Activity #13: Generalized Linear Models: Logistic, Poisson Regression

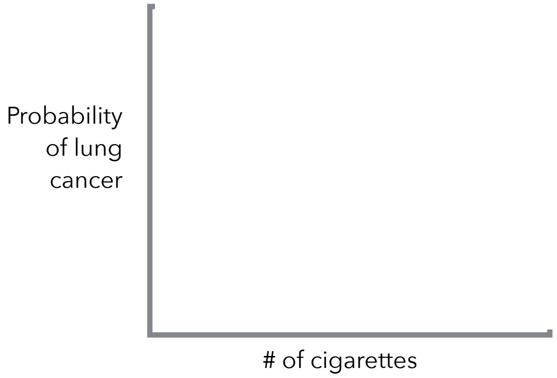
In every regression model we've used thus far, our dependent variable has been a continuous variable. In this activity, we'll learn how to use the *Generalized Linear Model* to model binary (or ordinal) dependent variables. We'll also learn how to model a dependent variable that represents a count.

1. Suppose we want to predict whether an individual will develop lung cancer. It might be reasonable to model lung cancer as a function of the number of cigarettes smoked each day. Suppose we sample 1,000 individuals for this study. To the right, sketch a scatterplot we might expect.

2. Could we fit a regression line to that scatterplot? Evaluate how well that line would model the data. Are there any problems in trying to fit a line to this data?

3. Our regular linear regression techniques won't work for binary dependent variables. Since our dependent variable can take only one of two values (0 or 1), the coefficients of our line won't make much sense.

   We might be more interested in modeling the probability of our dependent variable. In this case, we want to predict the probability that an individual gets lung cancer based on the number of cigarettes smoked each day. Using your intuition, fit a curve that you think would best describe this relationship.

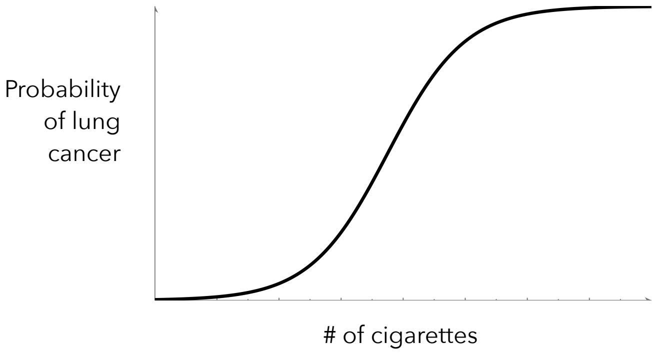


The relationship between a single predictor (x1) and a binary dependent variable (y) can be modeled by a **logistic function**:

$$P\left(y=1\right)=\frac{1}{1+e^{-\left(b_0+b_1x_1\right)}}$$

4. In this fictitious lung cancer example, we might find the logistic function displayed to the right.

   We could enter our observed data into a computer and have it find the best-fitting logistic curve (just as we fit nonlinear models in the previous activity). This method is tedious, though.



$$P(y=1)=\frac{1}{1+e^{-(b_0+b_1x_1)}}$$

Notice that logistic functions have three properties:

a) They are asymptotic with respect to P(y) = 0 and P(y) = 1. This allows us to use logistic functions to model probabilities (since probabilities range from 0-1).

b) The are monotonically increasing. This can be useful if we believe higher values of our independent variable correspond with higher probabilities of the dependent variable.

c) They are continuous.

If fitting the logistic curve as a nonlinear model is tedious, how else can we estimate the coefficients? To answer this, we need to take a quick detour back to the concepts of probabilities and odds.

A 1994 study of 17,096 college students found that 3,314 identified themselves as *frequent binge drinkers*.

Source: http://www.ndsn.org/jan95/college.html
Henry Wechsler, Ph.D., et.al., "Health and Behavioral Consequences of Binge Drinking in College," Journal of the American Medical Association, Dec. 7, 1994, p. 1672-1677; "The Disruptions of Campus Drunkenness," US News & World Report, Dec. 19, 1994, p. 12

5. If the students in this study are representative of students at St. Ambrose, what's the probability that a St. Ambrose student is a binge drinker?

6. Recall that we define odds to be: $\text{odds of an event} = \frac{P(\text{event})}{1-P(\text{event})}$. Calculate and interpret the following:

   a) The odds that a student is a binge drinker:

   b) The odds that a student is not a binge drinker:

7. Recall that with a logistic function, we can model the probability of an event by $P\left(y=1\right)=\frac{1}{1+e^{-\left(b_0+b_1x_1\right)}}$.

We can convert this to an odds function to get: $\frac{\frac{1}{1+e^{-\left(b_0+b_1x_1\right)}}}{1-\frac{1}{1+e^{-\left(b_0+b_1x_1\right)}}}$

To simplify the notation, let's substitute z for the exponent and simplify:

$$\text{odds} = \frac{\frac{1}{1+e^{-z}}}{1-\frac{1}{1+e^{-z}}} = \frac{\frac{1}{1+e^{-z}}}{\frac{1+e^{-z}}{1+e^{-z}}-\frac{1}{1+e^{-z}}} = \frac{\frac{1}{1+e^{-z}}}{\frac{e^{-z}}{1+e^{-z}}} = \frac{1}{1+e^{-z}}\left(\frac{1+e^{-z}}{e^{-z}}\right) = \frac{1}{e^{-z}} = e^{z} = e^{\left(b_0+b_1x_1\right)}$$

If we take the natural log of the odds, we get the log odds (also called the *logit):*

$$\ln\left(\text{odds}\right)=\ln\left(e^{\left(b_0+b_1x_1\right)}\right)=b_0+b_1x_1$$

The *logit* (log-odds) is simply a linear function. This means a computer can use iterative methods to fit a linear function to the log-odds. We can then transform those log-odds into odds and, finally, probabilities to interpret the model.

We'll use the logistic function to model the probability of an event: $\frac{1}{1+e^{-\left(b_0+b_1x_1\right)}}$

To estimate the coefficients, we convert this to a logit (log-odds): $\ln\left(\text{odds}\right)=b_0+b_1x_1$

Once we have our coefficients, we may want to convert log-odds back to odds: $\text{odds}=e^{b_0+b_1x_1}$

Then we may want to convert the odds back to a probability: $P\left(y\right)=\frac{\text{odds}}{\text{odds + 1}}=\frac{e^{b_0+b_1x_1}}{e^{b_0+b_1x_1}+1}=\frac{1}{1+e^{-\left(b_0+b_1x_1\right)}}$

Let's work through a simple example by hand. In this scenario, we'll try to see if gender predicts whether a student is more likely to be a frequent binge drinker.

8. The following table displays the results from that binge drinking study. Fill-in the table below.

	Frequent binge drinker	Not	Total
Male	1624	5528	7152
Female	1690	8254	9944
Total	3314	13782	17096

	Female	Male
P(binge drinker) =		
Odds(binge drinker) =		
ln(odds) =		
Odds ratio =		

9. Suppose we code our gender variable as female = 0 and male = 1. We'd get the following results:

	Female (gender = 0)	Male (gender = 1)
ln(odds) =	ln(0.205) = -1.59	ln(0.294) = -1.23

Recall that $\ln(\text{odds}) = b_0 + b_1x_1$

-1.59 = $b_0 + b_1(0)$ -1.23 = -1.59 + $b_1(1)$

Therefore, b_0 = _______________ Therefore, b_1 = _______________

Our model for the ln(odds) is: ____

Odds(binge drinker) =

___________________________ ___________________________

Odds ratio = ____

P(binge drinker) =

Relative probability = ___________________________________

10. We've just gone through our first example of a logistic regression. We use logistic regression when our dependent variable is dichotomous/binary. We'll see more (real) examples of logistic regression later. For now, let's discuss the *generalized linear model*.

When we fit a linear regression model, we're predicting the expected value of a dependent variable as a linear combination of a set of predictors. We interpret the coefficients of the model in such a way that a constant change in a predictor leads to a constant change in the dependent variable. This works when our dependent variable can, essentially, vary indefinitely in either direction.

Linear models are not appropriate for many other types of dependent variables. For example, suppose we want to model the number of students attending St. Ambrose as a function of the cost of tuition. Notice that our dependent variable (number of students) represents a **count** that must be positive. If we fit a linear model to this data, we may predict a negative number of students at some tuition cost. This is clearly impossible. In this case, it may be better to use a linear model to predict a constant rate of student enrollment (e.g., an increase in tuition of $1,000 leads to 1% fewer students attending). This type of model is called a *log-linear model*, since the logarithm of the dependent variable is predicted to vary linearly.

Like we concluded on the first page of this activity, it's not appropriate to use a linear model to predict a binary dependent variable (a Bernoulli variable). A line will predict values other than the 0 or 1 that are possible for the dependent variable.

Generalized linear models handle these situations by:

- allowing dependent variables to have arbitrary distributions (other than normal distributions)
- allowing an arbitrary (link) function of the dependent variable to vary linearly with the predicted values.

In the case of modeling student enrollment as a function of tuition, we may choose a Poisson model (to model counts) and a log link. In the binge drinking example, we chose a Binomial model (to model probabilities) and a logit (log-odds) link function.

Generalized linear models consist of three components:

1) Random component

This specifies the distribution of the dependent variable (y) given values of the predictor variables.

If y is a continuous variable, its probability distribution might be normal;
if y is binary, the distribution might be binomial;
if y represents counts, the distribution might be Poisson

2) Systematic (linear) component

This represents a linear combination of the predictor variables: $\eta_i = b_0 + b_1 x_1 + ... + b_k x_k$

These predictor variables may be continuous, categorical, transformations of continuous variables, polynomial terms, interactions, etc.

3) Link function

This links the random and systematic components by linking the expected value of y to the predictors: $g(E[y]) = \eta_i = b_0 + b_1 x_1 + ... + b_k x_k$. Some common link functions include:

- Identity link -- $g(E[y]) = \mu_y$ -- which is used in standard linear models.
- Log link -- $g(E[y]) = \ln(\mu_y)$ -- which is used for count data in log-linear models.
- Logit link -- $g(E[y]) = \ln(\mu_y / (1 - \mu_y))$ -- which is used for binary dependent variables.

Because the link function is invertible, we can write $E[y] = g^{-1}(\eta_i) = g^{-1}(b_0 + b_1 x_1 + \ldots + b_k x_k)$

With this, the generalized linear model can be thought of as
- a linear model for a transformation of the expected value of the dependent variable, or
- a nonlinear regression model for the dependent variable

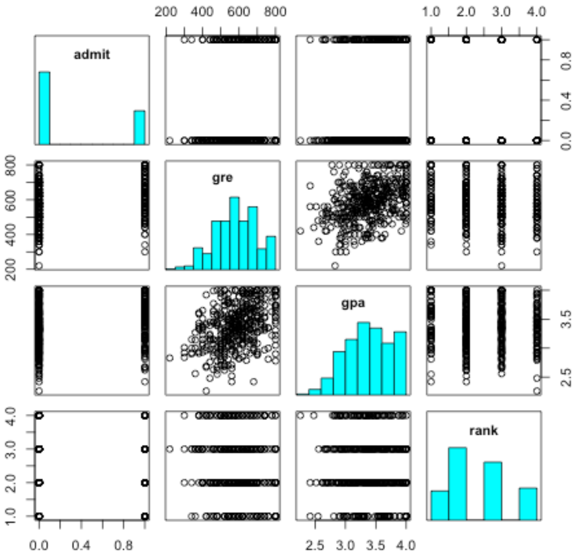
Scenario: A researcher is interested in how variables such as GRE scores, undergraduate GPAs, and the prestige of the undergraduate institution predict admission to graduate school.

The dataset consists of 400 observations on these 4 variables. The dependent variable, **admit**, is binary:

admit = 1 = the student was admitted
admit = 0 = the student was not admitted

46.6% of the students in the data were admitted.

The rank variable represents the prestige of the undergraduate institution on a scale from 1-4.



11. Since our dependent variable is binary, we'll use logistic regression. I'll first see if GRE scores predict admission. To do this, I entered the following model into R and obtained the following estimates:

Entered into R:

```
admit.gre <- glm(admit ~ gre, data=admit, family="binomial")
```

Output:

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)   95% CI
(Intercept) -2.901344   0.606038  -4.787 1.69e-06   (-4.09, -1.71)
gre          0.003582   0.000986   3.633  0.00028   (0.002, 0.005)
```

Note the p-values come from a Wald Test: http://en.wikipedia.org/wiki/Wald_test

Write out this estimated model: ln(odds of admission) = ______________________________

Based on the p-value and/or confidence interval, what can we conclude about GRE scores?

The mean GRE score (for the old version) was 500 with a standard deviation of 100. Given a student has a GRE score of 400, what are the odds that the student is admitted into this graduate school? Remember that the computer estimates coefficients for the **log-odds**.

What's the **probability** a student with a GRE score of 400 gets admitted?

12. This time, calculate the odds of admission for a student with a GRE score of 401. Convert that to a probability.

13. Now that we know the odds of admission for scores of 400 and 401, calculate the odds ratio. How much higher are the odds of admission for a student with a score of 401?

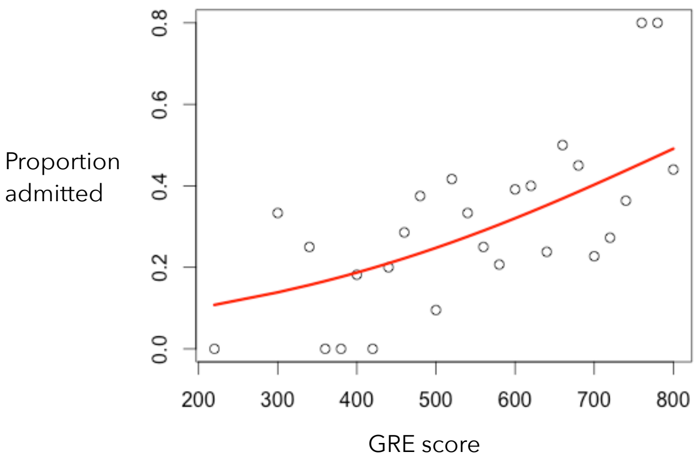
14. The computer output shows the coefficient of GRE is 0.003582. Calculate the exp(0.003582) and interpret what that coefficient represents. Below, I've pasted output from R when I asked for odds ratios from this data. Interpret.

```
                   OR      2.5 %    97.5 %
(Intercept) 0.0549493 0.01624471 0.1755632
gre         1.0035886 1.00168137 1.0055682
```

15. We could calculate the odds and probabilities of admission for various GRE scores. Verify some of these values.

GRE	Odds	P(admit)
200	0.112	0.101
300	0.161	0.139
400	0.23	0.187
500	0.329	0.248
600	0.471	0.32
700	0.674	0.403
800	0.965	0.491

Below, I've plotted the proportion of students admitted as a function of GRE scores. The red line represents our fitted logistic model. Evaluate how well the model fit the data.



We'll come back to this admissions example in a bit. For now, let's see another application of logistic regression.

Scenario: On January 28, 1986, the space shuttle Challenger exploded and seven astronauts died because two rubber O-rings leaked. These rings had lost their resiliency because the shuttle was launched on a very cold day. Ambient temperatures were in the low 30s and the O-rings themselves were much colder, less than 20 degrees Fahrenheit.

Flight	Temp	Damage
1	66	0
2	69	0
3	67	0
4	70	1
…	…	…
25	31	1

The table to the right displays a sample of the data from 25 trial launches of the space shuttle. The O-rings from each launch were examined to see if they sustained damage (1) or no damage (0) from the launch. The ambient temperature at the time of the launch was also recorded.

The estimated coefficients for a logistic regression analysis are displayed below:

.

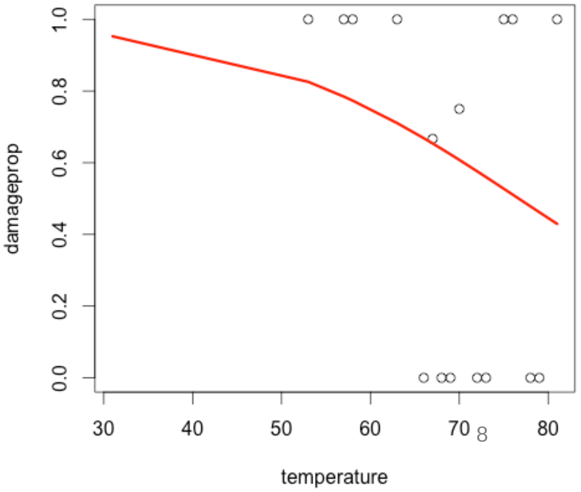
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    95% Confidence Interval
(Intercept)  5.03663    4.81697   1.046    0.296    (-4.404, 14.478)
temp        -0.06569    0.06819  -0.963    0.335    (-0.199, 0.0680)
```

Source: Tufte, E. (1997). Visual and Statistical Thinking: Displays of Evidence for Making Decisions.

16. Interpret the coefficient of temperature in our model.

17. Calculate the probability of sustaining damage if the temperature were 30-degrees. Then, calculate the probability if the temperature were 60-degrees. Interpret the relative probability.

18. Based on the output pasted above or the plot to the right, evaluate how well this model fits the data.



Scenario: Approximately 2-3% of Americans have *dermatophyte onychomycosis* (a toenail infection). The infection is caused by a fungus and does not only disfigure the nails but can also cause physical pain and impair the ability to work.

Researchers conducted a randomized, double-blind clinical trial of treatments for this toenail infection. 378 patients were randomly allocated into two oral antifungal treatments:

- 250 mg/day terbinafine, or
- 200 mg/day itraconazole.

The patients were them monitored over time. The data generated from this study include:

y = onycholysis (0 = none or mild; 1 = moderate or severe)
x_1 = treatment (0 = itraconazole; 1 = terbinafine)
x_2 = month = number of months since first treatment

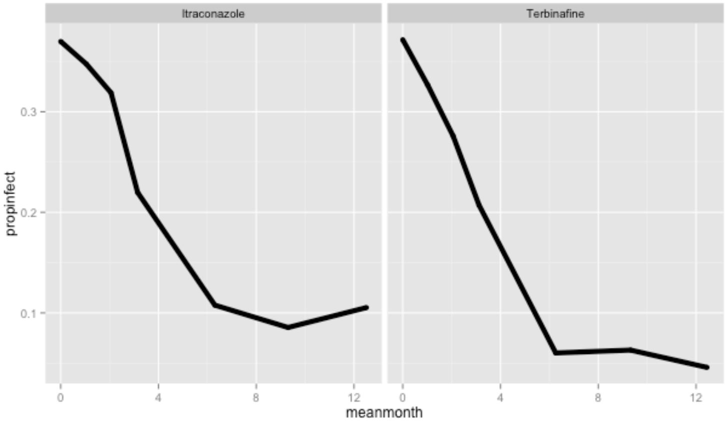
The main research question was whether the treatments differ in their efficacy. In other words, do patients receiving one treatment experience a greater decrease in their probability of having onycholysis than those receiving the other treatment?

Source: De Backer, M. et al. (1998). Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: a double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day. Journal of the American Academy of Dermatology. 38: 57-63.

19. Before we begin, let's take a look at the proportion of patients in each treatment with toenail infections over time.

Based on this plot, which treatment appears to be more effective?

From this plot, how can we tell what proportion of the 378 patients even had the toenail infection?



20. I used R to estimate the coefficients of following logistic regression model:

$$\ln(\text{odds}) = b_0 + b_1(\text{Terbinafine}) + b_2(\text{month}) + b_3(\text{Terbinafine})(\text{month})$$

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)   95% CI
(Intercept)                 -0.5566273  0.1089628  -5.108 3.25e-07   (-0.770, -0.343)
treatmentTerbinafine        -0.0005817  0.1561463  -0.004   0.9970   (-0.307,  0.305)
month                       -0.1703078  0.0236199  -7.210 5.58e-13   (-0.217, -0.124)
treatmentTerbinafine:month -0.0672216  0.0375235  -1.791   0.0732   (-0.141,  0.006)

                                  OR     2.5 %    97.5 %
(Intercept)                 0.5731389 0.4621244 0.7085917
treatmentTerbinafine        0.9994185 0.7358422 1.3574609
month                       0.8434052 0.8039199 0.8820652
treatmentTerbinafine:month 0.9349880 0.8676827 1.0055790
```

21. Let's rearrange some of the terms in this model to see if we can gain a better understanding.

$$\text{Original model: } \ln(\text{odds}) = b_0 + b_1(\text{Terbinafine}) + b_2(\text{month}) + b_3(\text{Terbinafine x month})$$

Model for itraconazole (terbinafine = 0)

$$\ln(\text{odds}) = b_0 + b_1(0) + b_2(\text{month}) + b_3(0 \text{ x month})$$
$$\ln(\text{odds}) = b_0 + b_2(\text{month})$$

Model for terbinafine (terbinafine = 1)

$$\ln(\text{odds}) = b_0 + b_1(1) + b_2(\text{month}) + b_3(1 \text{ x month})$$
$$\ln(\text{odds}) = b_0 + b_1 + b_2(\text{month}) + b_3(\text{month})$$
$$\ln(\text{odds}) = (b_0 + b_1) + (b_2 + b_3)(\text{month})$$

Based on these rewritten models, which model parameters represent the increased effectiveness of terbinafine?

Odds itraconazole (terbinafine = 0)

Odds for terbinafine (terbinafine = 1)

Odds ratio

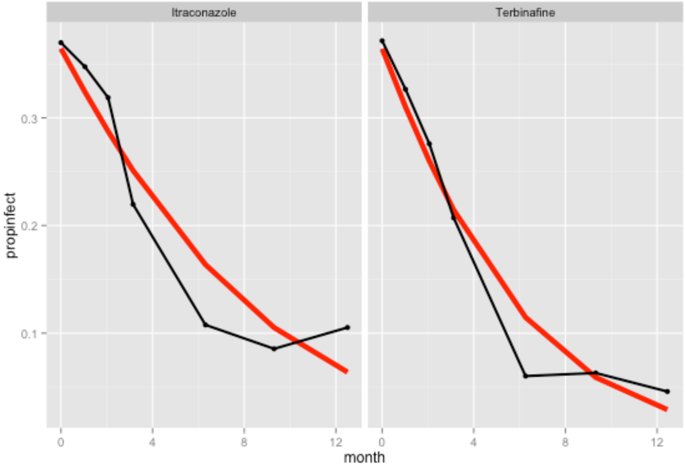
22. Sketch a graph of the odds ratio as a function of time. Try the reciprocal so you're comparing the odds of infection for terbinafine compare to the odds of infection for itraconazole. How can we interpret this graph?

23. Calculate the odds ratio comparing the effectiveness of the treatments at 0 months. What does this represent?

24. Calculate the odds ratio comparing the effectiveness of the treatments at 15 months.

25. Calculate and interpret the relative probability comparing the effectiveness of the treatments at 15 months.

26. Using R, I plotted our logistic model on top of the proportion of patients with infections at each month. Evaluate the fit of the model.



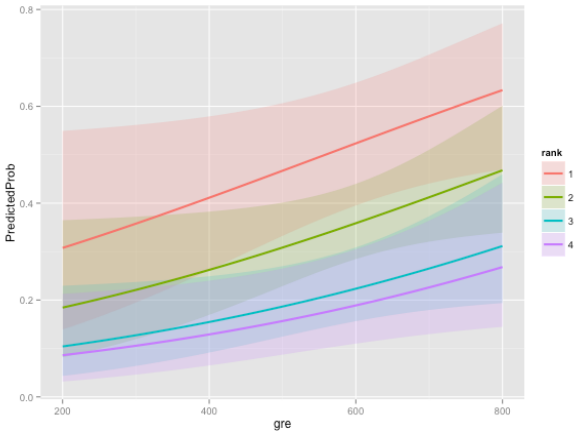
27. Let's go back to our graduate admissions example. We've already investigated the extent to which GRE scores predict admissions. Let's try to incorporate more predictors into our model.

Let's try a full model of: $\ln(\text{odds}) = b_0 + b_1(\text{GRE}) + b_2(\text{GPA}) + b_3(\text{rank})$

Interpret:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)   95% CI
(Intercept) -3.989979   1.139951  -3.500 0.000465   (-6.223, -1.756)
gre          0.002264   0.001094   2.070 0.038465   ( 0.0001, 0.004)
gpa          0.804038   0.331819   2.423 0.015388   ( 0.154,  1.454)
rank2       -0.675443   0.316490  -2.134 0.032829   (-1.296, -0.055)
rank3       -1.340204   0.345306  -3.881 0.000104   (-2.017, -0.663)
rank4       -1.551464   0.417832  -3.713 0.000205   (-2.370, -0.733)
```

```
                   OR       2.5 %    97.5 %
(Intercept) 0.0185001 0.001889165 0.1665354
gre         1.0022670 1.000137602 1.0044457
gpa         2.2345448 1.173858216 4.3238349
rank2       0.5089310 0.272289674 0.9448343
rank3       0.2617923 0.131641717 0.5115181
rank4       0.2119375 0.090715546 0.4706961
```



To see if the rank of a student's undergraduate school improves our prediction, we can conduct a test on terms 4-6 in our model. Interpret:

```
wald.test(b = coef(admitlogit), Sigma = vcov(admitlogit), Terms = 4:6)
Wald test: Chi-squared test:
X2 = 20.9, df = 3, P(> X2) = 0.00011
```

We could also compare rank 2 institutions to rank 3. Interpret:

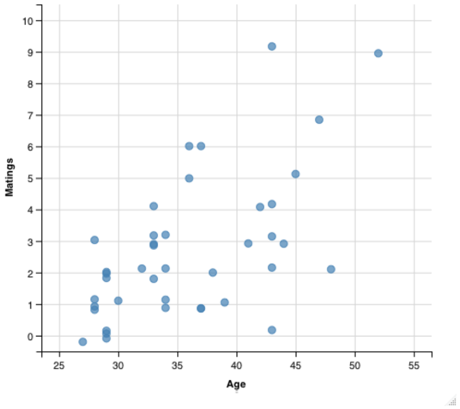
```
wald.test(b = coef(admitlogit), Sigma = vcov(admitlogit), L = l)
Wald test: Chi-squared test:
X2 = 5.5, df = 1, P(> X2) = 0.019
```

Finally, we can evaluate the fit of our model. If we compare it to a null model, we get a p-value of 0.000000076.

In the assignment for this activity, you'll figure out the factors that influenced the odds of surviving the Titanic. You'll also try to develop (and test) a model to predict the retention of this year's freshman class.

Scenario: When elephants reach maturity (around the age of 14), they have to compete with all adult males for mating opportunities. Elephants get bigger as they get older -- and females are generally more receptive to larger males -- so the number of matings should increase with age.

We're going to model the number of matings as a function of the age of the males.



Elephant #	Age	# of Matings
1	27	0
2	28	1
…	…	…
41	52	9
Mean =	35.85	2.682927
Variance =	43.28	5.071951

Source: gpk package: http://cran.r-project.org/web/packages/gpk/gpk.pdf

28. Here we have a scenario in which our dependent variable represents a count (number of matings). Counts tend to be: (1) discrete, (2) positive, (3) positively skewed with a high proportion of zeros.

If we try to fit an ordinary least squares regression line, we'll run into problems:

(1) the relationship between X and Y is nonlinear
(2) counts tend to be heteroskedastic
(3) our line will predict negative values (which cannot exist).

We can use the generalized linear model to predict counts. To do so, we'll use a natural log link function:

$$g\left(E\left[y\right]\right) = \ln\left(\mu_y\right) = b_0 + b_1 x_1$$

This link function ensures our predicted values for y are positive and positively skewed. To see this more clearly, we can use the exponential function:

$$e^{\ln\left(\mu_y\right)} = e^{b_0 + b_1 x_1}$$

$$\mu_y = e^{b_0 + b_1 x_1}$$

When we use this log link function, we're conducting a *Poisson regression*. If you took MATH 300, you might remember the Poisson distribution. We used it to calculate the probability of eating at least 5 bug parts when we eat peanut butter. We learned the Poisson distribution is like a binomial distribution with an infinite number of trials:

Poisson Distribution Conditions:
- P(# number of events occur in a fixed interval of time, space, distance, volume)
- The events occur with a known average rate
- The events occur independently of the time since the last event
- $E[x] = Var[x] = \lambda$

PMF: $P(X = x) = \left(\dfrac{\lambda^x}{x!}\right)e^{-\lambda}$

One thing to note is that in a Poisson distribution, the mean is assumed to equal the variance.

29. Based on the scatterplot for our elephant data, it looks as though when age increases:
    a. the (mean) number of matings increases
    b. the variability (dispersion) in number of matings increases

Since the dispersion increases with the mean for our count dependent variable, Poisson regression might be a good choice. We can then model: $\ln\left(\text{matings}\right) = b_0 + b_1\left(\text{age}\right)$

Before I do that, let's fit a null model: $\ln\left(\text{matings}\right) = b_0$

I used R to estimate the coefficients of this null model:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.98691    0.09535   10.35   <2e-16 ***
---
    Null deviance: 75.372  on 40  degrees of freedom
AIC: 178.82
```

So our null model is: $\ln\left(\text{matings}\right) = 0.98691$. Convert this to predict the number of matings under this null model. What does this number represent?

30. The deviance of 75.372 is a measure of how poorly our model fits the data. If the model fits the data, then we would expect the deviance to be approximately equal to the degrees of freedom. We could also compare the deviance to a chi-squared distribution to get a sense of model fit. Based on this value, does our null model fit the data well?

31. I then had R estimate the coefficients of a model including the age predictor using an iterative process:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.58201    0.54462  -2.905  0.00368
Age          0.06869    0.01375   4.997 5.81e-07
---
    Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.012  on 39  degrees of freedom
AIC: 156.46
```

Write out this model. I then used the model to predict the average number of matings at ages 30, 31, and 45. By what amount are the matings increasing each year? What are the predicted variances in matings at each age?

$$\text{matings at 30 years} = e^{-1.58201+0.06869(30)} = 1.614098$$

$$\text{matings at 31 years} = e^{-1.58201+0.06869(31)} = 1.728872$$

$$\text{matings at 45 years} = e^{-1.58201+0.06869(45)} = 4.522968$$

32. In general, we can interpret the parameters of a Poisson regression like this:

b_0: the mean of the Poisson distribution when our predictor equals zero
b_1: a 1-unit increase in X is associated with an e^{b1} increase in the expected count of Y

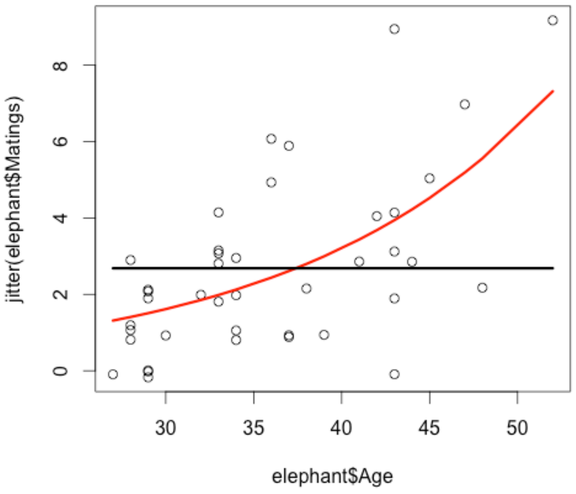
In our example, the b_0 parameter doesn't make sense, since age=0 is not meaningful.

What does the b_1 parameter represent in our scenario? Interpret that coefficient of 0.06869.

33. We can plot our predictions (from both the null and age-based models) on top of our data. Obviously, the model with age as a predictor appears to better fit the data.

We can also compare models by comparing the decrease in deviance values from our null to full models.

Using R, I conducted a chi-squared test that the difference in deviance values between our two models is zero. R reported a p-value of 0.000000799. Interpret this value.



Scenario: University professors typically need to publish articles in order to earn promotions and tenure. The number of articles a professor publishes could be predicted by a number of factors. In this scenario, we'll have the following data:

- art = number of articles published
- fem = gender
- mar = marriage status
- kid5 = # of kids under age 6
- phd = prestige of PhD program

Professor	art	fem	mar	kid5	phd
1	0	man	married	0	2.52
2	2	woman	single	1	2.05
…		…			…
915	19	man	married	0	1.86
Mean =	1.693				
Variance =	3.710				

Source: Long (1990)

34. To get a sense of our data, go ahead and predict the relationship between each predictor and our dependent variable. Explain why we shouldn't use ordinary least squares regression on this data.

35. Take a look at the mean and variance of our dependent variable. Should we conduct a Poisson regression?

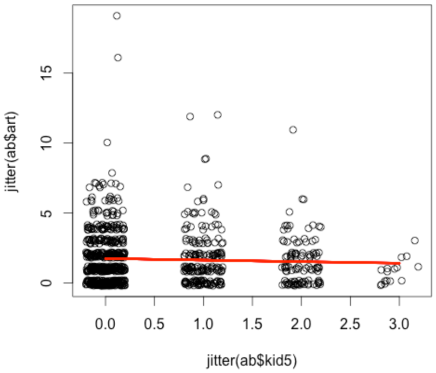
36. If we were to fit a null model with no predictors, what would be the value of our b_0 coefficient? Explain.

37. Using Poisson regression, I fit a series of models. Interpret the coefficients and evaluate model fit.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.63265    0.03279  19.293  < 2e-16 ***
femWomen    -0.24718    0.05187  -4.765 1.89e-06 ***
---

    Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1794.4  on 913  degrees of freedom
AIC: 3466.1

Exponentiated coefficients:
(Intercept)    femWomen
  1.8825911   0.7810027
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.63902    0.03513  18.190  < 2e-16 ***
femWomen    -0.24054    0.05353  -4.493 7.01e-06 ***
marSingle   -0.02815    0.05632  -0.500    0.617
---

    Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1794.1  on 912  degrees of freedom
AIC: 3467.9

Exponentiated coefficients:
(Intercept)    femWomen   marSingle
  1.8946194   0.7862031   0.9722457
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.55960    0.02988  18.728   <2e-16 ***
kid5        -0.06978    0.03450  -2.023   0.0431 *
---

    Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1813.2  on 913  degrees of freedom
AIC: 3485

Exponentiated coefficients:
(Intercept)        kid5
  1.7499725   0.9325973
```



38. Finally, I fit a model with all our predictors. Interpret the coefficients and evaluate model fit.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.50776    0.09411   5.395 6.85e-08 ***
femWomen    -0.27717    0.05428  -5.107 3.28e-07 ***
marSingle   -0.15211    0.06112  -2.489  0.01282 *
kid5        -0.16151    0.03934  -4.105 4.04e-05 ***
phd          0.08410    0.02602   3.233  0.00123 **
---

    Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1766.2  on 910  degrees of freedom
AIC: 3444

Exponentiated coefficients:
(Intercept)    femWomen   marSingle        kid5         phd
  1.6615728   0.7579236   0.8588924   0.8508577   1.0877412
```

Why does this model still not fit our data well? What could we do about this?

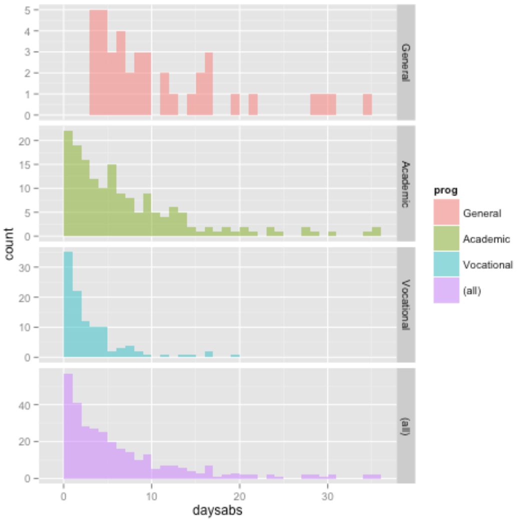
We can use our model to make predictions:

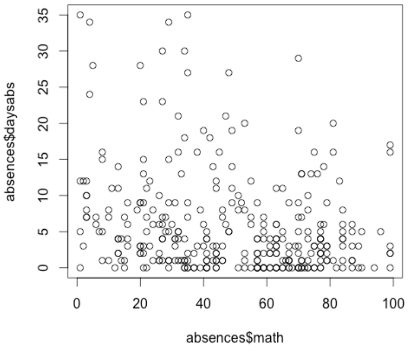
- Single female with 1 kid under the age of 6: Predicted publications = 1.0166
- Married man with no kids under the age of 6: Predicted publications = 2.1571

Scenario: Suppose we want to predict the number absences for students at a high school. We have this data:

- math = math test score
- daysabs = days absent
- prog = type of program

Student	gender	math	daysabs	prog
1	male	63	4	academic
2	female	20	2	general
…		…		
314	female	77	2	vocational
Mean =			5.955	
Variance =			49.519	





Source: UCLA TISE

39. Even though the variance is much higher than the mean (even within each program type), let’s try to fit a Poisson regression model with all our predictors.

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.7594786  0.0637731  43.270  < 2e-16 ***
gendermale     -0.2424762  0.0467765  -5.184 2.18e-07 ***
math           -0.0069561  0.0009354  -7.437 1.03e-13 ***
progAcademic   -0.4260327  0.0567308  -7.510 5.92e-14 ***
progVocational -1.2707199  0.0779143 -16.309  < 2e-16 ***
---

    Null deviance: 2217.7  on 313  degrees of freedom
Residual deviance: 1746.8  on 309  degrees of freedom
AIC: 2640.2

Exponentiated coefficients:
   (Intercept)     gendermale           math   progAcademic progVocational
    15.7916072      0.7846824      0.9930680      0.6530950      0.2806295
```

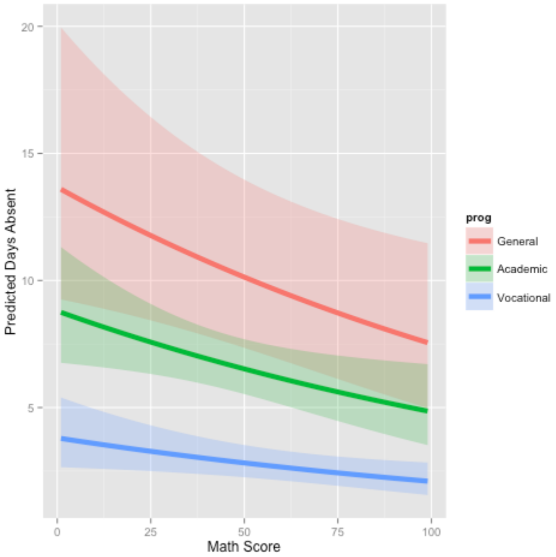
The fit is terrible because of that over-dispersion. So what could we do about this? One thing we could do is fit a negative binomial regression model. The negative binomial model is just like the Poisson model except it includes an extra parameter to model dispersion.

I fit this model using R and found:

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.707484   0.204275  13.254  < 2e-16 ***
gendermale     -0.211086   0.121989  -1.730   0.0836 .
math           -0.006236   0.002492  -2.502   0.0124 *
progAcademic   -0.424540   0.181725  -2.336   0.0195 *
progVocational -1.252615   0.199699  -6.273 3.55e-10 ***
---
(Dispersion parameter for Negative Binomial(1.0473) family taken to be 1)

    Null deviance: 431.67  on 313  degrees of freedom
Residual deviance: 358.87  on 309  degrees of freedom
AIC: 1740.3

Exponentiated coefficients:
   (Intercept)     gendermale           math   progAcademic progVocational
    14.9915148      0.8097046      0.9937839      0.6540708      0.2857565
```

The deviance decreased by a large amount, so this model does fit our data better.

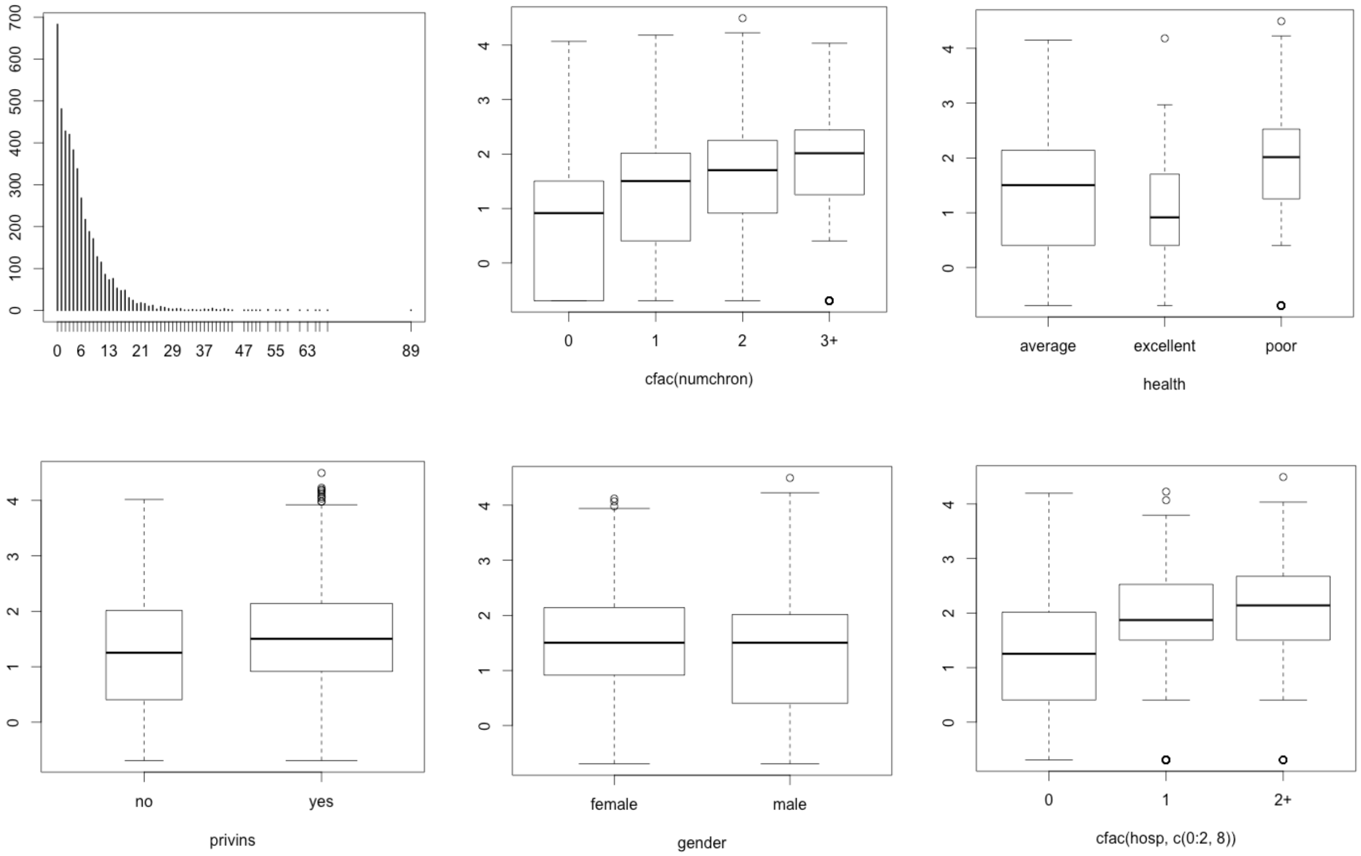
Suppose we’re interested in seeing the the type of program (academic, general, vocational) predicts absences. To do so, we can fit a reduced model that does not include the program predictor. We can then test the difference in deviance between the models. I did this in R and found a p-value of 0.000000000313. Interpret.



Scenario: Suppose we want to predict the number of times an individual visits a doctor's office.

Data from the U.S. National Medical Expenditure Survey (NMES) for 1987-88 include:

- ofp = number of physician office visits
- hosp = number of hospital stays
- health = self-perceived health status
- numchron = number of chronic conditions
- gender
- school = years of education
- privins = private insurance?



40. We can fit a Poisson regression model with all our predictors, but we'll have an over-dispersion problem (too many zeros in our data). We could also try other methods, such as quasi-poisson regression, zero-inflated poisson regression, or negative binomial regression.

    On the next page, I've pasted output comparing the coefficients and standard errors for each model.

    Below that, I've shown how many zeros each model predicts (compared to the observed number of zeros).

    Which model fits best?

```
                    ML-Pois  Quasi-Pois          NB
(Intercept)      1.02887420  1.02887420  0.92925658
hosp             0.16479739  0.16479739  0.21777223
healthexcellent -0.36199320 -0.36199320 -0.34180660
healthpoor       0.24830697  0.24830697  0.30501303
numchron         0.14663928  0.14663928  0.17491552
gendermale      -0.11231992 -0.11231992 -0.12648813
school           0.02614299  0.02614299  0.02681508
privinsyes       0.20168688  0.20168688  0.22440187
```

```
                    ML-Pois    Adj-Pois  Quasi-Pois          NB
(Intercept)     0.023784601 0.064529808 0.061593641 0.054591271
hosp            0.005997367 0.021945186 0.015531043 0.020176492
healthexcellent 0.030303905 0.077448586 0.078476316 0.060923623
healthpoor      0.017844531 0.054021990 0.046210977 0.048510797
numchron        0.004579677 0.012907865 0.011859732 0.012091749
gendermale      0.012945146 0.035343487 0.033523316 0.031215523
school          0.001843329 0.005084002 0.004773565 0.004393971
privinsyes      0.016859826 0.043128006 0.043660942 0.039463744
```

```
   Obs ML-Pois      NB
   683      47     608
```