + \hat{\gamma} w$	0.90	-0.40	5.06	3.76
(SE)	(0.20)	(0.49)	(0.98)	(1.08)
95% CI	(0.51,1.29)	(-1.37,1.86)	(3.14,6.98)	(1.64,7.18)
$\hat{\pi}(i,w)$	0.71	0.40	0.99	0.98
95% CI	(0.62,0.78)	(0.20, 0.87)	(0.96,1.00)	(0.84,1.00)

$R: tmp.outA2 < -glm(tmpyA \sim tmpCb + W, family = binomial(link = \text{"probit"}))$

	Estimate	Std. Error	z value	$Pr(>\lvert z \rvert)$	
(Intercept)	-6.98838	1.54195	-4.532	5.84e-06	***
tmpCb2	-0.76494	0.31341	-2.441	0.0147	*
W	0.28637	0.05924	4.834	1.34e-06	***

Null deviance: 225.76 on 172 degrees of freedom
Residual deviance: 187.72 on 170 degrees of freedom

AIC: 193.72

MLE and 95% CI for the prob of YesSa with lighter colored crabs and width=26.3cm:

- $\hat{\pi}(1, 26.3) = pnorm(\hat{\alpha} + \hat{\beta}_1^C + \hat{\gamma}26.3) = 0.706$
- CI: (0.624, 0.779)

B. Regression with Count Response

- **B.1 Poisson Regression**
- **B.2 Quasi-Poisson Regression**
- **B.3 Comparisons**

Preparation ... ...

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

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

	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

AIC: 927.93

Alternative ways of using the color variable?

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

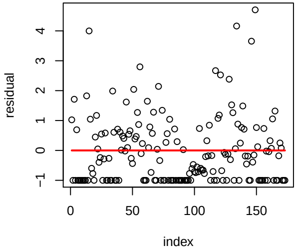
	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-2.51998	0.61063	-4.127	3.68e-05	***
tmpC	-0.16940	0.06184	-2.739	0.00616	**
W	0.14957	0.02068	7.233	4.72e-13	***

Null deviance: 632.79 on 172 degrees of freedom

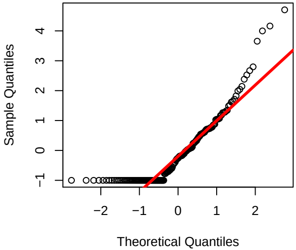
Residual deviance: 560.20 on 170 degrees of freedom

AIC: 921.5

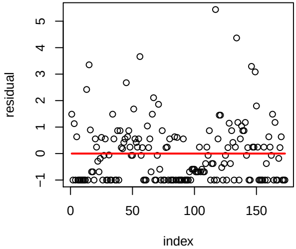
Model Checking: Residual Plots:



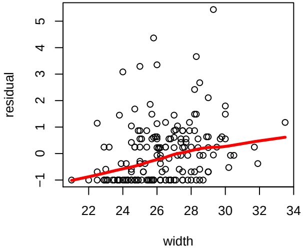
(1) Residuals with outB1c



(2) outB1c residual qqnorm



(3) Residuals with outB1c−W



(4) Residuals with outB1c−W vs W

What if the Poisson assumption is not appropriate?

$R : tmp.outB2a <- glm(tmpyB \sim tmpC + W, family = quasipoisson(link = \text{"log"}))$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-2.51998	1.09722	-2.297	0.0229	*
tmpC	-0.16940	0.11112	-1.524	0.1292	
W	0.14957	0.03716	4.025	8.55e-05	***

(Dispersion parameter for quasipoisson family taken to be 3.228764)

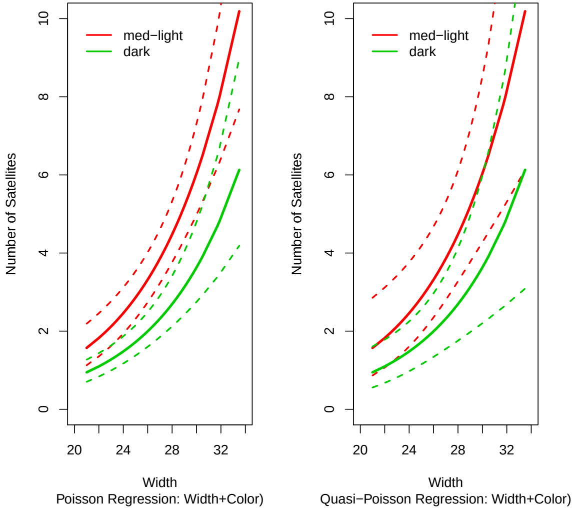
Null deviance: 632.79 on 172 degrees of freedom

Residual deviance: 560.20 on 170 degrees of freedom

AIC: NA

Comparisons between Poisson vs Quasi-Poisson:

- estm for the parameters: the same
- estm for the SE of the parameter estimators: different when the counts are overdispersed
  - Poisson Regression: under-estm the SE



What will we study next?

1. *Introduction and Preparation*
2. *Analysis with Binary Variables (Chp 1-2)*
3. *Analysis with Multicategory Variables (Chp 3)*
4. *Analsyis with Count Response (Chp 4)*
5. *Model Selection and Evaluation (Chp 5)*
6. **Additional Topics (Chp 6)**
   - *6.1 Exact Inference (Chp 6.2)*
   - *6.2 Revisit to Loglinear and Logistic Models for Contingency Tables: the Loglinear-Logit Connection*
   - **6.3 Revisit III to GLM and Advanced Topics (Chp 5.3, Chp 6.5)**
     - *6.3.1 Revisit III to GLM*
     - **6.3.2 Marginal Modeling: Quasi-Score, Generalized Estimating Equation (GEE)**
     - **6.3.3 Mixed Effect Models for Correlated Data**