Random	Assessment	Interpretation	Misc.
100	100	100	100
200	200	200	200
300	300	300	300
400	400	400	400
500	500	500	500

Subject 1	100

TRUE/FALSE: The variance of a gamma random variable increases cubically with the mean?

Question Answer	Done! Home

FALSE - it increases quadratically.

Subject 1	200

I can be thought of as the sum of Bernoulli trials. That is, I'm a count out of a total. What distribution do I have?

Question Answer	Done! Home

Binomial

So, if you identify a Binomially-distributed random variable, you would use code that looks something like:

proc genmod data = mydata;

model y/m = x1 x2 x3 / dist = binomial link = logit;

run;

Which of the following situations motivate the use of a GLM with a Binomial random component? Check all that apply.

1. The response tends to have a symmetric distribution around 0.
2. The response variable is a count that can be thought of as a sum of m Bernoulli trials.
3. The response can be any non-negative integer.
4. The variance appears to have a linear relationship with the mean.

1. ~~The response tends to have a symmetric distribution around 0.~~
2. The response variable is a count that can be thought of as a sum of m Bernoulli trials.
3. ~~The response can be any non-negative integer.~~
4. ~~The variance appears to have a linear relationship with the mean.~~

Here's our code.

```
proc genmod data = mydata;
model y = explanatory / dist = gamma link = log;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr >Ch
Intercept	1	−2.506	0.5446	21.22093	¡0.0
explanatory	1	0.0729	0.0418	3.041603	0.08

Provide a predicted value for Y if explanatory is equal to 39.

Log link.

$$\hat{\mu} = e^{-2.506+0.0729*39} = 1.400879$$

Subject 1

500

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Squar
Intercept		1	10.1446	0.0243	10.0967	10.1925	17389
sx	female	1	−0.0473	0.0162	−0.0792	−0.0153	8.5
sx	male	0	0.0000	0.0000	0.0000	0.0000	
rk	assistant	1	−0.4526	0.0214	−0.4947	−0.4106	449.4
rk	associate	1	−0.1968	0.0183	−0.2327	−0.1607	115.9
rk	full	0	0.0000	0.0000	0.0000	0.0000	
yd		1	0.0065	0.0009	0.0046	0.0083	48.3
Scale		1	112.5768	11.2410	91.9605	136.0716	

You fit a GLM with gamma random component and a log link to model professors' salaries. Provide a meaningful interpretation of the CI for the coefficient corresponding to female.

Question Answer

Done! Home

Subject 1

500

We're 95% confident that the mean salary for females is somewhere between exp(-.0792)=0.92 and exp(-0.0153)=0.98 times the mean salary for an otherwise identical male.

That is, female's salary is somewhere between 8% and 2% less than comparable males' salary.

For example, males for a particular rank + graduation year tended to make around $100,000, we'd expect the females to make between $92,000 and $98,000.

Assuming we have > 10 observations, will AIC or BIC favor larger models?

BIC: $-2l(\hat{\beta}) + log(n)(k+1)$, so $log(n)$ is the penalty for BIC.
AIC: $-2l(\hat{\beta}) + 2(k+1)$, so 2 is the penalty for AIC.
AIC has a smaller penalty (2 will always be less than log(n)), so it will tend to favor models with more explanatory variables (a larger k) than BIC.

Subject 2

200

You have just fit a GLM with a Binomial random component and logit link. List any/all test statistics you could potentially use to test for lack of fit.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	16	13.7526	0.8595
Pearson Chi-Square	16	13.5431	0.8464

Question
Answer

Done!
Home

Subject 2

200

For logistic regression (Binomial), we can use Pearson or deviance. So we can use d=13.7526 or P =13.5431. Recall, the null and alternative hyptheses have NOTHING to do with whether or not the betas are 0.

- H0: The binomial logistic regression is the true model
- HA: The binomial logistic regression is NOT the true model
- Use deviance test statistic, d = 13.7526
- null distribution is χ^2 distribution with $(n-(k+1))$ df
- p-value = (code below)
- `data test;`
- `pval = 1-CDF('CHISQUARE',13.7526,n-(k+1));`
- `run;`

Question Answer

Done! Home

Subject 2

300

You have just performed a Gamma regression with a log link. Provide a test statistic that can be used to test for lack of fit.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	47	0.6052	0.0129
Scaled Deviance	47	52.1007	1.1085
Pearson Chi-Square	47	0.6339	0.0135
Scaled Pearson X2	47	54.5672	1.1610
Log Likelihood		−480.5077	

Question Answer

Done! Home

Subject 2

300

For gamma random components, we only want to use the Pearson test statistic (never deviance). This can be found in the Scaled Pearson X2 row of the table. It is 54.5672. The null distribution is $\chi^2(n-(k+1))$. (This is always the null distribution when testing lack of fit.)
Similarly, we only want to look at pearson residuals. ex) if you find significant lack of fit in a Gamma model, your first step will be to plot the PEARSON standardized residual against predicted values to assess any problems with the mean.

Question
Answer

Done!
Home

DAILY DOUBLE - selector gets 1^{st} chance at $800

You have performed a Binomial logistic regression.
The standard error of $\hat{\beta}_1$ is 0.748 (e.g., $Var(\hat{\beta}_1)^{1/2}$ =0.748).

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	18	43.252	2.40
Pearson Chi-Square	18	46.26	2.57

You have reason to believe overdispersion is a problem so you account for it using the deviance-based estimate of ϕ. What is $Var(\hat{\beta}_1)$?

Subject 2

400

Using deviance,

$$\hat{\phi} = 43.252/18 = 2.40$$

We know that after accounting for overdispersion,

$$Var(\hat{\beta}_1) = \hat{\phi} 0.748^2 = 2.4 * 0.56 = 1.344$$

(Pearson would yield $Var(\hat{\beta}_1) = 2.57 * 0.56 = 1.4392$) Follow up: If we *do* account for overdispersion, what will/will not be affected?

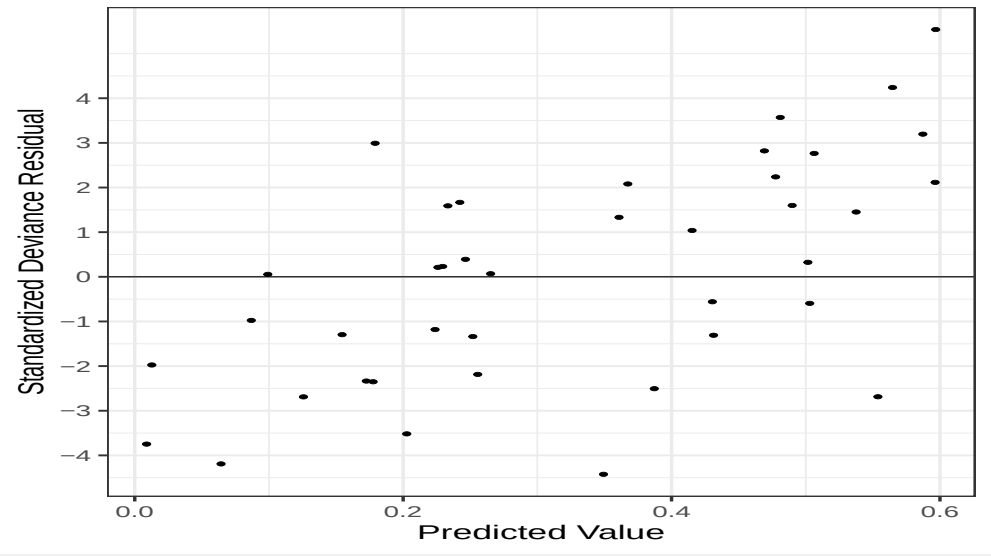
Question Answer

Done! Home

Subject 2

500

For a binomial logistic regression with a single numeric explanatory variable, this is a plot of standardized deviance residuals against predicted values. Make a statement about the estimated mean/prediction function as it relates to the data.



Question Answer

Done! Home

For low predicted values, the residuals tend to be negative. This indicates we're overpredicting in that area. For high predicted values, the residuals tend to be positive. This indicates we're underpredicting in that area. Perhaps our link function is a poor fit. In addition, there are extreme values we wouldn't expect if the model were correct.

Subject 3

100

Suppose Y_i is the count of females out of a nest of m_i turtle eggs, x_i is the incubation temperature. We assume:

$$Y_i \sim Binomial(m_i, \pi_i)$$

$$log(\frac{\pi_i}{1-\pi_i}) = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 x_{1i}$$

Provide a meaningful interpretation of β_1.

Question
Answer

Done!
Home

The odds of a given egg containing a female change by a factor of e^{β_1} for each unit increase in temperature.

A comprehensive discussion would quantify our uncertainty. Something along the lines of : This is our best estimate. We're 95% confident that the odds will change by a factor somewhere between exp(lower bound) and exp(upper bound). For example, they could increase by as little as 15% (for example) or as much as 27% (for example).

In addition, a comprehensive discussion would also incorporate statements about probability (e.g., the probability of a female incubated at 81 degrees is X while decreasing the temperature to 76 degrees results in a probability of a female changing to Y.)

Subject 3

200

Suppose Y_i is the count of females out of a nest of m_i turtle eggs, x_i is the incubation temperature. We assume:

$$Y_i \sim Binomial(m_i, \pi_i)$$

$$log(\frac{\pi_i}{1-\pi_i}) = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 x_{1i}$$

Suppose $\hat{\beta}_0 = -33, \hat{\beta}_1 = 0.46$. Provide a meaningful interpretation of $\hat{\beta}_1$ terms of a shift in temperature of 2 degrees.

Question Answer

Done! Home

Subject 3

200

The odds of a given egg containing a female increase by a factor of 2.51 for each two-degree increase in temperature. (The odds more than double!)

If you really want to engage your audience in a meaningful discussion, you would supplement this with a confidence interval, as we've done together.

In addition, you might drive home your point by calculating various probabilities: The probability of seeing a female at 72 degrees is 0.53. If it gets a little warmer - say 76 degrees - that probability increases to 0.87!

Question
Answer

Done!
Home

Subject 3

300

`hourly` is the hourly wage of employees and `age` is the age of the employees at a company. We have written the following code:

```
proc genmod data = mydata;
model hourly = age / dist = gamma link = identity;
run;
```

(WHITEBOARD) From this information, fully specify this model - the random component as well as the systematic component.

Question
Answer

Done!
Home

For $i = 1, \ldots, n$, let age_i represent the age of the i^{th} employee and Y_i be the random variable representing the hourly pay of the i^{th} employee Then,

Random component: $Y_i \sim Gamma(\mu_i, \sigma^2)$

Systematic component:

- Link Function
  (`link = identity` in code. this impacts β_1 interpretation)

$$\eta_i = g(\mu_i) = \mu_i$$

- Linear predictor

$$\eta_i = \beta_0 + \beta_1 age_i$$

Subject 3

400

We have written the following code:

```
proc genmod data = mydata;
model hourly = age / dist = gamma link = identity;
run;
```

Suppose the coefficient corresponding to age is estimated to be 1.13. Provide a meaningful interpretation of this estimate.

Question
Answer

Done!
Home

Subject 3

400

The mean hourly pay increases by $ 1.13 for each 1 year increase in age. Note: This is because of the identity link. Usually, we use a log-link with the Gamma random component, which requires us to exponentiate the estimate and interpret in a multiplicative way.

Question
Answer

Done!
Home

Subject 3

500

`mydata` has 235 observations. We have written the following code:

```
proc genmod data = mydata;
model y = explanatory / dist = gamma link = log;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr >ChiSq
Intercept	1	−2.506	0.5446	21.22093	¡0.0001
explanatory	1	0.0729	0.0418	3.041603	0.08123

Provide a 95% Wald-style confidence interval for the multiplicative effect of a 2-unit shift of explanatory on the mean response.

$z_{0.95} = 1.645, \quad z_{0.975} = 1.960, \quad t_{0.95,233} = 1.651, \quad t_{0.975,233} = 1.970$

Question
Answer

Done!
Home

Subject 3

500

A 95% Wald-style confidence interval for the coefficient corresponding to explanatory is $(0.0729 - 1.96 * (0.0418), 0.0729 + 1.96 * (0.0418))$. That is,

$$P(0.0729 - 1.96 * (0.0418) \leq \beta_1 \leq 0.0729 + 1.96 * (0.0418)) = 0.95$$

$$P(2*(0.0729-1.96*(0.0418)) \leq 2\beta_1 \leq 2*(0.0729+1.96*(0.0418))) = 0.95$$

$$P(e^{2*(0.0729-1.96*(0.0418))} \leq e^{2\beta_1} \leq e^{2*(0.0729+1.96*(0.0418)})) = 0.95$$

$$\Rightarrow (0.98, 1.36)$$

Thus, we're 95% confident that, when the explanatory increases by 2 units, the mean response changes by a factor that is somewhere between 0.98 and 1.36.

Question Answer

Done! Home

Subject 4	100

What are the practical differences between the gamma and inverse gaussian distributions?

Question Answer	Done! Home

The inverse Gaussian distribution dictates that $V(Y_i) \propto E(Y_i)^3$. That is, the variance increases cubically with the mean. This is opposed to the gamma, which dictates that the variance increases quadratically with the mean.

Let Y_i represent the number of shots made out of a total of 10 tries for student i. Let x_{1i} represent distance student i stood from the waste basket and x_{2i} represent the number of pets the student has. I fit a binomial logistic regression to this data with $Y_i \sim Binomial(m_i = 10, \pi_i)$ and $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$. I find $\hat{\beta}_0 = 3.2$, $\hat{\beta}_1 = -0.5$, and $\hat{\beta}_2 = 0.4$. Estimate the odds that the new student will make if she is 6 feet from the wastebasket and has 1 cat.

$$e^{3.2-0.5(6)+0.4(1)} = 1.822119 \ (= \frac{0.6456563}{1-0.6456563}).$$

Subject 4

300

We know about three possible link functions for Bernoulli/Binomial random components. What is the main advantage of using the logit link function over the other two?

Question
Answer

Done!
Home

Subject 4	300

The logit link function yields interpretation of (exponentiated) coefficients in terms of odds, while the other two **do not**.

Question Answer	Done! Home

Subject 4	400

Why do we not perform lack of fit tests for Bernoulli random components?

Question Answer	Done! Home

Subject 4

400

Lack of fit tests tend to fail when the counts are small. Thus, we never do this kind of testing for Bernoulli data (which is just Binomial data where $m_i = 1$.). We're cautious when $m_i < 5$ is common for Binomial data. (Also, we're cautious when counts < 5 are common for Poisson data.)

Question
Answer

Done!
Home

Subject 4

500

Let Y_i represent the number of shots made out of a total of 10 tries for student i. Let x_{1i} represent distance student i stood from the waste basket and x_{2i} represent the number of pets the student has. I fit a binomial logistic regression to this data with $Y_i \sim Binomial(m_i = 10, \pi_i)$ and $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$. I find $\hat{\beta}_0 = 3.2$, $\hat{\beta}_1 = -0.5$, and $\hat{\beta}_2 = 0.4$. Estimate the proportion of shots a new student will make if she is 6 feet from the wastebasket and has 1 cat.

Question
Answer

Done!
Home

The estimated proportion (probability) is:

$$\hat{\pi} = \frac{e^{3.2-0.5(6)+0.4(1)}}{1+e^{3.2-0.5(6)+0.4(1)}} = 0.6456563$$

Note: this is NOT the same as estimated odds, which is $e^{3.2-0.5(6)+0.4(1)} = 1.822119$ $(= \frac{0.6456563}{1-0.6456563})$.