Discrete Response Model Lecture 5

Models for Count Response, Discrete Response Model Evaluation, and Model Selection

datascience@berkeley

Poisson Probability Model

Count Response Data

Count responses can also arise from other mechanisms that have nothing to do with Bernoulli trials. Examples include:

- The number of credit cards an individual owns
- The number of arrests for a city per year
- The number of people arriving at an airport on a given day
- The number of cars stopped at the 33rd and Holdrege streets intersection
- The number of people standing in line at a specific Starbucks between 6 AM and 7 PM

For these settings, a Poisson distribution can be used to model the count responses.

Poisson PMF

Poisson PMF:

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

for y = 0, 1, 2, …, where

Y is a random variable
y denotes the possible outcomes of Y
μ is a parameter that is greater than 0

We often write Y ~ Po(μ) as shorthand notation
to mean that Y has a Poisson distribution with parameter μ.

Y will denote the number of occurrences of an event.

Properties

Below are characteristics of a Poisson distribution and associated items of interest:

- One can show the mean and variance of Y to be:

  $E(Y) = \mu$ and $Var(Y) = \mu$

  Having the mean and variance BOTH equal to μ is nice, but this is also limiting. Often, the actual variability observed in a sample is GREATER than μ.

  This is referred to as overdispersion.

  If $Y_1, \ldots, Y_n$ are independent with distribution $Po(\mu)$, then $\sum_{k=1}^{n} Y_k \sim Po(\sum_{k=1}^{n} \mu = n\mu)$.

  If each Y_k had a different mean, we would have $\sum_{k=1}^{n} Y_k \sim Po(\sum_{k=1}^{n} \mu_k)$

Properties

Likelihood function:

$$L(\mu; y_1 \ldots, y_n) = \prod_{k=1}^{n} \frac{e^{-\mu} \mu^{y_k}}{y_k!}$$

MLE for μ is $\hat{\mu} = n^{-1} \sum_{k=1}^{n} y_k$, i.e., the sample mean

The estimated variance for $\hat{\mu}$ is

$$\mathrm{Var}(\hat{\mu}) = -\left[E\left(\frac{\partial^2 \log[L(\mu \mid Y_1, \ldots, Y_n)]}{\partial \mu^2} \right) \right]^{-1} \Bigg|_{\mu=\hat{\mu}} = \frac{\hat{\mu}}{n}$$

Hypothesis Testing and Score CI for μ

$H_0: \mu = \mu_0$ vs. $H_a: \mu \neq \mu_0$

Wald test statistic: $$Z_0 = \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\mu} / n}}$$

Score test statistic: $$Z_0 = \frac{\hat{\mu} - \mu_0}{\sqrt{\mu_0 / n}}$$

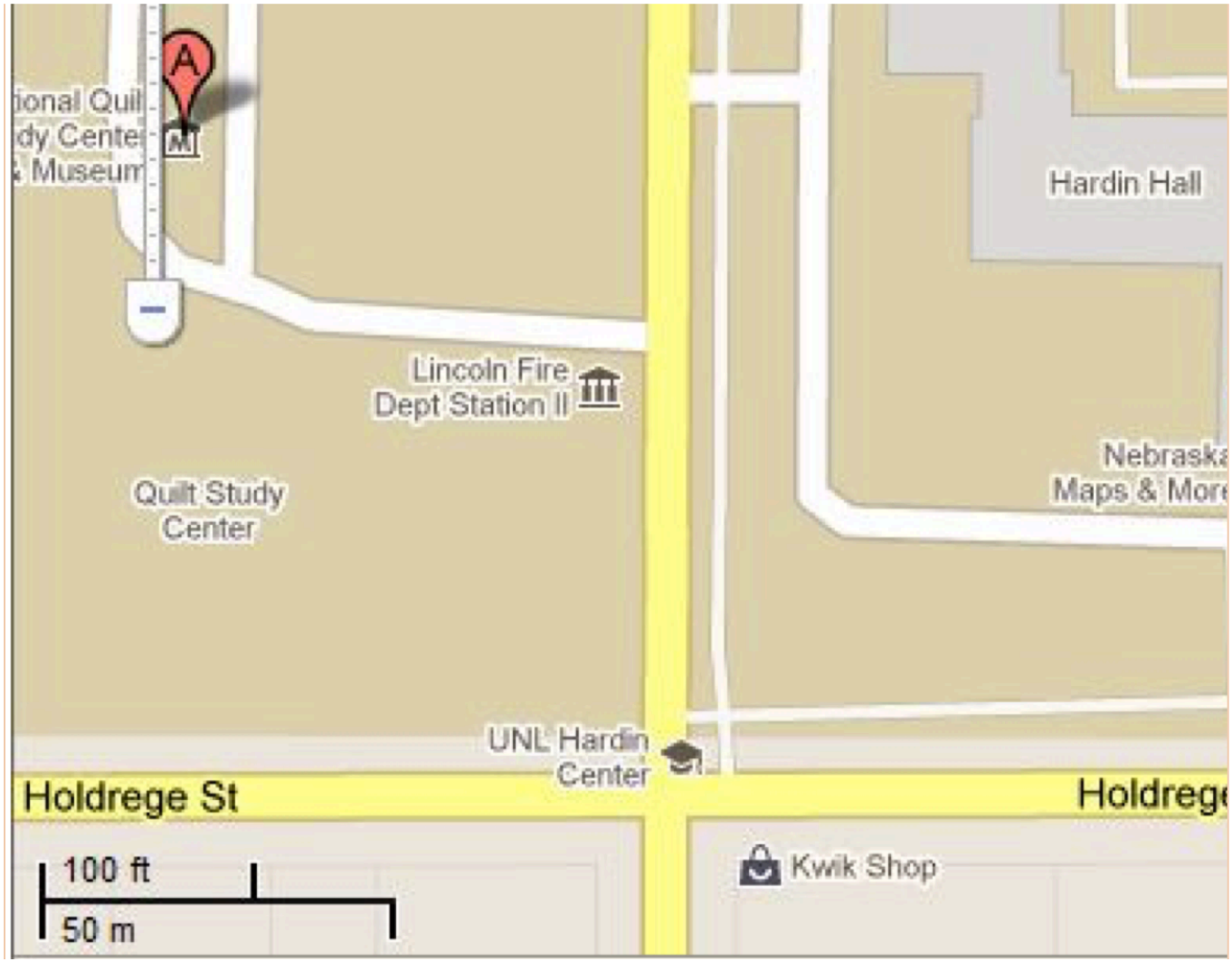
Score CI for μ

$$\left(\hat{\mu} + \frac{Z^2_{1-\alpha/2}}{2n} \right) \pm Z_{1-\alpha/2} \sqrt{\frac{\hat{\mu} + Z^2_{1-\alpha/2} / 4n}{n}}$$

An Example

Example: 33rd and Holdrege Streets

The intersection at 33rd and Holdrege Streets is a typical north-south/east-west, four-way intersection.



Example

A picture taken from a location northwest of the intersection.

- Approximately 150 feet north of the intersection is a fire station located on the west side of the street.

- A back-up of vehicles at the stoplight waiting to go south could block the fire station's driveway, which would prevent emergency vehicles from exiting the station.
- **Question: What is the probability that this could happen?**

Example

To examine this more closely, a sample of **40 consecutive stoplight cycles** from 3:25 PM to 4:05 PM on a non-holiday weekday was taken, and the number of vehicles stopped at the stoplight going south were counted.

```
> str(stoplight)
'data.frame':   40 obs. of  2 variables:
 $ Observation: int  1 2 3 4 5 6 7 8 9 10 ...
 $ vehicles   : int  4 6 1 2 3 3 0 4 2 8 ...
> head(stoplight)
  Observation vehicles
1           1        4
2           2        6
3           3        1
4           4        2
5           5        3
6           6        3
```

Note that there were no vehicles remaining in the intersection for more than one stoplight cycle. Why is this important to know?

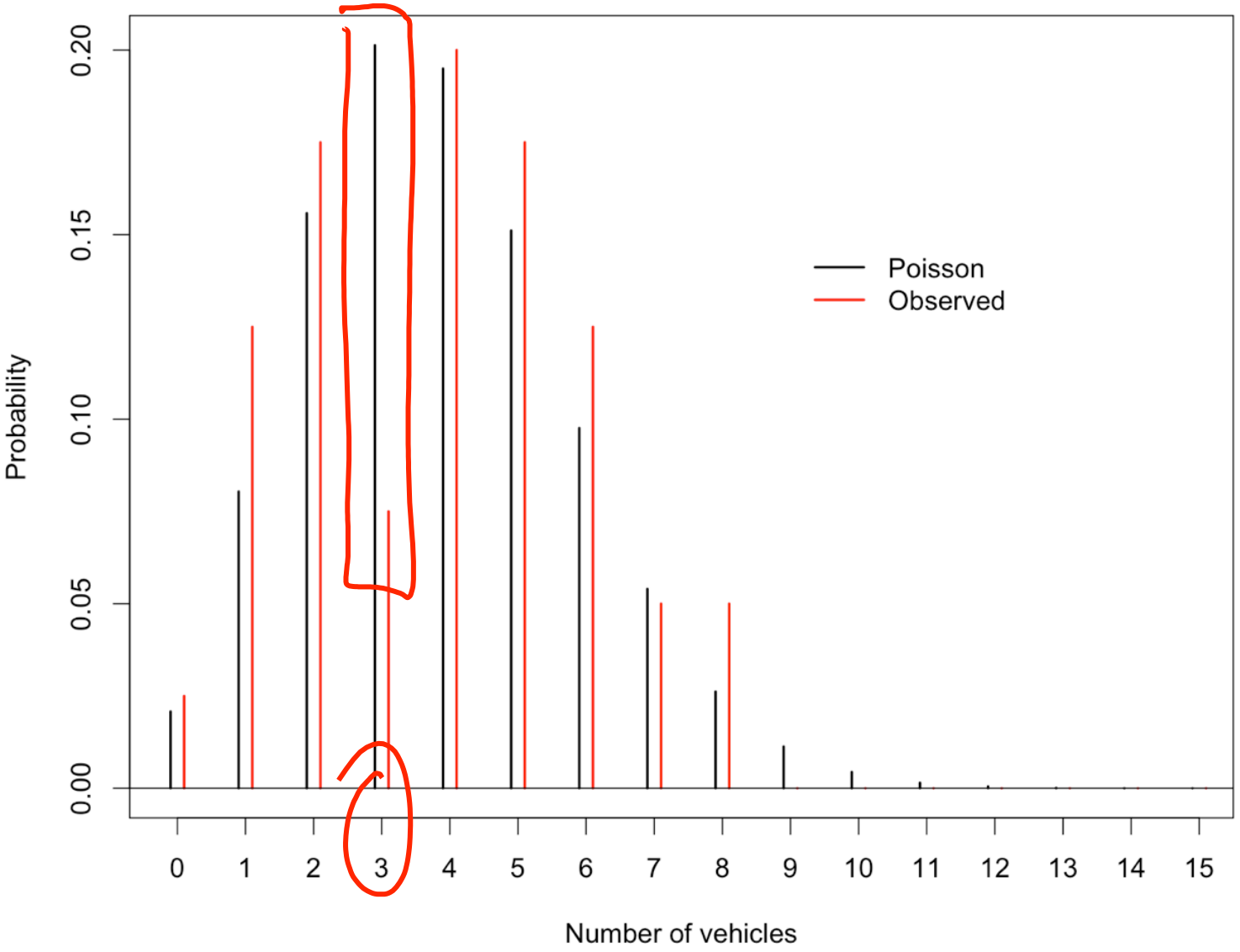
An Example (continue)

Example

```
> mean(stoplight$vehicles)
[1] 3.875
> var(stoplight$vehicles)
[1] 4.317308
> table(stoplight$vehicles) #Note that y = 0, 1, ..., 8 all have positive counts

0 1 2 3 4 5 6 7 8
1 5 7 3 8 7 5 2 2
```

```
> rel.freq <- table(stoplight$vehicles)/length(stoplight$vehicles)
> rel.freq2 <- c(rel.freq, rep(0, times = 7))
> y <- 0:15
> prob <- round(dpois(x = y, lambda = mean(stoplight$vehicles)), 4)
> data.frame(y, prob, rel.freq = rel.freq2)
    y   prob rel.freq
1   0 0.0208    0.025
2   1 0.0804    0.125
3   2 0.1558    0.175
4   3 0.2013    0.075
5   4 0.1950    0.200
6   5 0.1511    0.175
7   6 0.0976    0.125
8   7 0.0540    0.050
9   8 0.0262    0.050
10  9 0.0113    0.000
11 10 0.0044    0.000
12 11 0.0015    0.000
13 12 0.0005    0.000
14 13 0.0001    0.000
15 14 0.0000    0.000
16 15 0.0000    0.000
```

Example


Probability
0.00
0.05
0.10
0.15
0.20
Poisson
Observed
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of vehicles

Example

Wald confidence interval

```
> mu.hat <- mean(stoplight$vehicles)
> mu.hat + qnorm(p = c(alpha/2, 1 - alpha/2))*sqrt(mu.hat/n)
[1] 3.264966 4.485034
```

Note that the Wald interval using the log(μ) transformation is

$$e^{\log(\hat{\mu}) \pm Z_{1-\alpha/2}\sqrt{1/(\hat{\mu}n)}}$$

Exponentiate the log() transformation.

```
> exp(log(mu.hat) + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(1/(mu.hat*n)))
[1] 3.310561 4.535674
```

Score C.I.

```
> (mu.hat + qnorm(p = c(alpha/2, 1 - alpha/2))/(2*n)) + qnorm(p = c(alpha/2, 1 - alpha/2
)) * sqrt((mu.hat + qnorm(p = 1 - alpha/2)/(4*n))/n)
[1] 3.239503 4.510497
```

Poisson Regression Model: Model for Mean: Log-Link

Mean of the Poisson Distribution

Suppose the mean parameter of a Poisson distribution is now dependent on a function of explanatory variables. For example, suppose there is only one explanatory variable x. We could represent this dependence by

$$\mu = \beta_0 + \beta_1 x$$

Depending on the value of the parameters and x, we could obtain a negative value for μ which would not make sense for a count! Instead, we can use

$$\log(\mu) = \beta_0 + \beta_1 x$$

which alternatively can be written as

$$\mu = \exp(\beta_0 + \beta_1 x)$$

Now, μ is guaranteed to be greater than 0. This is referred to a Poisson regression model.

Mean as a Function of Explanatory Variables

When needed, we can emphasize that the mean changes as a function of the variable x for the i^{th} observation with

$$\mu_i = \exp(\beta_0 + \beta_1 x_i)$$

If there are p explanatory variables, we can write the model as

$$\mu = \exp(\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p)$$

or

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

Generalized Linear Model (GLM)

A Poisson regression model is a generalized linear model with the following components:

1. Random: Y has a Poisson distribution
2. Systematic: $\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$
3. Link: log

- A consequence of the log link function is that the explanatory variables affect the response mean in multiplicative way

Parameter Estimation and Inference

Maximum Likelihood

Maximum likelihood estimation is used again to find the MLEs. Suppose my sample is denoted as $(y_i, x_{i1}, \ldots, x_{ip})$ with $i = 1, \ldots, n$. The likelihood function is

$$L(\beta_0, \quad, \beta_p \mid y_1, \quad, y_n) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

where $\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \quad + \beta_p x_{ip})$.

For most situations, the likelihood function needs to be maximized using iterative numerical procedures. The glm() function in R completes this maximization where the family argument needs to be given as poisson(link = log).

- The covariance matrix for the parameter estimators follows from using standard likelihood procedures as outlined in the book's Appendix B.

- Wald and LR-based inference methods are performed in the same ways as for likelihood procedures in earlier weeks

Example: Horseshoe Crabs

The purpose of this example is to determine if the shell width of a female (x) is related to the number of satellites (Y) she has around her.

$$\log(\mu) = \beta_0 + \beta_1 x$$

where

Y = Number of satellites
x = Shell width of female (measured in cm)

can be used to estimate the mean number of satellites given a shell width.

```
> crab <- read.csv(file  = "HorseshoeCrabs.csv")
> str(crab)
'data.frame':   173 obs. of  5 variables:
 $ Color : int  2 3 3 4 2 1 4 2 2 2 ...
 $ Spine : int  3 3 3 2 3 2 3 3 1 3 ...
 $ Width : num  28.3 26 25.6 21 29 25 26.2 24.9 25.7 27.5 ...
 $ Weight: num  3.05 2.6 2.15 1.85 3 2.3 1.3 2.1 2 3.15 ...
 $ Sat   : int  8 4 0 0 1 3 0 0 8 6 ...
> head(crab)
  Color Spine Width Weight Sat
1     2     3  28.3   3.05   8
2     3     3  26.0   2.60   4
3     3     3  25.6   2.15   0
4     4     2  21.0   1.85   0
5     2     3  29.0   3.00   1
6     1     2  25.0   2.30   3
```

Model Estimation and Estimation Results

```
> mod.fit<-glm(formula = Sat ~ Width, data = crab,
+              family = poisson(link = log))
> summary(mod.fit)

Call:
glm(formula = Sat ~ Width, family = poisson(link = log), data = crab)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8526  -1.9884  -0.4933   1.0970   4.9221

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224  -6.095  1.1e-09 ***
Width        0.16405    0.01997   8.216  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 567.88  on 171  degrees of freedom
AIC: 927.18

Number of Fisher Scoring iterations: 6
```

The estimated Poisson regression model is

$$\hat{\mu} = \exp(-3.3048 + 0.1640x)$$

The model could also be written as:

$$\log(\hat{\mu}) = -3.3048 + 0.1640x$$

Parameter Estimation and Inference

Model Interpretation

- We are no longer modeling the log-odds anymore!
- We will not use odds ratios to interpret the effect of

Consider a model with one explanatory variable again:

$$\mu(x) = \exp(\beta_0 + \beta_1 x)$$

where I am using "$\mu(x)$" here to emphasize we are evaluating the model at a particular numerical value of x. The model evaluated at a c-unit increase in the explanatory variable is

$$\mu(x+c) = \exp(\beta_0 + \beta_1(x + c))$$

Model Interpretation

- Consider a Poisson regression model with one explanatory variable $\mu(x) = exp(\beta_0 + \beta_1 x)$.
- Increase the explanatory variable by c units, and the result is $\mu(x + c) = exp(\beta_0 + \beta_1(x + c)) = \mu(x)exp(c\beta_1))$.
- The ratio of the means at $x + c$ and x is

$$\frac{\mu(x + c)}{\mu(x)} = \frac{exp(\beta_0 + \beta_1(x + c))}{exp(\beta_0 + \beta_1 x)} = exp(c\beta_1)$$

- The interpretation is that "the percentage change in the mean response resulting from a c unit change in x is $100(e^{c\beta_1} - 1)$"
- As an example, if $exp(c\beta_1) = 1.1$, then it means that the percentage change in the mean response resulting from a c unit change in x is 10%
- Note that this interpretation is not dependent on the original value of x

Model Result Interpretation

```
> summary(crab)
     Color           Spine           Width           Weight           Sat
 Min.   :1.000   Min.   :1.000   Min.   :21.0   Min.   :1.200   Min.   : 0.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:24.9   1st Qu.:2.000   1st Qu.: 0.000
 Median :2.000   Median :3.000   Median :26.1   Median :2.350   Median : 2.000
 Mean   :2.439   Mean   :2.486   Mean   :26.3   Mean   :2.437   Mean   : 2.919
 3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:27.7   3rd Qu.:2.850   3rd Qu.: 5.000
 Max.   :4.000   Max.   :3.000   Max.   :33.5   Max.   :5.200   Max.   :15.000
```

```
Call:
glm(formula = Sat ~ Width, family = poisson(link = log), data = crab)
```

The expected number of satellites when the shell width is 23:

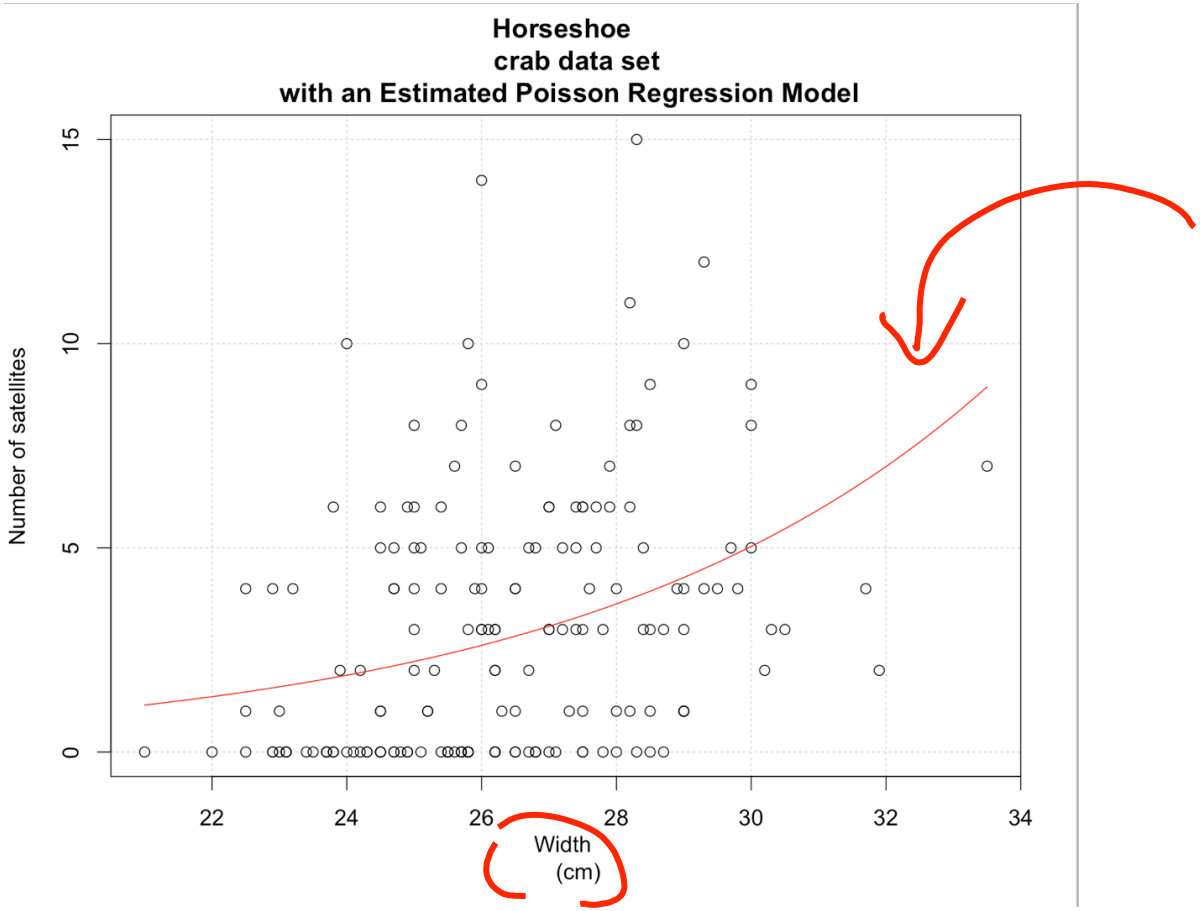
$$\hat{\mu} = \exp(-3.3048 + 0.1640 \times 23) = 1.60$$

- The positive slope indicates that the number of satellite increases with shell width
- So, a 1-unit increase in shell width leads to an 17.8% estimated increase in the number satellites

Example

When there is only one explanatory variable in the model, we can easily examine the estimated model through a plot.

When there are more than one explanatory variables, we will have to make the plot conditional on specific values on all the other variables in the model.



Example

```
#Function to find confidence interval
ci.mu<-function(newdata, mod.fit.obj, alpha) {
    lin.pred.hat<-predict(object = mod.fit.obj, newdata =
      newdata, type = "link", se = TRUE)
    lower<-exp(lin.pred.hat$fit - qnorm(1 - alpha/2) *
      lin.pred.hat$se)
    upper<-exp(lin.pred.hat$fit + qnorm(1 - alpha/2) *
      lin.pred.hat$se)
    list(lower = lower, upper = upper)
  }
```

```
Browse[1]> ci.mu(newdata = data.frame(Width = 23), mod.fit.obj =
+     mod.fit, alpha = 0.05)
$lower
       1
1.332135

$upper
       1
1.915114
```

Profile Likelihood Ratio Confidence Interval

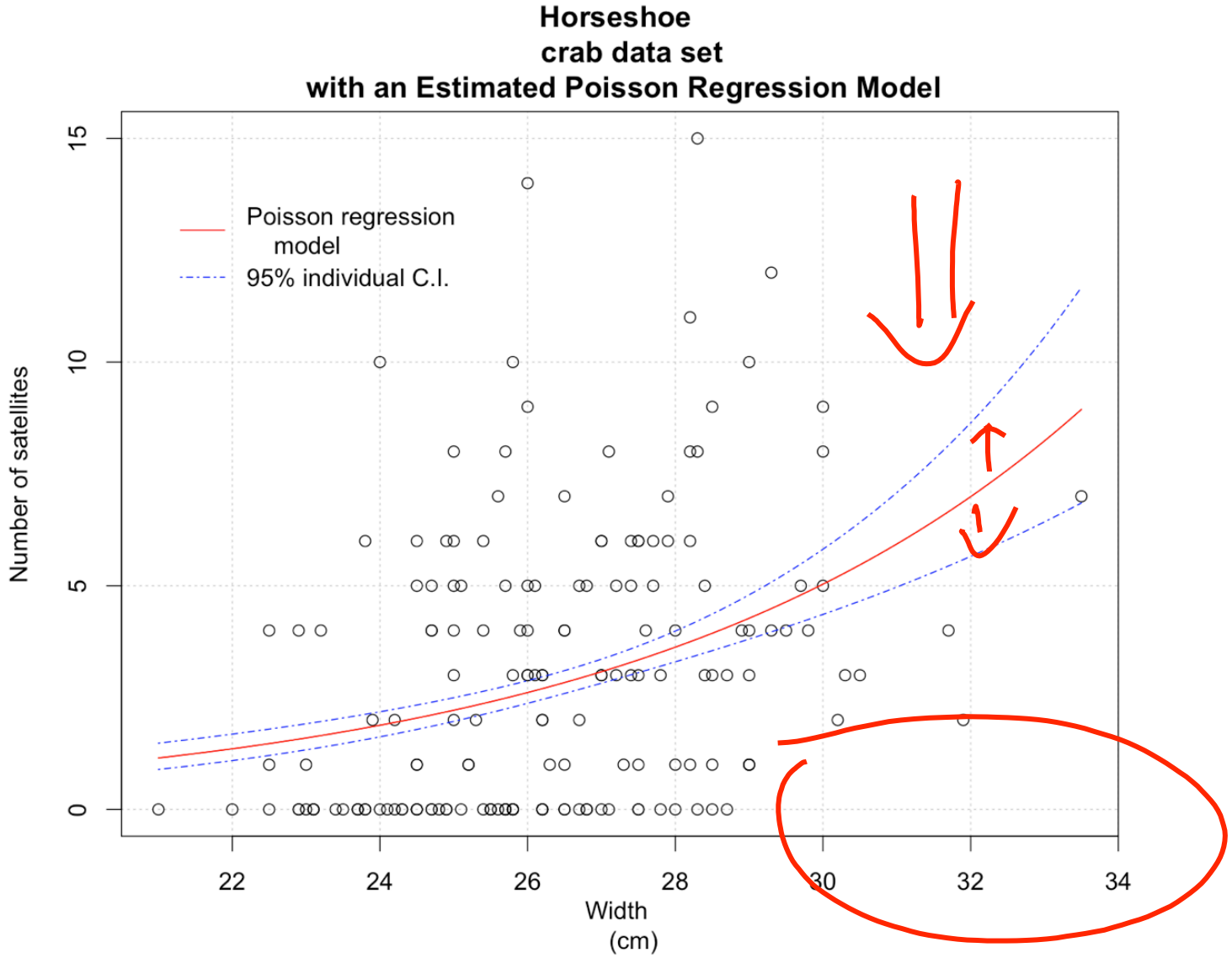
```
library(mcprofile)
linear.combo<-mcprofile(object = mod.fit, CM = K)
#CI for beta_0 + beta_1 * x
ci.logmu.profile<-confint(object = linear.combo, level = 0.95)
ci.logmu.profile
```

```
   mcprofile - Confidence Intervals

level:            0.95
adjustment:       single-step

   Estimate lower upper
C1    0.468 0.284 0.647
```

```
> ci.logmu.profile$confint
      lower     upper
1 0.2841521 0.6471587
> exp(ci.logmu.profile)

   mcprofile - Confidence Intervals

level:            0.95
adjustment:       single-step

   Estimate lower upper
C1      1.6  1.33  1.91
```

The 95% interval is **1.33 < μ < 1.91**, which is quite similar to the Wald interval.

Estimated Model and Confidence Bands



Binning an Explanatory Variable

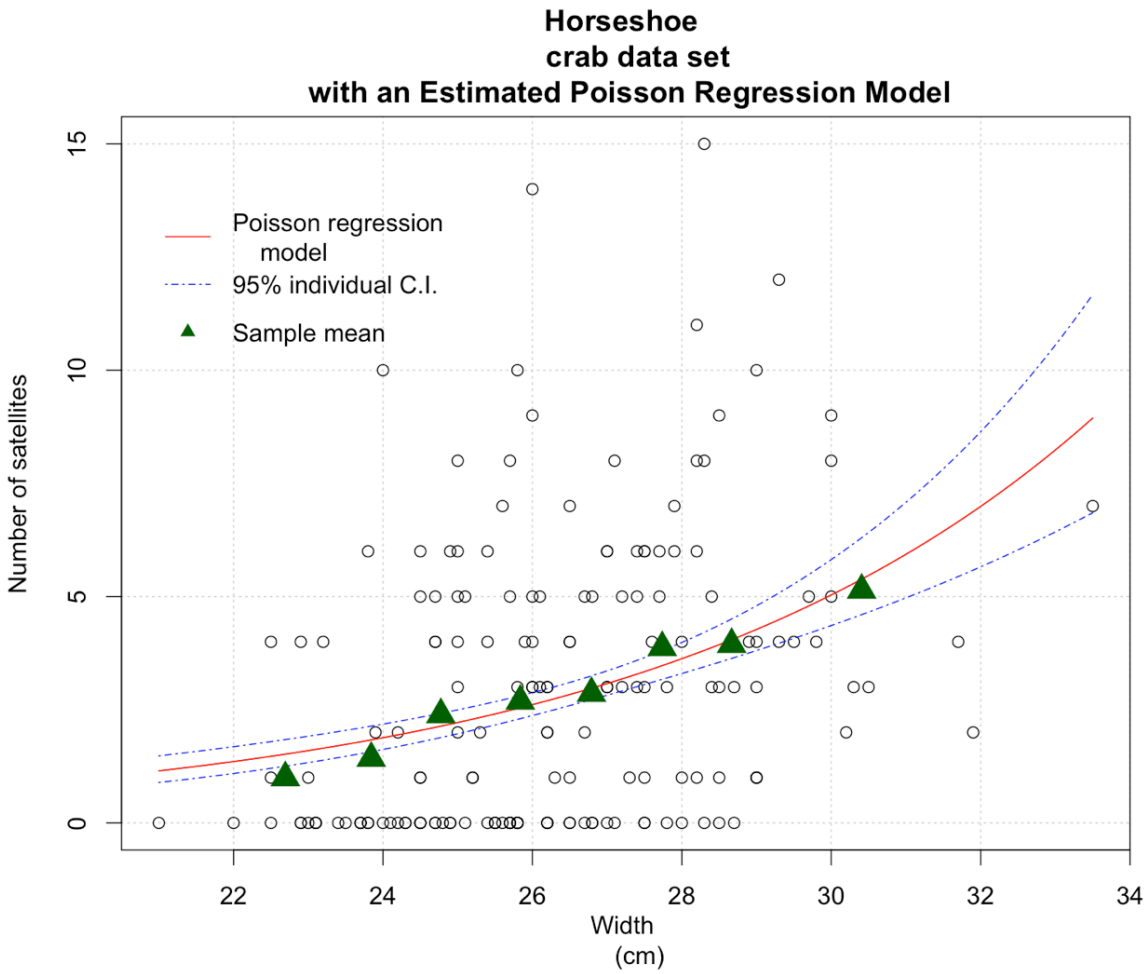
- The data show somewhat of an upward trend. The model captures this through displaying similar qualities.

- You may be alarmed by the number of plotting points far from the estimated model. However, remember that the model is **trying to estimate the "average" number of satellites given the width.**
- We can examine this more closely where we add the average number of satellites for a "width group."

```
Browse[1]> groups<-ifelse(test = crab$Width<23.25, yes = 1, no =
+         ifelse(test = crab$Width<24.25, yes = 2, no =
+         ifelse(test = crab$Width<25.25, yes = 3, no =
+         ifelse(test = crab$Width<26.25, yes = 4, no =
+         ifelse(test = crab$Width<27.25, yes = 5, no =
+         ifelse(test = crab$Width<28.25, yes = 6, no =
+         ifelse(test = crab$Width<29.25, yes = 7, no = 8)))))))
Browse[1]> crab.group<-data.frame(crab,groups)
Browse[1]> head(crab.group)
  Color Spine Width Weight Sat groups
1     2     3  28.3   3.05   8      7
2     3     3  26.0   2.60   4      4
3     3     3  25.6   2.15   0      4
4     4     2  21.0   1.85   0      1
5     2     3  29.0   3.00   1      7
6     1     2  25.0   2.30   3      3
```

```
Browse[1]> ybar<-aggregate(formula = Sat ~ groups, data = crab, FUN = mean)
Browse[1]> xbar<-aggregate(formula = Width ~ groups, data = crab, FUN = mean)
Browse[1]> data.frame(ybar, xbar$Width)
  groups      Sat xbar.Width
1      1 1.000000   22.69286
2      2 1.428571   23.84286
3      3 2.392857   24.77500
4      4 2.692308   25.83846
5      5 2.863636   26.79091
6      6 3.875000   27.73750
7      7 3.944444   28.66667
8      8 5.142857   30.40714
```

Comparing the model mean with the the Average" number of satellites given the width



Notice how the red line goes through the middle of the green triangles (group means).

Final Recap

- This interpretation is not dependent on the original value of x!
- Choose a value of c appropriate for the data.
- The estimate of $100(e^{c\beta_1} - 1)$ is $100(e^{c\hat{\beta}_1} - 1)$.
- Wald and LR confidence intervals can be found using the usual methods.
- If there is more than one explanatory variable in the model, the same result holds but need to included the "conditional" interpretation.
- If there are interactions or transformations of explanatory variables or categorical explanatory variables, similar types of adjustments need to be made as we introduced in previous week.
- The students are referred to the textbook for more information about Poisson regression models.

Variable Selection

Stepwise Selection

- **Stepwise methods** for variable selection:
  - Use with caution.
  - For datasets with not too many variables (say, no more than a couple hundred), doing EDA is important.
  - For datasets with thousands-plus variables, a selection method is likely needed.
  - Always remember that theory, subject matter knowledge, and contextual information are important.
  - More details are covered in the text.

- Notice that all of these variable selection methods assume a "given a set of variables." In practice, it is common to create additional variables.
- As such, when building a model, one may have to:
  1. Examine the given set of variables.
  2. Consider various transformations of a selected set of variables.
  3. Consider create additional variables.
  4. Select a set of variables among the given, transformed, and the created variables.

LASSO

- The least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)) has evolved since.
  - Basic idea: Add a penalty to the log-likelihood function and then maximize it to obtain estimates.
  - This penalty is chosen to help extenuate the effects of those explanatory variables that are truly important, while keeping parameter estimates close to 0 for those parameters that are not truly important.
  - The model with the smallest residual deviance is considered to the "best."

The LASSO parameters estimate $\hat{\beta}_{0,LASSO}, \hat{\beta}_{1,LASSO}, \ldots, \hat{\beta}_{p,LASSO}$ maximize

$$log\left(L\left(\beta_0, \beta_1, \ldots, \beta_p | y_1, \ldots, y_n\right)\right) - \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ is a .

Model Evaluation

Model Assumptions

- Any statistical model comes with a set of statistical assumptions.
- When estimating a GLM, we make the following assumptions:

1. The data are generated independently by an identical distribution we specify, be that binomial, Poisson, and so on.
2. The mean of the distribution is linked to the explanatory variables by the "link" function we specify.
3. The link relates to the explanatory variables in a linear fashion.

On Residuals

- "Raw" residuals, the difference between observed and predicted values, are not useful to evaluate the underling assumptions of GLM and count response models because the residuals of these models depend on their means.
- As such, we introduce the concepts of Pearson residuals and standardized Pearson residuals
- The idea is to account for the variance related to the response variables, be that categorical or count response.

$$y_i - \hat{y}_i$$

$$e_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{Var}(Y_i)}}$$

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{Var}\left(Y_i - \hat{Y}_i\right)}} = r_m = \frac{y_i - \hat{y}_i}{\sqrt{\hat{Var}(Y_i -)(1 - h_i)}},$$

where $i = 1, \ldots, n$ and h_i is the i^{th} diagonal element of the hat matrix.

On Residuals

- Although there are other kinds of residuals, such as *deviance residuals*, we will use *standardized residuals*.
  - It is easy to compute.
  - It is easy to interpret: It can be viewed approximately as "observations from a standard normal distribution."
- Of course, we have to use this "approximation" with caution.
- Residuals from binomial models when there are only a small number of exploratory variables combinations are poorly approximated by a normal distribution.
  - The normal approximation is also not appropriate when is near 0 or 1.
  - A similar situation happens with Poisson model when the estimated means are very small.

Computing Residuals in R

- The generic function **residuals()** has methods functions that work on model-estimated objects produced by **glm()**.
- These functions use the **type** argument to select the type of residuals computed.
  - "pearson" or "response" are used to selected Pearson and Raw residuals.
- **Residuals.glm()** methods also has a "deviance" argument value to produce deviance residuals.
- Standardized Pearson and deviance residuals are avaiable for the glm-class objects using **rstandard()**.

Residual Diagnostic Assessment

- The information content contained in a set of residuals can be easily revealed in graphs.
- Many of these graphs are similar to those used in the context of classical linear regression models.
- However, interpretation in GLM can be different.
- Residuals in GLM can be used to diagnose problems in the conditional mean model, possible outliers, and model assumptions, such as the choice of the probability family, that are inappropriate.
- As mentioned above, residual plots of count response model is subject to the caveats already noted, that is, in situation where there are a limited number of possible responses.

Residual Diagnostic Assessment

1. Residuals against each of the explanatory variables

- A plot of standardized residuals against each explanatory variable can show whether the form for the explanatory variable is appropriate.
- The plot should show:
  - Same variance throughout the range of the explanatory variable
  - No serious fluctuations in the mean value
- Add a smoother, such as *loess* smoother, to the plot to add the visualization.
- Watch out for a clear pattern of curvature.
- In binomial models, the loess curve can be weighted by so that larger numbers of trails contribute relatively more to the curve placement than those with relatively few.
- Note: Loess curves are highly variable when data are sparse or near the extreme values.
  - ➔ Do not overreact to the changes in the curve at the edge of the plot.

Residual Diagnostic Assessment

2. Residuals against the fitted values (or in the binomial model)

- It is useful to examine if the link function is appropriate.
- The plot should have:
  - Constant variance throughout the range of the response
  - No clear curvature
- A plot of the residuals against the linear predictor,:
  - Shows patterns of change in the mean residuals more clearly
  - Help diagnose how the link function g(.) should be changed to better fit the data

Residual Diagnostic Assessment

3. Any of these plots should be used to check for extreme residuals

- As noted before, only about 5% of standardized residuals should be beyond +/-2 and typically none beyond +/-3.
- Presence of a large amount of extreme residuals indicates overdisperson, meaning that there is more variability to the counts than what the model assumes.
- It may also indicate that there are missing explanatory variables from the model.

Slight different in interpretation for binomial models:

- Extreme residuals occur more often in regions of high or low estimated probability of success.
- To account for this situation, calculate the binomial probability that such an extreme value of could occur trails with probability

Berkeley SCHOOL OF INFORMATION