Solutions

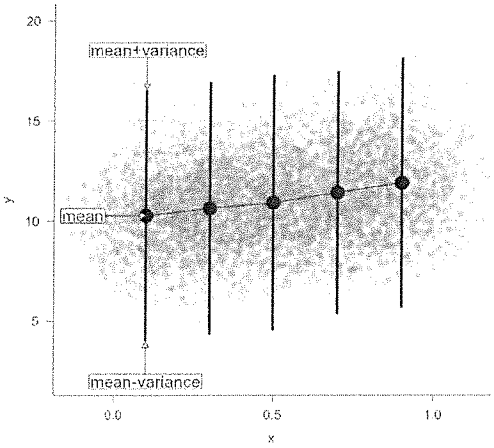
Midterm 2

S&DS 361

2023-04-04

1. Modeling

On the next 3 pages, there are 12 figures. In each figure, there is a scatter plot of y versus x. Points were jittered using `geom_jitter(height=0.1, width=0.2)`, so when points are slightly spread out around discrete values like 0, 1, 2, etc. (see for example Figures 3-5), assume they are exactly 0, 1, 2, etc. The big dots show the mean for different subsets of the predictor x. The vertical lines show the interval (mean − variance, mean + variance) for each subset. See the example figure below for a visual description.



Typically, the subsets are [0,2], (2,4], (4, 6], (6, 8], (8, 10], except in the cases where x is discrete, in which case the subsets are the individual discrete values.

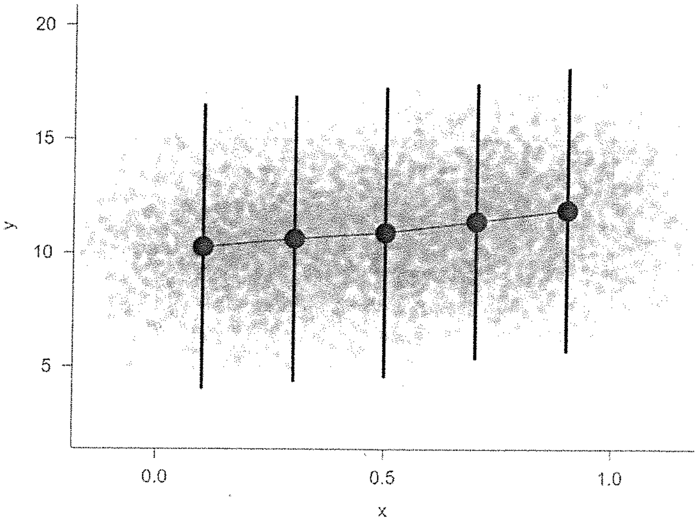
For each figure, use your understanding of the assumptions of the generalized linear models we discussed in class (listed below) to determine the appropriate model (or models) for the data shown in the figure. Next to each figure's title, write the letter (or letters) of the model (or models) you choose. Give a brief description of your answer in 10 words or fewer.

- A. Linear Regression
- B. Linear Regression with log transform of the outcome
- C. Poisson Regression
- D. Binary Logistic Regression
- E. Binomial Logistic Regression
- F. Negative Binomial Regression
- G. None of the above

3

Figure 1. Answer: A (Linear)

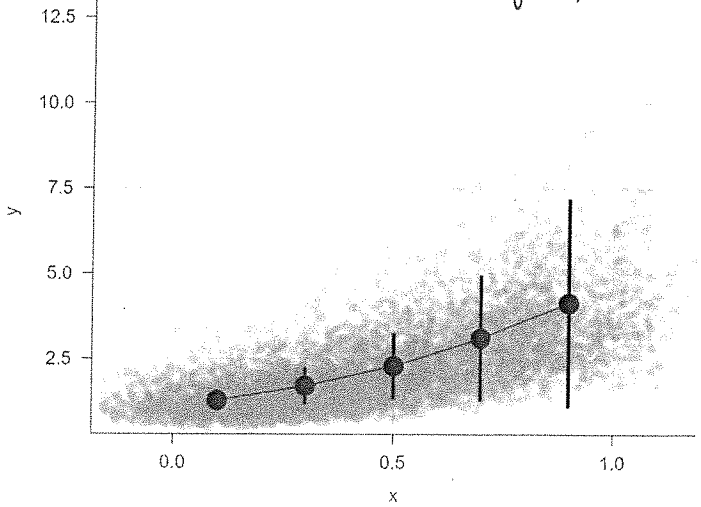
Explanation: Linear relationship, constant variance continuous outcome



3

Figure 2. Answer: B (Linear w/ Log)

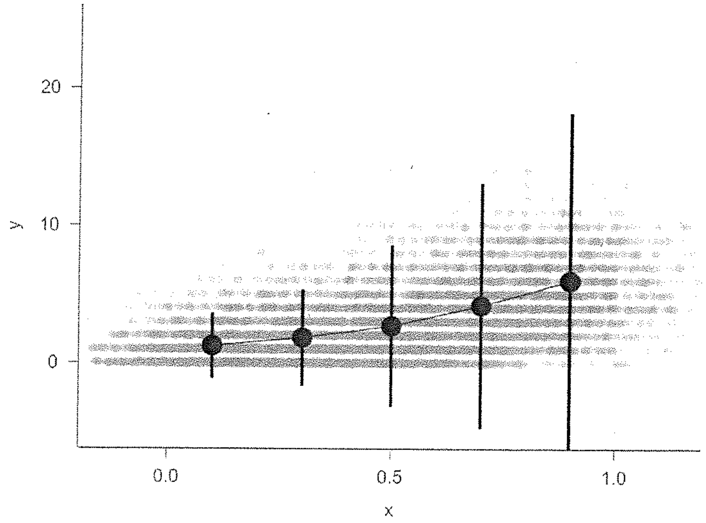
Explanation: exponential relationship, continuous outcome, increasing variance (so A isn't good)



3

Figure 3. Answer: F (Neg Bin)

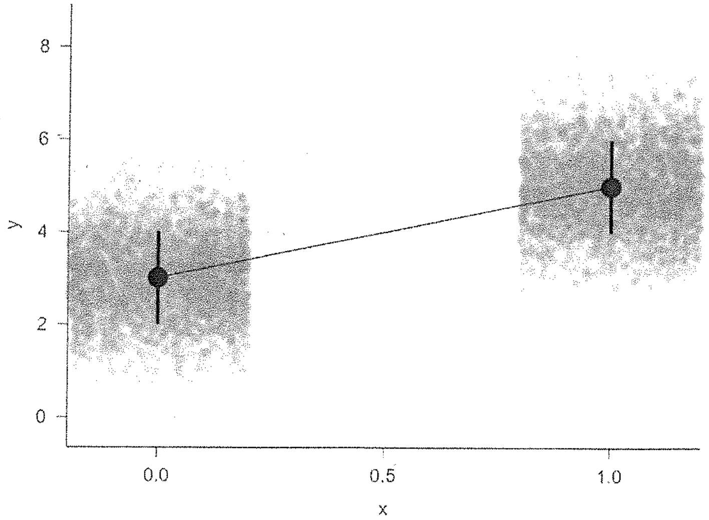
Explanation: Count outcome, variance > mean exponential relationship



3

Figure 4. Answer: A (Linear)

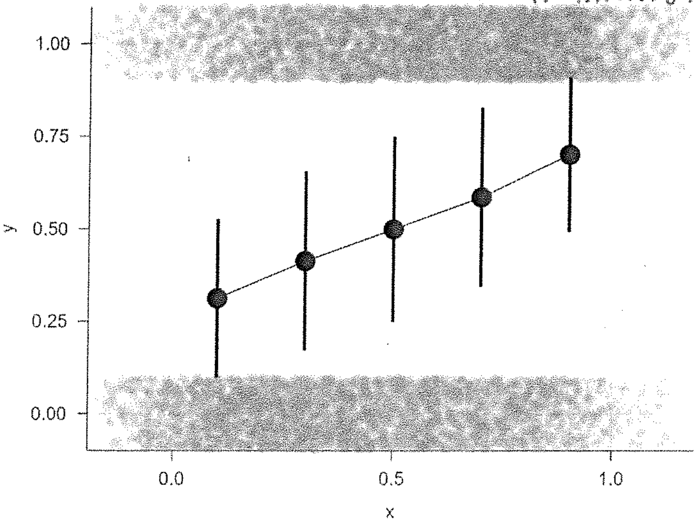
Explanation: Constant variance, continuous outcome



3

Figure 5. Answer: D (Binary Logistic)

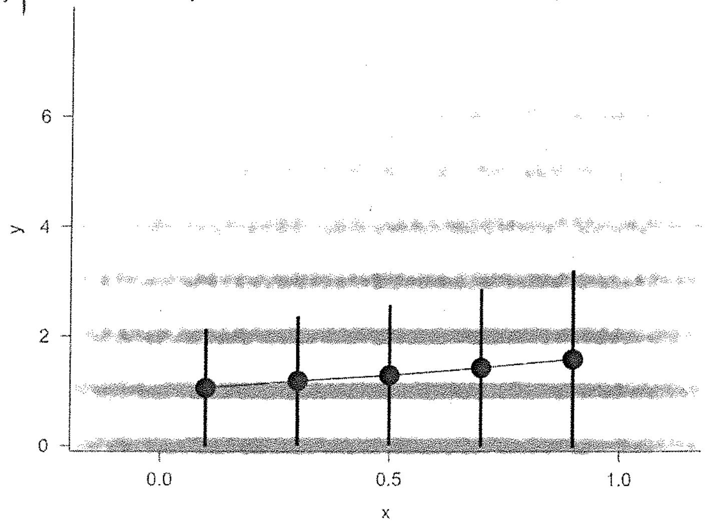
Explanation: Binary outcome, could reasonably be a logistic relationship (we are only seeing near the inflection point)



3

Figure 6. Answer: C (Poisson)

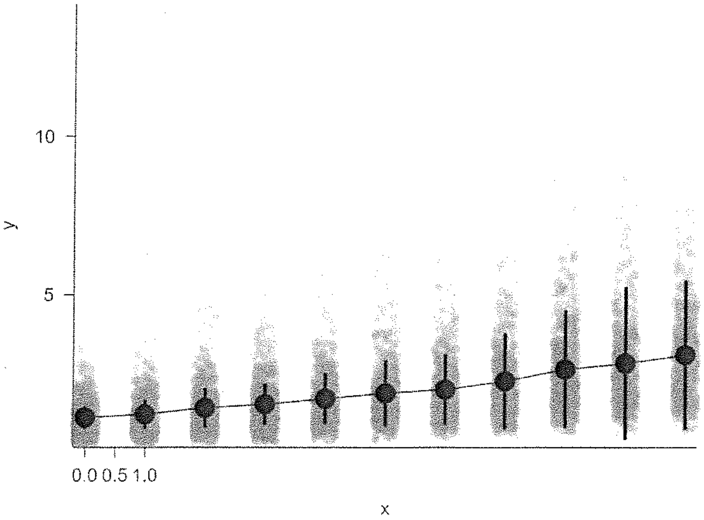
Explanation: Count outcome, mean = variance Exponential relationship



3

Figure 7. Answer: B (Linear w/ Log)

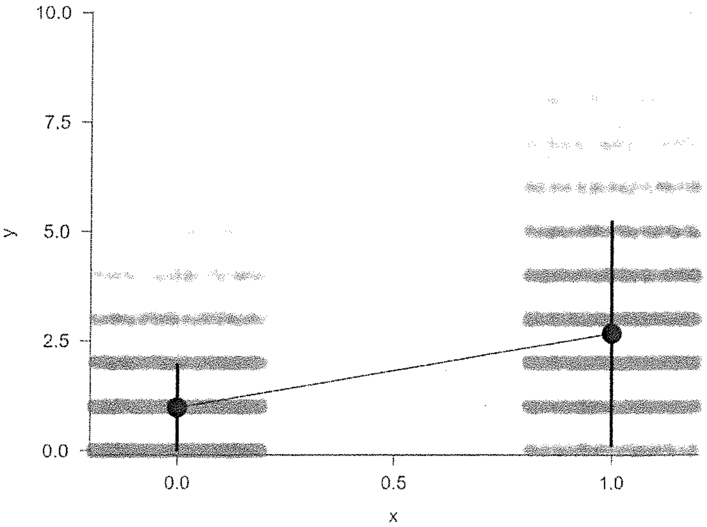
Explanation: Continuous outcome, non-constant variance, exponential relationship



3

Figure 8. Answer: C (Poisson)

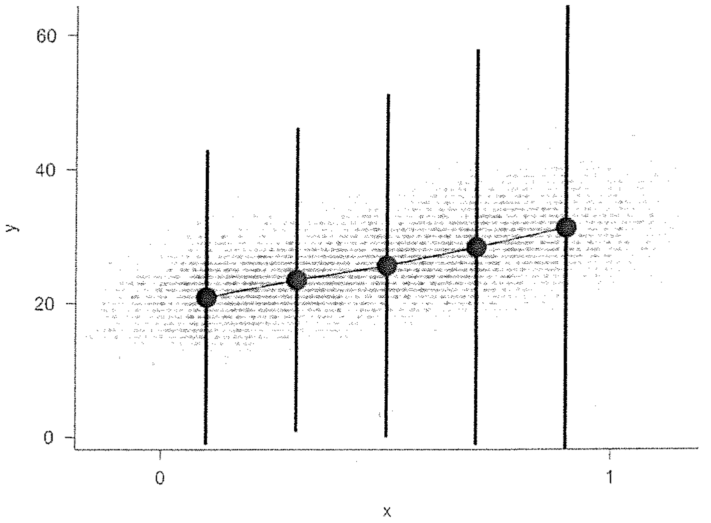
Explanation: Count outcome, mean = variance



3

Figure 9. Answer: C (Poisson)

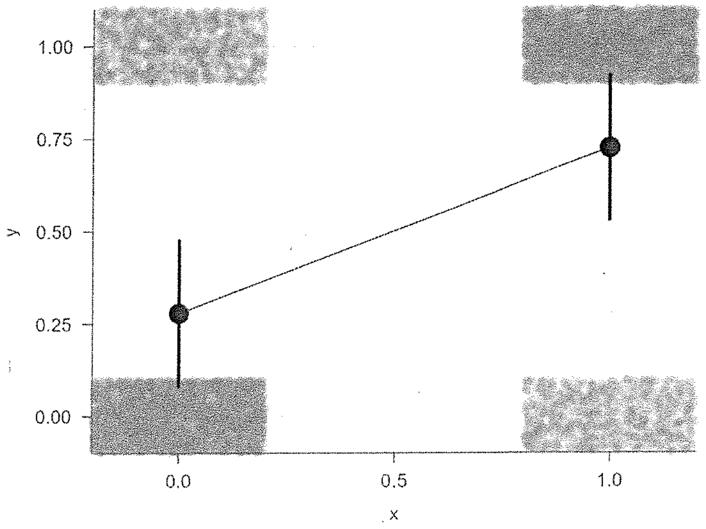
Explanation: Count outcome, mean = variance



3

Figure 10. Answer: D (Binary Logistic)

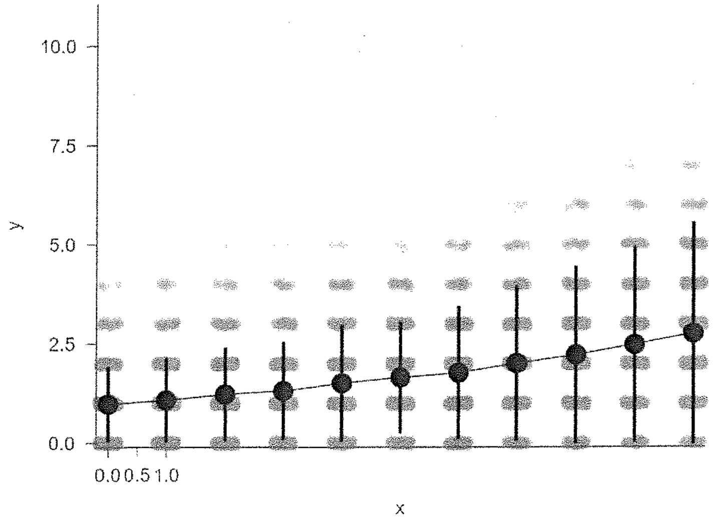
Explanation: Binary outcome



3

Figure 11. Answer: C (Poisson)

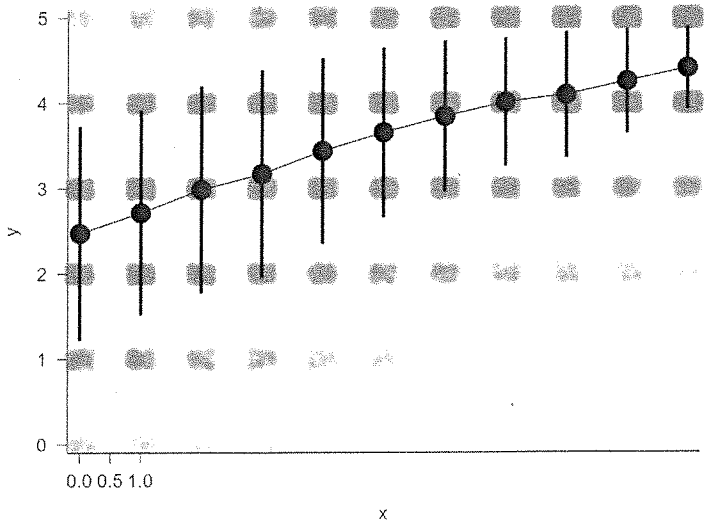
Explanation: Count Outcome, mean = variance



3

Figure 12. Answer: E (Binomial Logistic)

Explanation: Count outcome with clear max of 5, variance < mean, and variance decreasing



λ

2. Poisson

10 Consider the Poisson regression model with a single predictor x_1. Show why e^{β_1} is the percent change in y for a unit increase in x_1.

$$\frac{\lambda_{x+1}}{\lambda_x} = \frac{e^{\beta_0+\beta_1(x_1+1)}}{e^{\beta_0+\beta_1 x_1}} = \frac{\cancel{e^{\beta_0}}\,\cancel{e^{\beta_1 x_1}}\,e^{\beta_1}}{\cancel{e^{\beta_0}}\,\cancel{e^{\beta_1 x_1}}} = e^{\beta_1}$$

percent change

3. Logistic regression

10 a. Consider the logistic regression model with a single predictor x_1. Show why e^{β_1} is the odds ratio corresponding to a unit increase in x_1 (or percent change in odds for a unit increase in x_1.)

$$\log\left(\frac{p_{x+1}}{1-p_{x+1}}\right) = \beta_0+\beta_1(x+1)$$

$$-\ \log\left(\frac{p_x}{1-p_x}\right) = \beta_0+\beta_1 x_1$$

$$\log\left(\frac{p_{x+1}}{1-p_{x+1}}\right) - \log\left(\frac{p_x}{1-p_x}\right) = \boxed{\cancel{\beta_0}} + \cancel{\beta_1 x} + \beta_1 - (\boxed{\cancel{\beta_0}} + \cancel{\beta_1 x}) = \beta_1$$

$$\log\frac{\frac{p_{x+1}}{1-p_{x+1}}}{\frac{p_x}{1-p_y}} = \beta_1 \qquad \frac{\frac{p_{x+1}}{1-p_{x+1}}}{\frac{p_x}{1-p_x}} = e^{\beta_1}$$

odds ratio

10 b. Suppose $p = \frac{e^x}{1+e^x}$. Find the inverse by solving for x.

$$p(1+e^x) = e^x$$
$$p + pe^x = e^x$$
$$p = e^x - pe^x$$
$$p = (1-p)e^x$$

$$\frac{p}{1-p} = e^x$$
$$\log\frac{p}{1-p} = x$$

10 c. **Suppose we fit a logistic regression model with one predictor to some data, and estimate the coefficients** to be $\hat{\beta}_0 = 0$, $\hat{\beta}_1 = 1$. Use an example to show that, for a unit increase in x_1, **the difference in probability** is not constant but depends on the value of x_1.

$$p = \frac{e^x}{1+e^x}$$

$x=0$: $p = \frac{e^0}{1+e^0} = \frac{1}{2} = .5$

difference is .23

$x=1$: $p = \frac{e^1}{1+e^1} = .73$

difference is .15

$x=2$: $p = \frac{e^2}{1+e^2} = .88$

4. Exponential family form

10

Show that the negative binomial distribution, where r is known, is in the exponential family.

$$f(y;\theta) = e^{a(y)b(\theta) + c(y) + d(\theta)}.$$

$$f(y) = \binom{y+r-1}{r-1} p^r (1-p)^y \quad , \quad y = 0, 1, \ldots$$

$$e^{\log\left[\binom{y+r-1}{r-1} p^r (1-p)^y\right]} = e^{\log\binom{y+r-1}{r-1} + r\log p + y\log(1-p)}$$

$$= e^{\underset{\overset{\uparrow}{a(y)}}{y} \; \underset{\overset{\uparrow}{b(p)}}{\log(1-p)} + \underset{\overset{\uparrow}{c(y)}}{\log\binom{y+r-1}{r-1}} + \underset{\overset{\uparrow}{d(p)}}{r\log p}}$$

5. MLE

10

Write the likelihood for multiple linear regression with p predictors and show that the maximum likelihood estimates are the same as the least squares estimates. Assume σ is known.

$$\max_{\beta_j} \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}\right) \qquad \mu_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$$

$$\max_{\beta_j} \sum_{i=1}^{n} \underbrace{\log\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{constant}} - \underbrace{\frac{1}{2\sigma^2}}_{\text{constant}}(y_i-\mu_i)^2$$

$$\max_{\beta_j} -\sum_{i=1}^{n}(y_i-\mu_i)^2$$

$$\underbrace{\min_{\beta_j} \sum_{i=1}^{n}(y_i-\mu_i)^2}_{\text{same as minimizing SSE, aka same as Least Squares}}$$