BST 210 Lab: Week 11
Poisson Regression

So far in this course, we've seen regression methods for a wide variety of outcomes: linear regression for continuous outcomes, logistic regression for binary outcomes, multinomial regression for categorical outcomes, and generalized ordinal regression and proportional odds models for ordinal outcomes.

Another type of outcome that often appears in public health research—and particularly in epidemiological studies—is count data:

- How many patients relapse within the first year after discharge from an addiction treatment facility?
- How many patients are diagnosed with breast cancer within a particular part of the country?

This past week in class, we discussed how to incorporate count data into our existing regression/generalized linear models framework; the key to doing so is the Poisson distribution & Poisson regression!

The Poisson Distribution

The Poisson distribution allows us to describe the probability of seeing a certain number of events (Y) over a specific period of time. Let λ be the rate at which events occur per unit time (in epidemiology, this is referred to as the **incidence rate**!), and let t be some time interval of interest. Then:

$$P(Y = y) = \frac{e^{-\lambda t}(\lambda t)^y}{y!}, \qquad \text{for } y = 0, 1, 2, \ldots$$

Two key things to know about the Poisson distribution:

- $E[Y] = \lambda t = \mu$
- $Var(Y) = \lambda t = \mu$

As we saw in class, the Poisson distribution belongs to the exponential dispersion family, with $\theta = \log(\mu)$, $b(\theta) = e^{\theta} + log(y!) = e^{\log(\mu)} + log(y!) = \mu$, $\phi = 1$, and $c(y, \phi) = 0$. For those curious to see the math, I've included it below. However, it is not anything you will need to know for an exam, so feel free to avert your eyes and quickly turn the page!

$$\begin{aligned} f(y \mid \mu) &= \frac{e^{-\mu}\mu^y}{y!} \\ &= \exp\left\{-\mu + \log\left(\frac{\mu^y}{y!}\right)\right\} \\ &= \exp\left\{\frac{y\log(\mu) - \mu - \log(y!)}{1}\right\} \\ &= \exp\left\{\frac{y\theta - b(\theta)}{\phi + c(y, \phi)}\right\}. \end{aligned}$$

Poisson Regression

Since the Poisson distribution belongs to the exponential dispersion family, we can use our existing GLM framework to create a regression model for count data!

A Quick Refresher:

The generalized linear model (GLM) framework helps us write out regression models for any outcome Y, where Y comes from an exponential dispersion family member. It includes the regression types we've already seen in class, including linear regression, logistic regression, and multinomial regression. Typically, we relate

(1) the expected value of the outcome, $E[Y]$, to

(2) a linear combination of our covariates of interest, $\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p$, through

(3) a link function, $g(\cdot)$.

So generally speaking, our model has the form $g(E[Y]) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p$.

Let Y_i be the observed number of events for observation or covariate pattern i. We assume $Y_i \sim Poisson(\lambda_i t_i)$, where λ_i is the incidence rate, and t_i is the observed person-time of exposure. We'd like to be able to relate the expected number of events, $E[Y_i]$, to some covariates of interest, $X_1, \ldots, X_p$.

What link function $g(\cdot)$ might we want to consider for $E[Y_i]$? Why?

The link function $g(\cdot)$ relates our expected number of events, $E[Y_i]$, to a linear combination of our covariates, $\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p$. When we fit this "linear predictor" component, we don't make any restrictions on the range of our outcome–we're implicitly assuming that it takes on any value between $-\infty$ and ∞. However, intuitively, we'd only like fitted values of $E[Y_i]$ that are greater than or equal to 0; it doesn't make much sense to have a negative number of expected events!

So we'd like to choose a link function that transforms our outcome of interest, $E[Y_i]$, so we don't run into this problem. For that reason, a $\log(\cdot)$ link might be a good choice, as the $\log(\cdot)$ transformation takes things that only have positive range and "stretches them out" so that their range includes the entire real line!

If we use the log link, our transformed outcome is then

$$\log(E[Y_i]) = \log(\lambda_i t_i) = \log(\lambda_i) + \log(t_i). \tag{1}$$

Since t_i is a fixed/observed constant, it doesn't depend on our predictors! So the expected number of events $E[Y_i]$ depends on our covariates of interest only through the incidence rate, λ_i:

$$\log(\lambda_i) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p. \tag{2}$$

Combining the expressions in equations (1) and (2), what is the final Poisson regression model that we fit?

$$\begin{aligned} \log(E[Y_i]) &= \log(\lambda_i) + \log(t_i) \\ &= \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p + \log(t_i) \end{aligned}$$

What are some of the assumptions we make in fitting this model?

In fitting this model, we assume: (1) that the observed counts Y_i are independent, (2) that the $Y_i \sim Pois(\lambda_i t_i)$ and so the mean equals the variance, (3) that the mean model is correctly specified, and (4) that the incidence rate λ_i does not depend on the total person-time of exposure, t_i (i.e. that the incidence rate is time invariant).

What part of this model is the "offset"? Why is the offset important, and what does it help us adjust for?

The offset is the name we give to the $\log(t_i)$ term in the Poisson regression model. It adjusts for the fact that the total number of observed events depends both on the underlying incidence rate (the rate/number of events per some unit amount of time) and the total person-time of exposure.

Within our dataset, it may be the case (and, in fact, often is the case) that each individual or covariate pattern has a different amount of exposure time. So when we model the expected number of events in each individual/covariate pattern, it's important that we take this differing exposure time into account. The offset allows us to do so!

Why is it that the coefficient of the offset set to be 1?

The offset is a known constant for each individual, and so we don't need to estimate a coefficient for it. It's also the case that the offset is part of the expression for $\log(E[Y_i]) = \log(\lambda_i) + \log(t_i)$. Setting the coefficient to be anything other than one would change this relationship.

Poisson Regression: An Example

Let's actually try fitting a Poisson regression model in SAS! To do so, we'll use the data found in `melanoma.csv`. The data comes from a 1975 study of the geographic variation in the incidence of melanoma, and contains information on the number of incident melanoma cases (`inccases`) in nine different US cities (`locale`) and six different age categories (`ageg`), as well as the total person-years of exposure in each city/age group (`persyrs`).

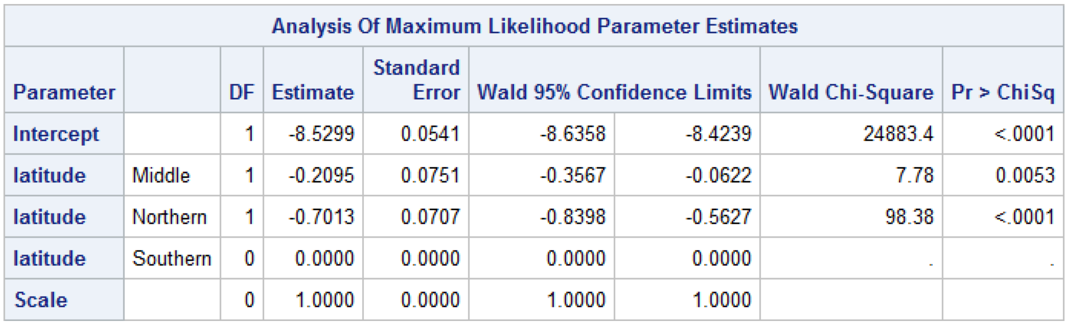
The nine US cities have also been grouped by their latitudes as being either "Northern", "Southern", or "Middle" (`latitudes`), while age has been categorized as "$<$ 35 years", "35-44 years", "45-54 years", "55-64 years", "65-74 years", and "$>$ 75 years". We will focus primarily on the association between latitude, age category, and melanoma incidence.

First, let's examine the relationship between latitude and melanoma incidence without adjusting for age:

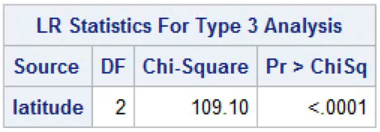
```
* Reading in the dataset;
proc import file="melanoma.csv" out=melanoma dbms=csv;
        getnames=YES;
run;

* Creating the log offset term;
data melanoma2;
        set melanoma;
        logt = log(persyrs);
run;

* Fitting a Poisson regression model for the relationship between latitude and melanoma
* incidence;
proc genmod data = melanoma2;
        class latitude / param=glm;
        model inccases = latitude / dist = poisson offset = logt type3;
        store lat1;
run;
```



Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-8.5299	0.0541	-8.6358	-8.4239	24883.4	<.0001
latitude	Middle	1	-0.2095	0.0751	-0.3567	-0.0622	7.78	0.0053
latitude	Northern	1	-0.7013	0.0707	-0.8398	-0.5627	98.38	<.0001
latitude	Southern	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		



LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
latitude	2	109.10	<.0001

What is the form of the Poisson model we just fit?

The model has the form

$$\log(E[\texttt{inccases}]) = \beta_0 + \beta_1 \cdot I(\texttt{Middle}) + \beta_2 \cdot I(\texttt{Northern}) + \log(\texttt{persyrs}).$$

Given our particular coefficient estimates, the model is

$$\log(E[\texttt{inccases}]) = -8.53 - 0.210 \cdot I(\texttt{Middle}) - 0.701 \cdot I(\texttt{Northern}) + \log(\texttt{persyrs}).$$

How would you interpret the intercept of the above model?

The incidence rate of melanoma cases within the population of individuals living in Southern latitudes is estimated to be $e^{-8.53} \approx 0.0002$ cases per person-year.

Effect Estimation

When we compare counts of events between two different populations, we face the added challenge that differences in counts may be partially/mostly/totally attributable to differences in the amount of exposure time between the two populations, rather than to any covariates of interest. This is a problem! For that reason, we choose to instead compare the incidence rates, since those measures are standardized by person-time of exposure. Our effect measure is then the incidence rate ratio (IRR):

$$IRR = \frac{\lambda_1}{\lambda_2} = \frac{\text{incidence rate (IR) in population 1}}{\text{incidence rate (IR) in population 2}}.$$

Using the above Poisson regression output, please provide an estimate for and interpretation of the two IRRs for the association between latitude and incidence of melanoma that are given directly by the model.

Among the population of individuals living in Northern latitudes, the incidence rate of melanoma is estimated to be $e^{-0.701} \approx 0.496$ times that among the population of individuals living in Southern latitudes.

Similarly, we estimate that the incidence rate of melanoma is $e^{-0.210} \approx 0.811$ times smaller among the populations of individuals living in Middle latitudes as compared to those living in Southern latitudes.

What test can we use to assess the significance of the association between latitude and melanoma incidence? Given the output above, is the association statistically significant?

Since the `latitude` covariate is treated as a factor variable, and thus is represented by two indicators, we want to test whether those indicators are collectively needed. Thus, we want to use a Likelihood Ratio Test, where our null hypothesis is that the model without latitude (i.e. including only the intercept and the offset term) is sufficient, while our alternative hypothesis is that the full model including latitude is needed. Given the above output, we reject the null hypothesis, and conclude that the association between melanoma and latitude is statistically significant ($p < 0.0001$).

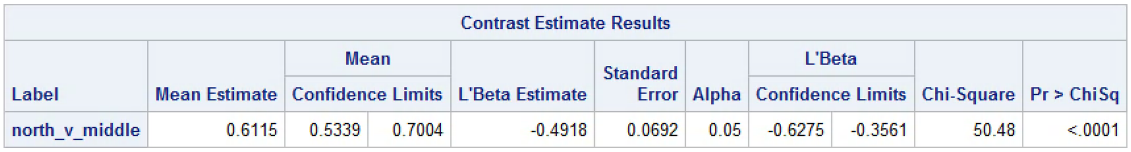
Suppose that we're specifically interested in the incidence rate ratio comparing the population of individuals living in northern latitudes to the population in middle latitudes. How can we use our model to arrive at an estimate for this IRR?

Given our model, we can write

$$\begin{aligned}\log(IRR) = \log\left(\frac{\lambda_{North}}{\lambda_{Mid}}\right) &= \log(\lambda_{North}) - \log(\lambda_{Mid})\\ &= \big(-8.53 - 0.701\big) - \big(-8.53 - 0.210\big)\\ &= 0.210 - 0.701 \approx -0.49.\end{aligned}$$

Note that, more abstractly, the log of the incidence rate ratio comparing melanoma incidence between populations in the northern and middle latitudes is simply $\beta_2 - \beta_1$: the coefficient for the indicator of `Northern` minus the coefficient for the indicator of `Middle`.

```
* Testing significance of and finding a confidence interval for Northern versus Middle
* latitudes;
proc genmod data=melanoma2;
        class latitude;
        model inccases = latitude / dist=poisson offset=logt type3;
        estimate 'north_v_middle' latitude -1 1;
run;
```



Contrast Estimate Results

		Mean			Standard		L'Beta			
Label	Mean Estimate	Confidence Limits		L'Beta Estimate	Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
north_v_middle	0.6115	0.5339	0.7004	-0.4918	0.0692	0.05	-0.6275	-0.3561	50.48	<.0001

Please interpret both the IRR and the 95% confidence interval comparing melanoma incidence between populations of individuals in the northern and middle latitudes.

The incidence rate of melanoma is estimated to be $(1 - 0.6115) \times 100\% \approx 38.85\%$ smaller among the population of individuals living in northern latitudes as compared to those living in the middle latitudes. With 95% confidence, the incidence rate ratio for the association between melanoma cases and northern versus middle latitudes is between 0.534 and 0.700.

Finally, using our fitted model, what is the estimated mean number of melanoma cases within a population of individuals living in cities with middle latitude and with a combined 200,000 person-years of exposure?

The estimated mean number of melanoma cases within this population is

$$\begin{aligned}&\log(E[\texttt{inccases}]) = -8.53 - 0.210 + \log(200000) \approx 3.47\\ \implies\ &E[\texttt{inccases}] = e^{3.47} \approx 32.03 \text{ cases.}\end{aligned}$$

Confounding & Effect Modification

Confounding and effect modification function much the same for Poisson regression as they did for linear, logistic, and multinomial/ordinal models!

For confounding:

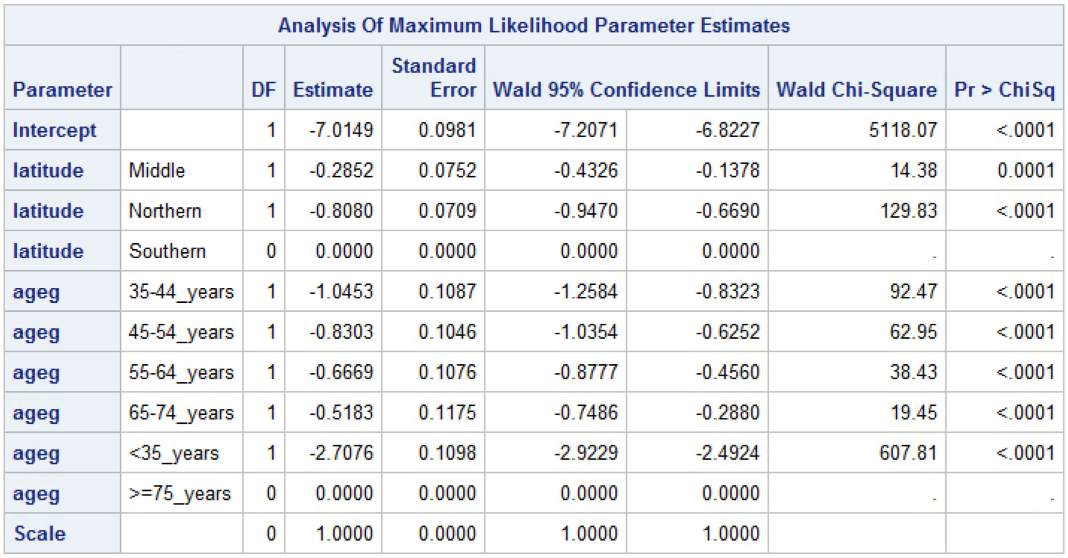
- There is no formal statistical test that we can perform to assess confounding, since confounding is not a statistical concept. Be careful not to call anything a *significant* confounder!
- We'll use a 10% difference between the adjusted and unadjusted coefficients as our rule-of-thumb for something being a *meaningful* confounder.

For effect modification:

- We're interested in determining whether the interaction term is statistically significant.
- Remember, if we're testing an interaction that includes more than one term, we need to test all of the coefficients *collectively*, as opposed to looking at their individual p-values!

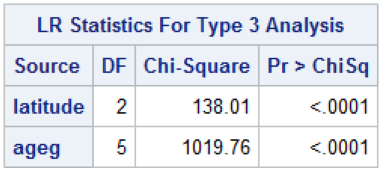
Let's examine the role of age as a potential confounder and effect modifier of the relationship between latitude and incident melanoma! We'll start by adding the main effect of age to the Poisson regression model:

```
* Adding in the main effect of age category;
proc genmod data=melanoma2;
        class latitude ageg / param = glm;
        model inccases = latitude ageg / dist=poisson offset=logt type3;
        store lat2;
run;
```



Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.0149	0.0981	-7.2071	-6.8227	5118.07	<.0001
latitude	Middle	1	-0.2852	0.0752	-0.4326	-0.1378	14.38	0.0001
latitude	Northern	1	-0.8080	0.0709	-0.9470	-0.6690	129.83	<.0001
latitude	Southern	0	0.0000	0.0000	0.0000	0.0000	.	.
ageg	35-44_years	1	-1.0453	0.1087	-1.2584	-0.8323	92.47	<.0001
ageg	45-54_years	1	-0.8303	0.1046	-1.0354	-0.6252	62.95	<.0001
ageg	55-64_years	1	-0.6669	0.1076	-0.8777	-0.4560	38.43	<.0001
ageg	65-74_years	1	-0.5183	0.1175	-0.7486	-0.2880	19.45	<.0001
ageg	<35_years	1	-2.7076	0.1098	-2.9229	-2.4924	607.81	<.0001
ageg	>=75_years	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		



LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
latitude	2	138.01	<.0001
ageg	5	1019.76	<.0001

Does age appear to be a meaningful confounder of the association between latitude and melanoma incidence? Why or why not?

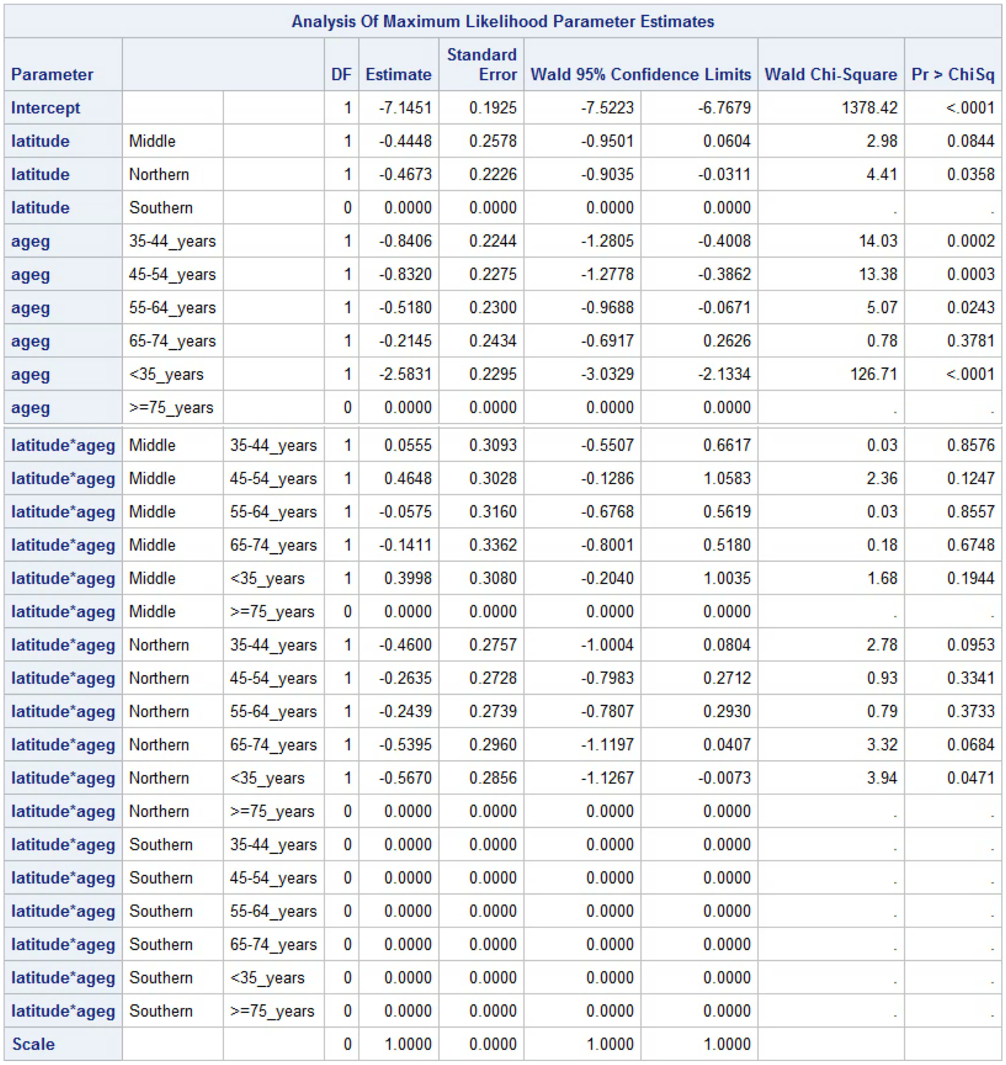
Yes, age does appear to be a meaningful confounder. Age meets our definition for a confounder, as age is associated with latitude (a number of individuals in the US retire to states in the South, since the winters are milder) and associated with melanoma (as individuals age they are at higher risk for all types of cancer), but is clearly not a downstream consequence of latitude. That age is a meaningful confounder is evidenced by our 10% rule-of-thumb for confounding. The coefficient estimate for $I(\texttt{Middle})$ changed by $\frac{-0.2095+0.2852}{-0.2095} \approx 36.1\%$ and $I(\texttt{Northern})$ changed by $\frac{-0.7013+0.8080}{-0.7013} \approx 15.2\%$ from the unadjusted to the adjusted analysis; both of these changes are greater than 10%.

Is age an independent predictor of melanoma incidence? How would you interpret the IRR for the association between age (specifically being under 35 versus over 75 years of age) and melanoma?

Yes, age is a significant independent predictor of melanoma incidence ($p < 0.0001$). We estimate that the incidence rate of melanoma among the population of individuals under 35 years of age is $e^{-2.708} \approx 0.0667$ times that among the population of individuals over 75 years of age, holding latitude constant.

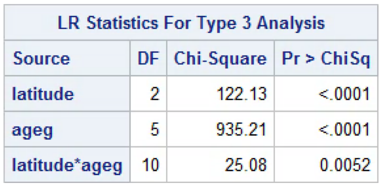
To check for the presence of effect measure modification by age, we must first add an age by latitude interaction into the model:

```
* Adding in an interaction term between age and latitude;
proc genmod data=melanoma2;
        class latitude ageg / param = glm;
        model inccases = latitude ageg latitude*ageg / dist=poisson offset=logt type3;
        store lat3;
run;
```



Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	-7.1451	0.1925	-7.5223	-6.7679	1378.42	<.0001
latitude	Middle		1	-0.4448	0.2578	-0.9501	0.0604	2.98	0.0844
latitude	Northern		1	-0.4673	0.2226	-0.9035	-0.0311	4.41	0.0358
latitude	Southern		0	0.0000	0.0000	0.0000	0.0000	.	.
ageg	35-44_years		1	-0.8406	0.2244	-1.2805	-0.4008	14.03	0.0002
ageg	45-54_years		1	-0.8320	0.2275	-1.2778	-0.3862	13.38	0.0003
ageg	55-64_years		1	-0.5180	0.2300	-0.9688	-0.0671	5.07	0.0243
ageg	65-74_years		1	-0.2145	0.2434	-0.6917	0.2626	0.78	0.3781
ageg	<35_years		1	-2.5831	0.2295	-3.0329	-2.1334	126.71	<.0001
ageg	>=75_years		0	0.0000	0.0000	0.0000	0.0000	.	.
latitude*ageg	Middle	35-44_years	1	0.0555	0.3093	-0.5507	0.6617	0.03	0.8576
latitude*ageg	Middle	45-54_years	1	0.4648	0.3028	-0.1286	1.0583	2.36	0.1247
latitude*ageg	Middle	55-64_years	1	-0.0575	0.3160	-0.6768	0.5619	0.03	0.8557
latitude*ageg	Middle	65-74_years	1	-0.1411	0.3362	-0.8001	0.5180	0.18	0.6748
latitude*ageg	Middle	<35_years	1	0.3998	0.3080	-0.2040	1.0035	1.68	0.1944
latitude*ageg	Middle	>=75_years	0	0.0000	0.0000	0.0000	0.0000	.	.
latitude*ageg	Northern	35-44_years	1	-0.4600	0.2757	-1.0004	0.0804	2.78	0.0953
latitude*ageg	Northern	45-54_years	1	-0.2635	0.2728	-0.7983	0.2712	0.93	0.3341
latitude*ageg	Northern	55-64_years	1	-0.2439	0.2739	-0.7807	0.2930	0.79	0.3733
latitude*ageg	Northern	65-74_years	1	-0.5395	0.2960	-1.1197	0.0407	3.32	0.0684
latitude*ageg	Northern	<35_years	1	-0.5670	0.2856	-1.1267	-0.0073	3.94	0.0471
latitude*ageg	Northern	>=75_years	0	0.0000	0.0000	0.0000	0.0000	.	.
latitude*ageg	Southern	35-44_years	0	0.0000	0.0000	0.0000	0.0000	.	.
latitude*ageg	Southern	45-54_years	0	0.0000	0.0000	0.0000	0.0000	.	.
latitude*ageg	Southern	55-64_years	0	0.0000	0.0000	0.0000	0.0000	.	.
latitude*ageg	Southern	65-74_years	0	0.0000	0.0000	0.0000	0.0000	.	.
latitude*ageg	Southern	<35_years	0	0.0000	0.0000	0.0000	0.0000	.	.
latitude*ageg	Southern	>=75_years	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale			0	1.0000	0.0000	1.0000	1.0000		



LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
latitude	2	122.13	<.0001
ageg	5	935.21	<.0001
latitude*ageg	10	25.08	0.0052

How do we formally test for the presence of effect modification? In particular, what are the null and alternative hypotheses, and what test and null distribution should we use?

Testing for the presence of effect modification is tantamount to assessing whether the interaction term (in this case the interaction terms between age and latitude) is significant. The null hypothesis is that the reduced model (with the main effects of age and latitude, but without the interaction) is sufficient, while the alternative hypothesis is that the full model (including the interaction term) is preferred.

We can test these hypotheses using a Likelihood Ratio Test. Since we need to estimate 10 coefficients in order to fit the interaction term, we compare our LRT test statistic to a χ^2_{10} distribution.

What are your conclusions?

The SAS output shows us the results of the Likelihood Ratio test for the significance of the interaction term between age and latitude. We reject the null hypothesis that the model without this interaction is sufficient, and thus conclude that the interaction term is needed ($p = 0.005$). As such, age is a significant effect modifier of the relationship between latitude and melanoma incidence.

Overdispersion

In assuming that the observed counts Y follow a Poisson distribution, we are implicitly making a very strong assumption about the relationship between the mean and the variance of the outcome, namely that $E[Y] = Var[Y] = \lambda t$. However, there are several scenarios in which this assumption does not hold, including when:

- The observed counts are not independent between individuals/covariate patterns
- The incidence rate λ varies over time.

Both of these situations lead to **overdispersion**, meaning that the actual variability in the count outcomes is much greater than the variability predicted by the model.

What are some of the problems that result from underestimating the variability in the counts?

Underestimating the variability of the counts leads to inferences (statistical conclusions) that are not valid. In particular, we will end up with confidence intervals that are too narrow, since the width of the interval depends directly on the estimated standard error. If the intervals are too narrow, what we think is a 95% confidence interval may only cover the true parameter value 93% or 80% of the time—we will think we are more confident in our point estimate than we actually are. Similarly, if we underestimate the variability in the counts, we will end up with p-values that are smaller than they should be. As a result, we will be more likely to incorrectly conclude an association is significant when it is, in fact, not.

Detecting Overdispersion

We can check for the presence of overdispersion by examining the goodness-of-fit of the model! Suppose Y_j is the observed number of events for the jth covariate pattern (or observation), $j = 1, \ldots, J$, and that $\hat{Y}_j$ is the expected number of events for that same covariate pattern from the Poisson model with p covariates $X_1, \ldots, X_p$. Then there are two key statistics that we use to examine goodness-of-fit for Poisson regression:

- Deviance

Test Statistic:	$2\sum_{j=1}^{J} Y_j \log\left(\frac{Y_j}{\widehat{Y}_j}\right) \sim \chi^2_{J-(p+1)}$
Intuition:	How do our predicted counts compare to the saturated model?

- Pearson χ^2 statistic

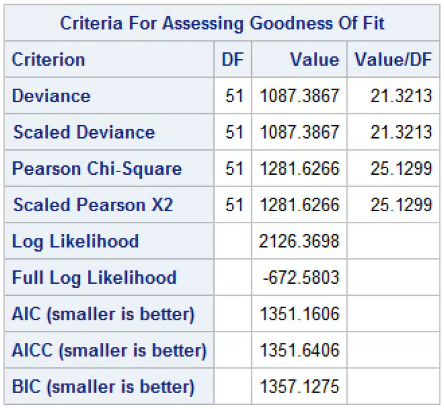
Test Statistic:	$\sum_{j=1}^{J} \frac{(Y_j - \widehat{Y}_j)^2}{\widehat{Y}_j} \sim \chi^2_{J-(p+1)}$
Intuition:	How off were our predictions by assuming the counts were Poisson distributed?

Assuming that we have enough observations in each covariate pattern, both the deviance and the Pearson χ^2 statistic follow a $\chi^2_{J-(p+1)}$ distribution. So we can do a formal goodness-of-fit test to assess whether overdispersion is present!

However, when we have a large number of covariate patterns, or a small sample size, it is often the case that we don't have quite enough observations for the $\chi^2_{J-(p+1)}$ distribution to be correct. In that case, we rely on more empirical checks! In particular, we look at either the Pearson χ^2 statistic or the deviance statistic divided by their degrees of freedom! If no substantial overdispersion is present, then

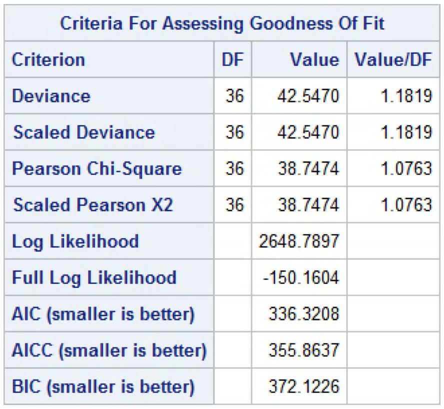
- $\frac{\text{Deviance}}{J-(p+1)} \approx 1$
- $\frac{\text{Pearson } \chi^2 \text{ statistic}}{J-(p+1)} \approx 1$

SAS automatically returns goodness-of-fit information with each PROC GENMOD call! Below are the outputs for the model with latitude only (left) and the model with latitude, age, and their interaction (right):



Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	51	1087.3867	21.3213
Scaled Deviance	51	1087.3867	21.3213
Pearson Chi-Square	51	1281.6266	25.1299
Scaled Pearson X2	51	1281.6266	25.1299
Log Likelihood		2126.3698	
Full Log Likelihood		-672.5803	
AIC (smaller is better)		1351.1606	
AICC (smaller is better)		1351.6406	
BIC (smaller is better)		1357.1275	

(a) Latitude-only model.



Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	36	42.5470	1.1819
Scaled Deviance	36	42.5470	1.1819
Pearson Chi-Square	36	38.7474	1.0763
Scaled Pearson X2	36	38.7474	1.0763
Log Likelihood		2648.7897	
Full Log Likelihood		-150.1604	
AIC (smaller is better)		336.3208	
AICC (smaller is better)		355.8637	
BIC (smaller is better)		372.1226	

(b) Interaction model.

Based on this output, should we be concerned about overdispersion in the interaction model? What about in the latitude only model?

Although the interaction model appears to be okay (the deviance and Pearson chi-square divided by their degrees of freedom are both near one), we definitely seem to have a problem in the latitude-only model! The ratios are instead near 20-25, which certainly indicates overdispersion!

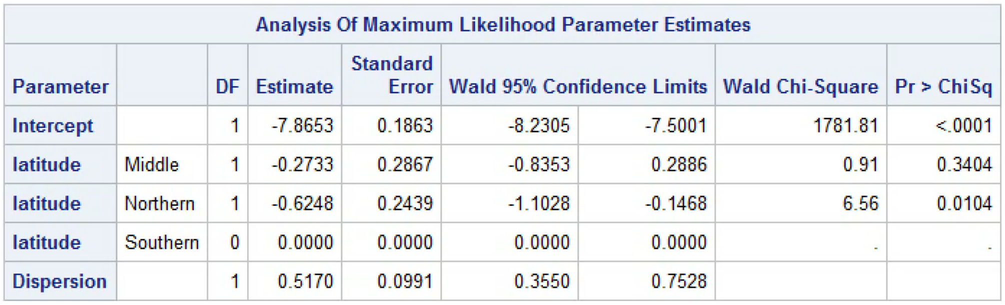
Adjusting for Overdispersion

If overdispersion appears to be present in the data, there are several things we can do! Let's focus on the latitude-only model, and go through two of the approaches we discussed in class.

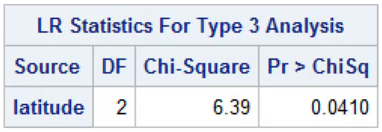
Negative Binomial Regression

The negative binomial distribution also handles counts, but doesn't require as strict of an assumption about the mean-variance relationship. We can fit a Negative Binomial regression model in SAS by using the following code:

```
* Negative Binomial Regression;
proc genmod data=melanoma2;
        class latitude;
        model inccases = latitude / dist=negbin offset=logt type3;
run;
```



Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.8653	0.1863	-8.2305	-7.5001	1781.81	<.0001
latitude	Middle	1	-0.2733	0.2867	-0.8353	0.2886	0.91	0.3404
latitude	Northern	1	-0.6248	0.2439	-1.1028	-0.1468	6.56	0.0104
latitude	Southern	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	0.5170	0.0991	0.3550	0.7528		



LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
latitude	2	6.39	0.0410

How do the point estimates compare to the estimates from the Poisson regression model? What about the standard errors?

Both the point estimates and standard errors differ from the Poisson regression model. For example, the coefficient for Middle latitude has gone from -0.210 (in the standard model) to -0.273 (in the Negative Binomial model). We also see that the standard errors are estimated to be much larger in the Negative Binomial model than in the standard Poisson regression model—further supporting our suspicion that we had underestimated the variances in the "standard" model.

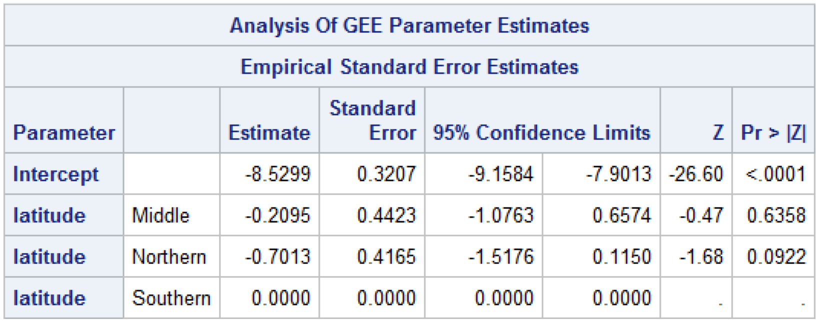
Robust variance estimation

In robust variance estimation, we assume that our mean model is correctly specified, meaning that we still assume

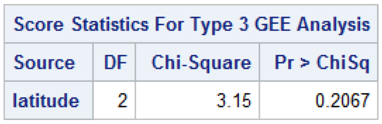
$$\log(E[Y_i]) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p + \log(t_i).$$

However, we use a robust method to estimate the variance separate from this mean. We can implement this in SAS using the `repeated` option in PROC GENMOD:

```
* Robust Variance Estimation;
proc genmod data=melanoma2;
        class latitude VAR1;
        model inccases = latitude / dist=poisson offset=logt type3;
        repeated subject=VAR1 / covb;
run;
```



Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-8.5299	0.3207	-9.1584	-7.9013	-26.60	<.0001
latitude	Middle	-0.2095	0.4423	-1.0763	0.6574	-0.47	0.6358
latitude	Northern	-0.7013	0.4165	-1.5176	0.1150	-1.68	0.0922
latitude	Southern	0.0000	0.0000	0.0000	0.0000	.	.



Score Statistics For Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
latitude	2	3.15	0.2067

Once again, how do the point estimates compare to the estimates from the Poisson regression model? What about the standard errors?

When we use robust variance estimation, we get the exact same point estimates we did when simply using a standard Poisson regression model! The standard error estimates are once again much higher under the model with robust variance estimation, given that it does handle/correct for overdispersion.

What are some differences between the Negative Binomial and the robust variance estimation approaches?

The key difference between these two approaches is in how they handle the mean model (the model for $E[Y]$). When we use a robust variance estimation approach, we still assume that this mean model is correctly specified—meaning that we still assume the counts are Poisson distributed, and that the functional form that we gave for $E[Y]$ is correct. We simply estimate the variance separately from the mean. As such, we arrive at the exact same point estimates for the coefficients.

When we fit a Negative Binomial model, we are actually changing our assumptions about the distribution of the counts. In particular, we now assume that the counts follow a Negative Binomial distribution, rather than a Poisson distribution. And since we're assuming a different underlying distribution for the Y_i, we arrive at slightly different point estimates.

Zero-Inflated Poisson

The Zero-Inflated Poisson (often abbreviated as ZIP!) allows us to account for a second kind of "departure" from a standard Poisson distribution! In some cases—especially when we're interested in a rare outcome, we may find that our dataset contains a much higher number of zero counts than we would expect if the counts were actually Poisson distributed. To handle this extra number of "structural" zeroes, the Zero-Inflated Poisson distribution consists of two parts:

- A **binary** component that generates the structural zeroes
- A regular **Poisson** component that generates some zeroes, as well as all other non-zero counts.

$$P(Y = y \,|\, \lambda, \pi) = \underbrace{\pi \cdot I(y = 0)}_{\text{structural zeroes}} + \underbrace{(1 - \pi)\frac{e^{-\lambda t}(\lambda t)^y}{y!}}_{\text{Poisson counts}}$$

Notice that, based on the Zero-Inflated Poisson model, we now have a higher probability of observing $Y = 0$!

Now that we have this model, we can imagine extending Poisson regression so that our count outcome Y follows a Zero-Inflated Poisson distribution instead. However, it's important to note that our expected number of counts $E[Y]$ now depends on both the incidence rate λ and the probability of seeing a structural zero, π.

- We have the option of using logistic regression to model π and Poisson regression to model λ as a function of our observed covariates $X_1, \ldots, X_p$!
- We don't need to use the same covariates to model both λ and π. In fact, the model is often easier to estimate if we use different covariates to model λ than we do to model π.

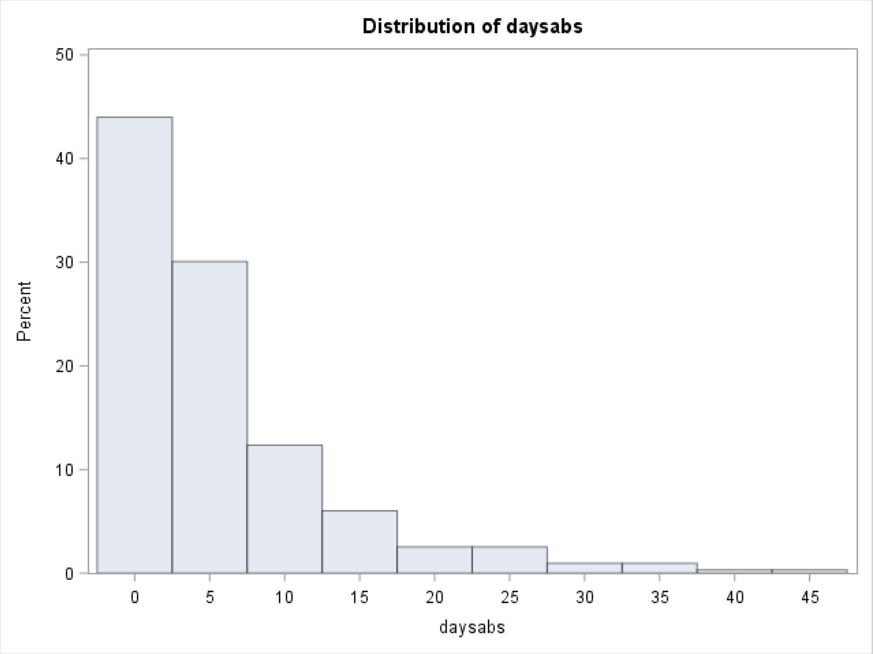
Zero-Inflated Poisson: An Example

For this example, we'll use the data found in `school_data.csv`. The dataset contains information on 316 students, with the outcome of interest being the number of absences recorded during the school year, and possible predictors being math standardized test scores, language standardized test scores, and gender.

Let's first look at a histogram of the number of recorded absences!

```
* Reading in the dataset;
proc import file='school_data.csv' out=school dbms=csv;
        getnames=YES;
run;

* Creating a histogram of student absences;
title 'Student Absences over the Course of a School Year';
proc univariate data=school noprint;
        histogram daysabs;
run;
```



Given the histogram above, there appears to be an unusually high/larger than expected number of zero counts in the dataset; we could confirm this by looking at a frequency table of recorded absences.

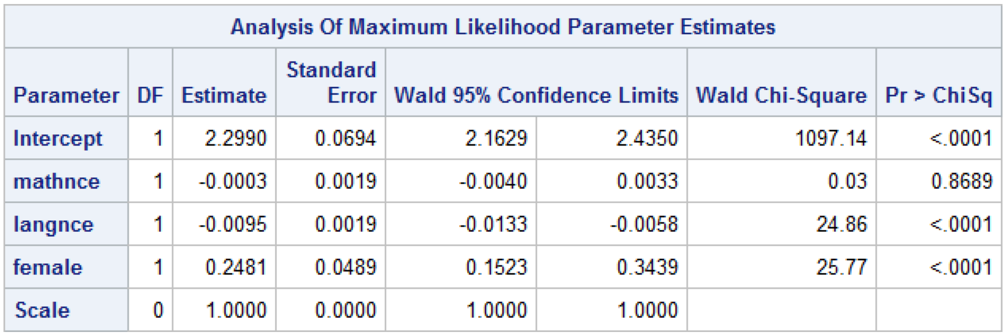
In this case, we may want to consider fitting a Zero-Inflated Poisson model! For a ZIP model, we could consider modeling both the probability of being a structural zero (π) and the incidence rate (λ) as functions of our covariates of interest:

$$\text{logit}(\pi) = \gamma_0 + \gamma_1 \cdot X_1 + \ldots + \gamma_p \cdot X_p$$
$$\log(\lambda) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p$$

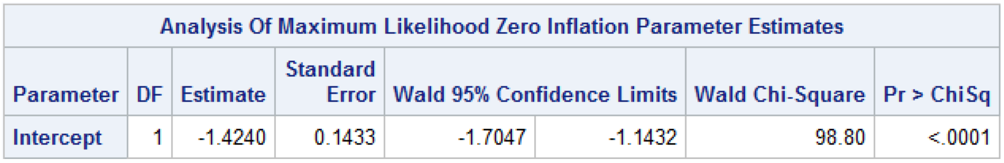
Note that among the population that is not a structural zero, we also have $\log(E[Y]) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p + \log(t_i)$.

Let's first model just the incidence rate as a function of math standardized test scores, language standardized test scores, and gender. Note that this means we are assuming that π is a constant.

```
* Fitting a ZIP with constant structural zero probability;
proc genmod data = school;
        model daysabs = mathnce langnce female / dist=zip;
        zeromodel / link=logit;
run;
```



Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	2.2990	0.0694	2.1629	2.4350	1097.14	<.0001
mathnce	1	-0.0003	0.0019	-0.0040	0.0033	0.03	0.8689
langnce	1	-0.0095	0.0019	-0.0133	-0.0058	24.86	<.0001
female	1	0.2481	0.0489	0.1523	0.3439	25.77	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		



Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4240	0.1433	-1.7047	-1.1432	98.80	<.0001

What is the form of the model that we fit?

We first observe that no follow-up times were recorded for any of the students, as they were assumed to have all been followed for the same amount of time. As such, we did not need to include any sort of offset term in our model, and so the model for the incidence rate λ is equivalent to the model for the expected number of counts, $E[Y_i]$. (Alternatively, we can think of the entire school year as being one unit of person-time.) In that case, the model we fit is simply

$$\log(E[Y_i]) = \log(\lambda_i) = 2.299 - 0.0003 \cdot \texttt{mathnce} - 0.0095 \cdot \texttt{langnce} + 0.248 \cdot \texttt{female}$$
$$\text{logit}(\pi_i) = -1.424.$$

Based on this model fit, what is the estimated probability of being a structural zero?

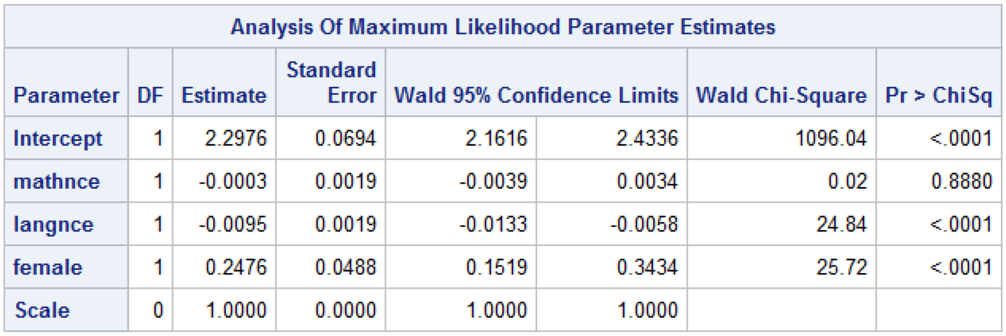
The probability of being a structural zero is estimated to be $\frac{e^{-1.424}}{1+e^{-1.424}} \approx 19.4\%$ for all students.

Please interpret the `langnce` coefficient. Is its effect significant?

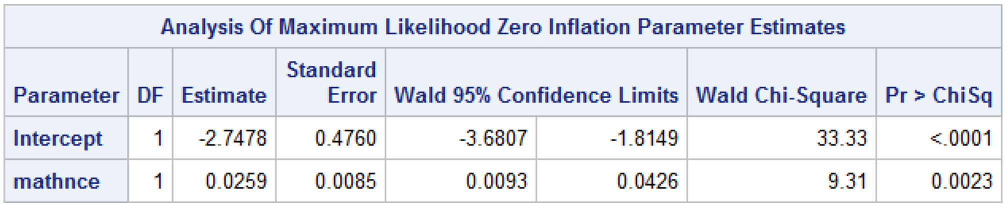
Among those students who are not "certain zeroes"/"structural zeroes", a one point increase in language standardized testing score is associated with a 0.95% decrease in the expected number of (equivalently, incidence rate of) school absences, holding math standardized test score and gender constant. The association between language score and number of absences is statistically significant ($p < 0.0001$).

Suppose that we believe the probability of being a certain/structural zero depends on math standardized test scores. In that case, we can fit the corresponding ZIP model using the following code:

```
* Fitting a ZIP with structural zero probability dependent on mathnce;
proc genmod data=school;
        model daysabs = mathnce langnce female / dist=zip;
        zeromodel mathnce / link=logit;
run;
```



Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	2.2976	0.0694	2.1616	2.4336	1096.04	<.0001
mathnce	1	-0.0003	0.0019	-0.0039	0.0034	0.02	0.8880
langnce	1	-0.0095	0.0019	-0.0133	-0.0058	24.84	<.0001
female	1	0.2476	0.0488	0.1519	0.3434	25.72	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		



Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7478	0.4760	-3.6807	-1.8149	33.33	<.0001
mathnce	1	0.0259	0.0085	0.0093	0.0426	9.31	0.0023

Please write down the new model that was fit, and interpret both `mathnce` *coefficients.*

Our new model is given by

$$\log(E[Y_i]) = \log(\lambda_i) = 2.298 - 0.0003 \cdot \texttt{mathnce} - 0.0095 \cdot \texttt{langnce} + 0.2476 \cdot \texttt{female}$$
$$\text{logit}(\pi_i) = -2.748 + 0.0259 \cdot \texttt{mathnce}.$$

A one point increase in standardized math scores is associated with a $e^{0.0259} \approx 1.026$ times increase in the odds of belonging to the structural zero group ($p = 0.0023$).

Among those students who are not certain/structural zeroes, a one point increase in standardized math score is associated with a $e^{-0.0003} \approx 0.9997$ times change in the expected number of (equivalently, incidence rate of) school absences, holding language test scores and gender constant ($p = 0.888$).