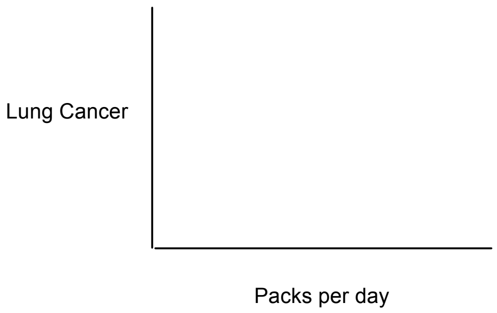
Activity #19: Logistic Regression

Resources: Placement.sav

Thus far, we've used regression to predict the value of a dependent variable based on a (possibly nonlinear) combination of independent variables. Every dependent variable we've looked at (GPA, SAT scores, Neuron activity, NBA Performance) has been a continuous variable.

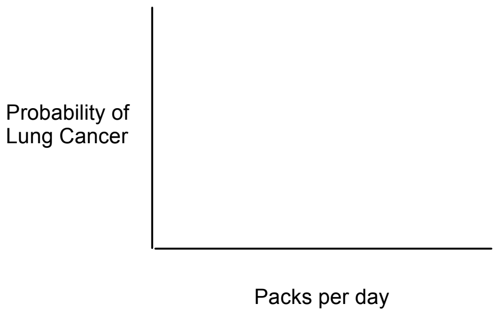
We can use a modified form of linear regression to predict values of a binary dependent variable (a dependent variable that can only take on two possible values). This procedure is called *logistic regression*.

1) Suppose we want to predict whether or not an individual will develop lung cancer. It might be reasonable to assume that the number of packs of cigarettes smoked each day might be a good predictor. If we sampled 100 individuals, what would the scatterplot of lung cancer vs. cigarette packs look like? Sketch it below.

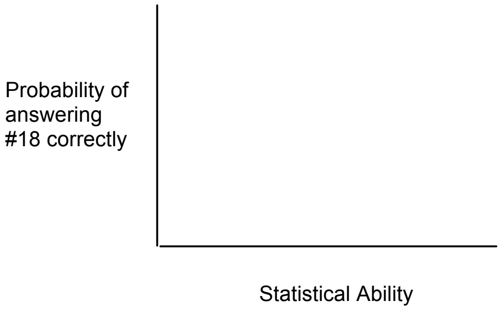


2) Can you fit a regression line to that scatterplot? How accurate is your line at predicting lung cancer?

3) Traditional linear regression won't work for binary dependent variables. We're not really interested in predicting a "score" on the dependent variable (what does a predicted lung cancer of 0.43 mean?). We're more interested in predicting the probability of an individual developing lung cancer based on the number of packs they smoke per day. What kind of model can be fit to the following set of axes?



4) Suppose we have another situation. Suppose I know your underlying statistical ability. Could I use that to predict the probability that you will answer item #18 on the final exam correctly? What would that function look like? What if I wanted to predict your performance on an easier item?

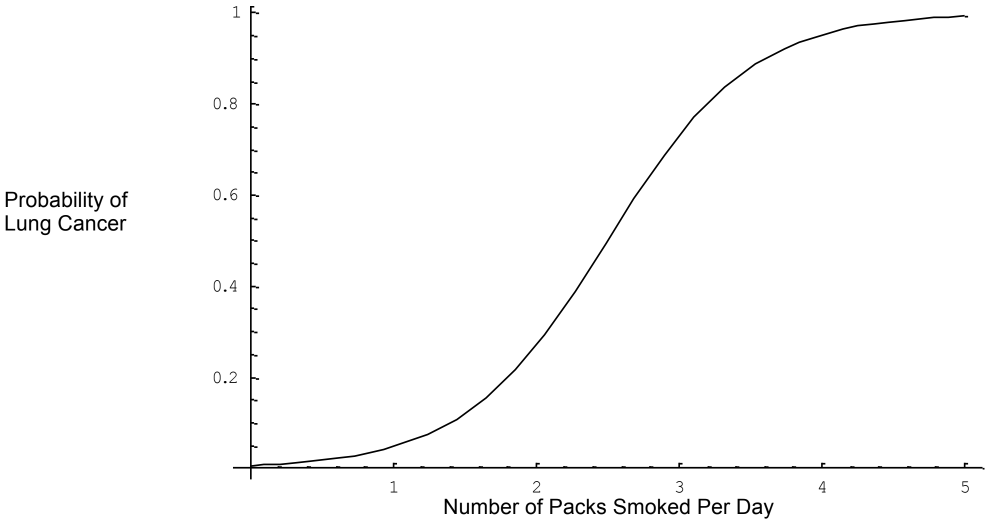


The relationship between a set of predictors and a binary dependent variable can be modeled by a **logistic function** of the form:

$$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$$

In our lung cancer example, $P(\text{lung cancer})=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$, where X_1 represents the number of packs smoked per day. We would use a computer to find the best values for the beta coefficients.

The graph of a logistic function typically looks like this: (note: this is not based on any actual lung cancer data)



Logistic functions have the following properties:

1. They are asymptotic with respect to P(Y) = 0 and P(Y) = 1. This is because probabilities range from zero to one.
2. They are monotonically increasing (because we would expect higher values of X to correspond with higher probabilities)
3. They are continuous.

How do we find the beta values in the logistic function? To answer this, we need to take a quick detour through the world of odds.

5) 17,096 students were surveyed with respect to their drinking habits. 3,314 of the students identified themselves as "frequent binge drinkers." If this sample is representative of students at SAU, what's the probability that a SAU student is a binge drinker?

Recall that we calculate odds as: $\text{ODDS} = \dfrac{\text{P(event)}}{\text{1 - P(event)}}$.

6) What are the odds that a student is a binge drinker? What are the odds that a student is not a binge drinker? Interpret these odds.

In a logistic function, the probability of an event is modeled by: $P(Y=1) = \dfrac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$

If we convert this into an odds function, we get: (we'll substitute z for the beta weights to simplify the notation)

$$\text{ODDS} = \frac{\text{P(event)}}{\text{1 - P(event)}} = \frac{\frac{1}{1+e^{-Z}}}{1-\frac{1}{1+e^{-Z}}} = \frac{\frac{1}{1+e^{-Z}}}{\frac{1+e^{-Z}}{1+e^{-Z}} - \frac{1}{1+e^{-Z}}} = \frac{\frac{1}{1+e^{-Z}}}{\frac{1+e^{-Z}-1}{1+e^{-Z}}} = \frac{1+e^{-Z}}{(1+e^{-Z})(e^{-Z})} = \frac{1}{e^{-Z}} = e^{Z}$$

If we take the natural logarithm of the odds, we get the *log odds*:

$$\text{LN(ODDS)} = LN\left(\frac{\text{P(event)}}{\text{1 - P(event)}}\right) = \ln(e^{Z}) = Z = \beta_0 + \beta_1 X_1$$

Notice that the log odds function is a linear function. Thus, we can use the concept of linear regression to find the beta values for the log odds. We then must simply convert log odds into regular odds and then into probabilities to find our logistic regression model.

Let's go through a quick example of this using our binge drinking example:

7) The following table summarizes the results of the binge drinking survey. Calculate the probabilities, odds, and log odds for males and females:

	Binge Drinker	Not Binge Drinker	Total
Males	1624	5528	7152
Females	1690	8254	9944
Total	3,314	13,782	17,096

Women		**Men**

8) We're trying to use gender to predict binge drinking. If we code the gender variable as: GENDER = 0 for females and GENDER = 1 for males, we get the following results:

Women		**Men**
$\text{Ln(Odds)} = LN(.205) = -1.59$		$\text{Ln(Odds)} = LN(.294) = -1.23$
Recall: $\text{LN(ODDS)} = \beta_0 + \beta_1 X_1$		Recall: $\text{LN(ODDS)} = \beta_0 + \beta_1 X_1$

Logistic Regression Inference:

1. To find a confidence interval for β_1: $b_1 \pm z(SE_{b_1})$

2. To find a confidence interval for the odds ratio: $e^{b_1 \pm z(SE_{b_1})}$

3. To test $H_0 : \beta_1 = 0$: $\chi_1^2 \sim \left(\frac{b_1}{SE_{b_1}} \right)^2$

The Math Department here at SAU is responsible for placing freshmen in the most appropriate courses. Based on scores from a 10-point placement test, students are either admitted or denied admission into MATH 151 College Algebra. Last year, 20 students took the placement test. Their scores and admission decisions are displayed in the following table:

Y (admission)	X (Placement Score)
0	2
0	2
0	2
0	3
0	3
1	3
0	3
0	3
0	4
0	4
1	4
1	5
1	5
1	5
1	5
0	7
1	7
1	7
1	8
1	8

Note: Some low-scoring students were admitted into College Algebra due to high ACT scores
Some high-scoring students were not admitted into College Algebra due to low ACT scores and poor high school grades

I had a computer run a logistic regression on this data (beta coefficients are calculated via maximum likelihood estimates). The following output was obtained:

Logistic Function: $\mathrm{L}N(ODDS) = \text{-}4.095 + 0.946(\mathrm{X}_1)$

$$SE_{b_1} = 0.423$$

We can use these results to answer the following questions:

9) Does a student's placement test score predict the probability that a student is admitted into College Algebra? To answer this question, we need to test the significance of the b_1 coefficient.

$$\chi_1^2 \sim \left(\frac{b_1}{SE_{b_1}}\right)^2 = \left(\frac{0.946}{0.423}\right)^2 = 5.002$$

This is significant at the 0.05 level. We conclude that placement scores predict admission into College Algebra

10) Given a student has a placement score of 6, what are the odds that the student is admitted into College Algebra? To answer this question, we need to remember the following relationship:

$$\text{LN(ODDS)} = \text{LN}\left(\frac{\text{P(event)}}{\text{1 - P(event)}}\right) = \ln\left(e^{\beta_0+\beta_1 X_1}\right) = \beta_0 + \beta_1 X_1$$

11) Given a student has a placement score of 6, what is the probability that the student is admitted into College Algebra? To answer this question, we need to remember the following relationship:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$$

12) The following table displays the odds and probabilities of students being admitted into College Algebra for students receiving placement scores from 1-10:

Placement Score	Odds of Admission	Probability of Admission
1	0.04	0.04
2	0.11	0.10
3	0.28	0.22
4	0.73	0.42
5	1.89	0.65
6	4.85	0.83
7	12.50	0.93
8	32.20	0.97
9	82.93	0.99
10	213.58	1.00

Calculate and interpret the odds ratio for students earning placement scores of 6 versus 4. Calculate the relative probability.

Another Logistic Regression Example

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.

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

The following table displays 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.

A logistic regression analysis yielded the results to the right:

FLIGHT	TEMP	DAMAGE
STS-1	66	0
STS-3	69	0
STS-5	68	0
STS-7	72	0
STS-8	73	0
STS-9	70	0
STS_41-G	78	0
STS_51-A	67	0
STS_51-J	79	0
STS-2	70	1
STS-6	67	1
STS_41-B	57	1
STS_51-F	81	1
STS_51-I	76	1
STS_61-B	76	1
STS_41-C	63	1
STS_41-D	70	1
STS_51-C	53	1
STS_51-D	67	1
STS_51-B	75	1
STS_51-G	70	1
STS_61-A	75	1
STS_61-C	58	1
STS-4	80	(missing)
STS_51-L	31	(missing)

```
-----------------------------------------------------------------
         any |      Coef.   Std. Err.       z    P>|z|
-------------+---------------------------------------------------
        temp |  -.0656892    .0681892    -0.96   0.335
       _cons |   5.036627    4.816972     1.05   0.296
-----------------------------------------------------------------
```

$$\mathrm{LN(ODDS)} = 5.036627 - 0.656892(\mathrm{temp})$$

```
Number of obs   =      23
LR chi2(1)      =    1.01
Prob > chi2     = 0.3145
Pseudo R2       = 0.0329
```

13) Calculate the odds that the O-rings would have been damaged if the ambient temperature was 30 degrees. Convert these odds into a probability. Calculate the odds ratio for a 30-degree day compared to a 60-degree day.

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.

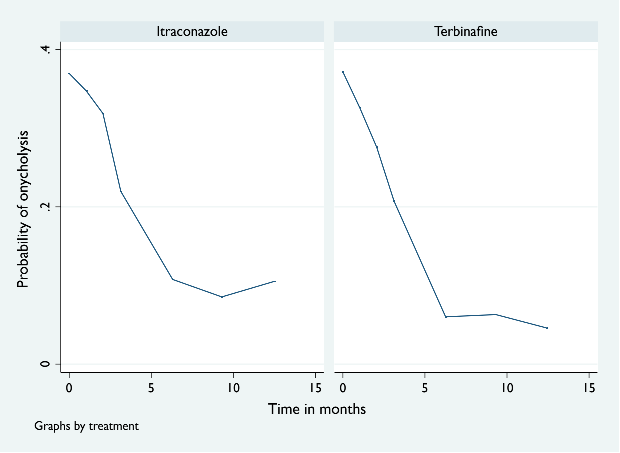
14) Approximately 2-3% of Americans have *dermatophyte onychomycosis* (toenail infection). The infection is caused by a fungus and does not only disfigure the nails but can also cause physical pain and impair the ability to work. Researchers conducted a randomized, double-blind clinical trial of treatments for *dermatophyte onychomycosis* (toenail infection). 378 patients were randomly allocated into two oral antifungal treatments (250 mg/day terbinafine and 200 mg/day itraconazole) and evaluated over time.

The variables of interest in this study are:

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 is 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?

Before we begin our analysis, let’s examine a plot of the proportion of patients in each treatment with toenail infections over time. What can we conclude?



15) I had Stata fit the following model to the data:

$$\text{logit}\{\text{P}(\text{y}_{ij} = 1 \middle| x_{ij}) = \ln(\text{odds of onycholysis}) = \text{b}_0 + b_1 x_1 + b_2 x_2 + b_3 x_2 x_3$$

I obtained the following output:

```
Logistic regression                          Number of obs   =       1908
                                             LR chi2(3)      =     164.47
                                             Prob > chi2     =     0.0000
Log likelihood = -908.00747                  Pseudo R2       =     0.0830

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |  -.0005817   .1561466    -0.00   0.997    -.3066235      .30546
       month |  -.1703078   .0236199    -7.21   0.000     -.216602   -.1240136
   trt_month |  -.0672216    .037524    -1.79   0.073    -.1407673    .0063242
       _cons |  -.5566273   .1089628    -5.11   0.000    -.7701904   -.3430642
------------------------------------------------------------------------------
```

Write out the best-fitting model:

16) Let's rearrange some terms in this model to see if we can gain a better understanding.

Original model: $\ln(\text{odds of onycholysis}) = \mathrm{b}_0 + b_1(\text{drug}) + b_2(\text{month}) + b_3(\text{drug x month})$

Model for itracconazole (drug = 0):

Model for terbinafine (drug = 1):

Simplified into "slope/intercept":

Simplified into "slope/intercept":

Based on these rewritten models, which model parameter(s) represent the increased effectiveness of terbinafine?

Odds for itracconazole (drug = 0):

Odds for terbinafine (drug = 1):

Odds ratio (odds for itraconazole / odds for terbinafine):

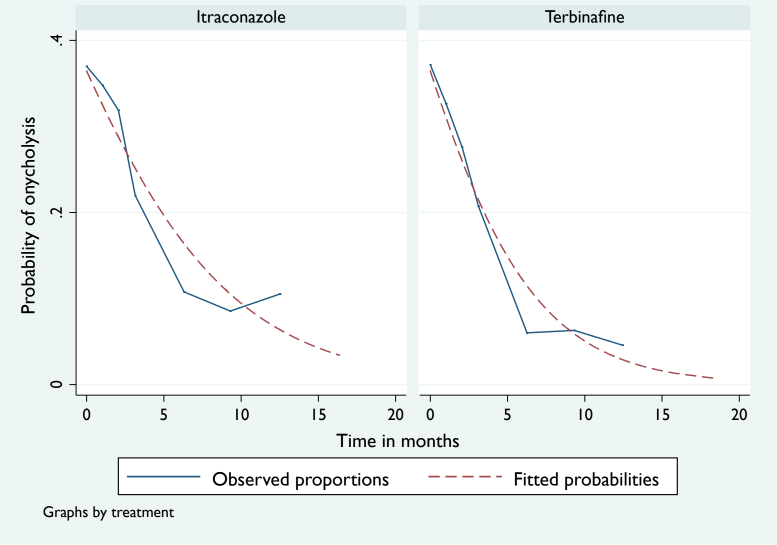
17) Sketch a graph of the odds ratio as a function of time. Try the reciprocal so that you are comparing the odds of infection for terbinafine compared to the odds of infection for itraconazole. How can we interpret this graph?

18) Calculate the odds ratio comparing the effectiveness of itraconazole to terbinafine at 0 months. What does this represent?

19) Calculate the odds ratio comparing the effectiveness of itraconazole to terbinafine at 15 months.

20) Calculate the relative probability comparing the effectiveness of itraconazole to terbinafine at 15 months.

Here is a graph showing our predicted probabilities (from our logistic model) to our observed proportions.



21) Recall from MATH 300 the concept of the standard error. Suppose we have an unknown population distribution with a mean of 500 and a standard deviation of 100. If we repeatedly sample 64 observations from that population and calculate the mean from each sample, the distribution of those means will…

Have a mean of: $\mu_{\overline{X}} = \mu_x = 500$

Have a standard error of: $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}} = \dfrac{100}{\sqrt{64}} = 12.5$

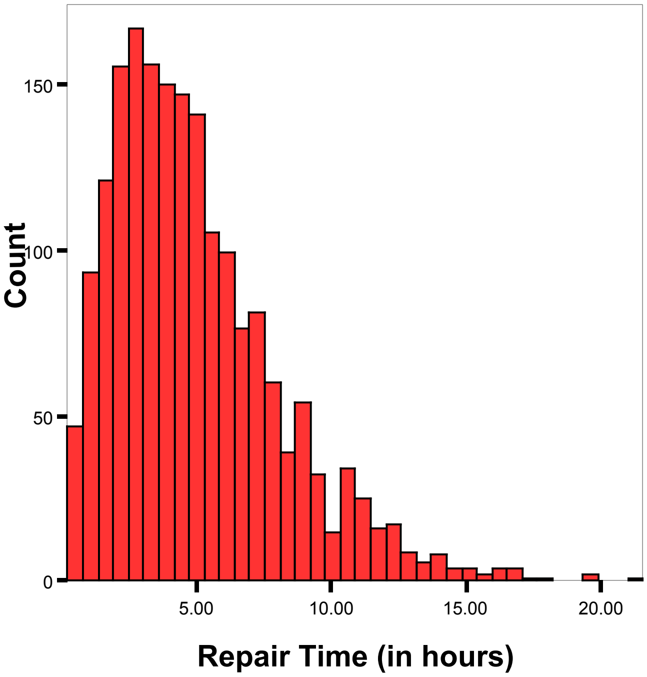
In calculating this standard error, we're assuming the Central Limit Theorem applies. How can we calculate the standard error if the CLT does not apply. In other words, how can we determine the standard error of a sampling distribution if our population is heavily skewed and/or we have a small sample size?

One way to estimate the standard error of any statistic (in this case, the sample mean) is to use the **bootstrap method**.

Example: In most states, many different companies offer local telephone service. The primary local company in each region is required to share its lines (for a fee) with its competitors.

Verizon is the primary company for a large area in the eastern U.S. As such, it must provide repair service for the customers of the other phone companies in this region. Some regulators are worried that if a customer of another phone company has a problem, Verizon may not make repairs as quickly as they would if it was a Verizon customer. The local Public Utilities Commission requires the use of significance tests to compare repair times for the two groups of customers.

At first glance, it might seem reasonable to run an independent samples t-test on this data (or possibly a regression analysis controlling for the distance a service request is from the company headquarters). Unfortunately, the distribution of repair times is far from normal. The following histogram displays the distribution of a random sample of 1876 repair times for Verizon's own customers. The distribution has a heavy positive skew. We would hesitate to run parametric tests on this data, especially since the sample sizes of customers from other phone companies is so small (compared to the number of Verizon customers).



Even though the CLT may not apply, we're still interested in determining the sampling distribution of means from this population. The next page demonstrates an example of the bootstrap method.

Suppose we administer a 10-item test to students in a statistics class. We observe the following scores: 2, 4, 5, 9

We want to determine the standard error of the sampling distribution of sample means from an unknown population of student test scores. If the CLT holds, we could calculate the standard error by noting:

Scores	Mean	Std. Deviation
2 4 5 9	5	2.55

We can then use results from the CLT to calculate: $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.55}{\sqrt{4}} = 1.275$

Bootstrap Method for determining the standard error of a statistic:

Step 1: Treat your sample as if it is the entire population. In other words, randomly sample *n* observations **with replacement** from your population of size *n*. In applying this method, we usually have a computer gather at least 1000 samples. Note that since we are sampling with replacement, the same observation could show up in our bootstrap sample multiple times.

In this simple example, we'll only gather 5 bootstrap samples. These are listed in the second column of the following table:

Bootstrap Sample	Sampled Observations	Sample Mean
1	4 2 9 4	4.75
2	2 9 5 2	4.50
3	5 5 4 9	5.75
4	5 2 4 2	3.25
5	9 9 5 5	7.00

Step 2: Calculate the sample mean for each bootstrap sample. These are recorded in the third column of the above table.

Step 3: We're interested in the standard error (which, conceptually, is the standard deviation of the sample means). This means we should calculate the standard deviation of the means calculated from the bootstrap samples.

Sample Mean	Standard Deviation
4.75	1.258967831
4.50	
5.75	
3.25	
4.25	

Step 4: This standard deviation is our estimate of the standard error. This example worked out nicely because I designed it that way. Normally with only 5 bootstrap replications, our estimated standard error would not be close to the actual standard error.