Chapter 12: Logistic Regression

So far we used linear models to predict a *continuous* reponse variable y using a set of *continuous* or *discrete* predictors $x_1, \dots, x_n$. Now we turn our attention to predicting a *discrete* reponse variable using both continuous and discrete predictors. There are many applications where the response variable is discrete, such as predicting 1.) if a team will win a sporting event, 2.) if a patient will have a heart attack, or 3.) if a medical device will fail.

Let's assume the response variable y has two possible outcomes, 0 and 1. (This is not really a limitation, as response variable with multiple outcomes can be modeled as a series of binary outcomes.) Linear models are unbounded, real-valued functions. The responses predicted by linear models can therefore range from negative to positive infinity. There is no way to use only field algebra or linear algebra to fit a function that predicts only two events. Instead, we use a *link function* to translate the continuous output of a linear model into a binary prediction.

Rather than predict the binary response, we could try predicting the probability that the response is equal to 1. Probabilities are continous, but they are bounded to the interval [0,1]. We don't have a method to force a linear models to only make predictions in this range. We could instead predict the *odds* of the response variable, which is the probability that the response equals 1 divided by the probability that the response equals 0.

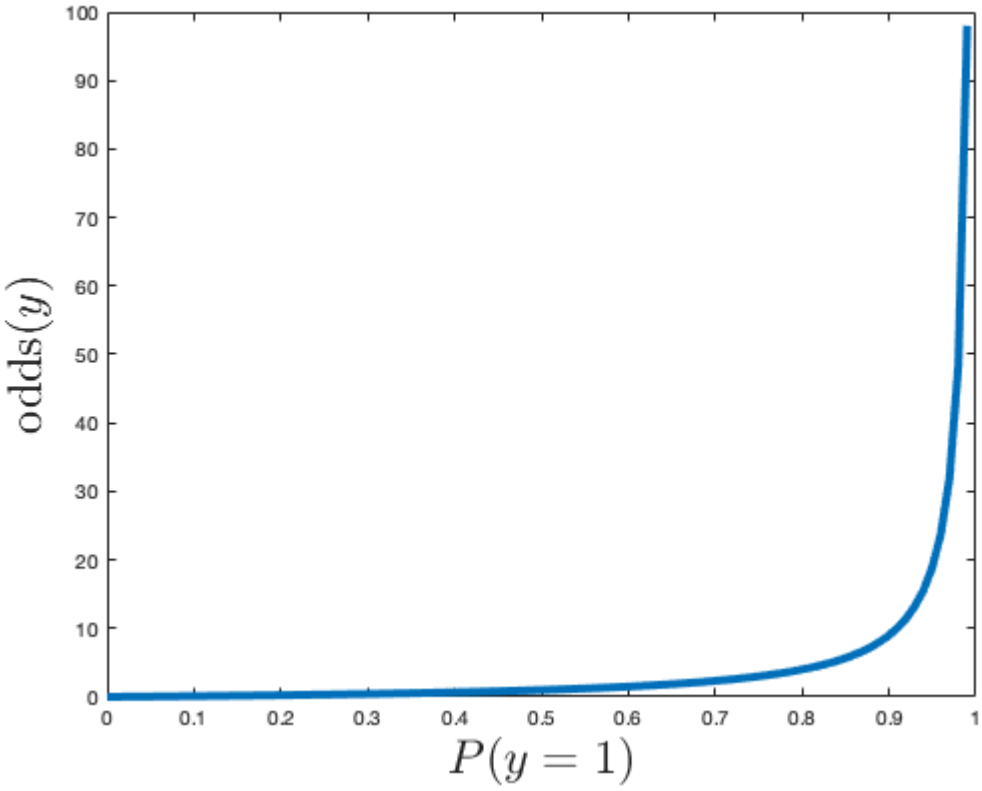
$$\text{odds}(y) = \frac{P(y=1)}{P(y=0)}$$

If the $\text{odds}(y) = 2$, then the probability that y equals 1 is twice as large as the probability that y equals 0. (Odds are usually expressed as a proportion, so an odds of 2 is written as 2:1, or "two to one".) We can convert between probabilities and odds by remembering that probabilities sum to 1, or $P(y=0) + P(y=1) = 1$. Then

$$\text{odds}(y) = \frac{P(y=1)}{P(y=0)} = \frac{P(y=1)}{1 - P(y=1)} \Rightarrow P(y=1) = \frac{\text{odds}(y)}{1 + \text{odds}(y)}$$

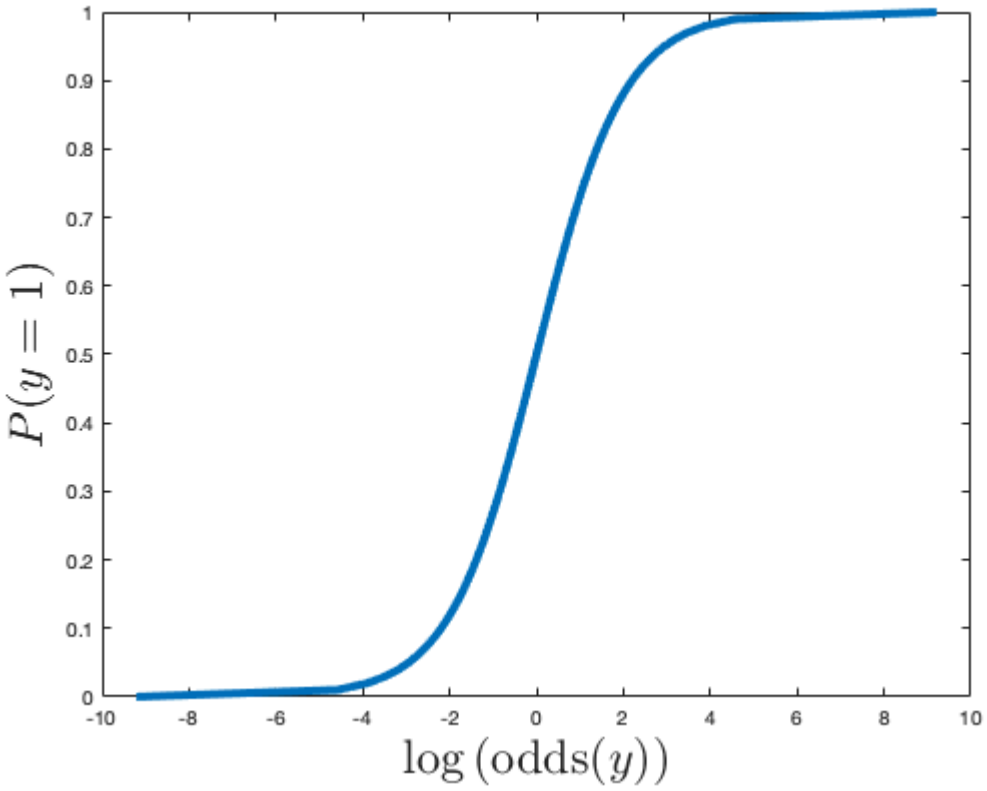
The odds of a discrete variable are always nonnegative and tend toward infinity as $P(y=0)$ approaches 0. The odds function is therefore unbounded, but only in the positive direction.

```
axargs = {'Interpreter','latex','FontSize',24};
pargs = {'LineWidth',4};
P = linspace(0,1);
plot(P,P./(1-P),pargs{:})
xlabel('$$P(y=1)$$',axargs{:})
ylabel('$$\mathrm{odds}(y)$$',axargs{:})
```



The natural logarithm of the odds, however, is a continuous variable. As shown below, the "log odds" is a sigmoid function with horizontal asymptotes at 0 and 1.

```
P = linspace(1e-4,1-1e-4,100);
plot(log(P./(1-P)),P,pargs{:})
ylabel('$$P(y=1)$$',axargs{:})
xlabel('$$\log\left(\mathrm{odds}(y)\right)$$',axargs{:})
```



This is exactly the type of function we're looking for. We can interpret the output of a linear model as the log of the odds of the response variable. Our linear model becomes

$$\log(\text{odds}(y)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Looking at the log odds function above, we see that it transitions from 0 to 1 when $\log(\text{odds}(y)) = 0$. Thus if our continuous linear model predicts a negative value, we say the binary response variable is 0. If the linear model predicts a positive value, we say the binary response variable is 1. (If the linear model predicts exactly zero, we are equally sure that the response is 0 and 1.) Using a linear model to predict the log-odds of a discrete variable is called *logistic regression*. The function $L(y) = \log(\text{odds}(y))$ is called the *logit* function. Because it links the response variable to the linear models, we refer to the logit (and other similar functions) as *link functions*.

Example: Predicting risk of Huntington's Disease

Huntington's Disease is an inherited genetic condition caused by repeated CAG sequences in the Huntingtin (*HTT*) gene. Too many CAG repeats create a "glutamine knot" in the protein, causing toxic protein aggregates in neurons. Symptoms of Huntington's appear later life, and an individual's risk for developing the disease correlates with the number of CAG repeats.

Huntingtin (*HTT*)

Leu Lys Ser Phe Gln Gln ... Gln Gln Gln Gln Pro

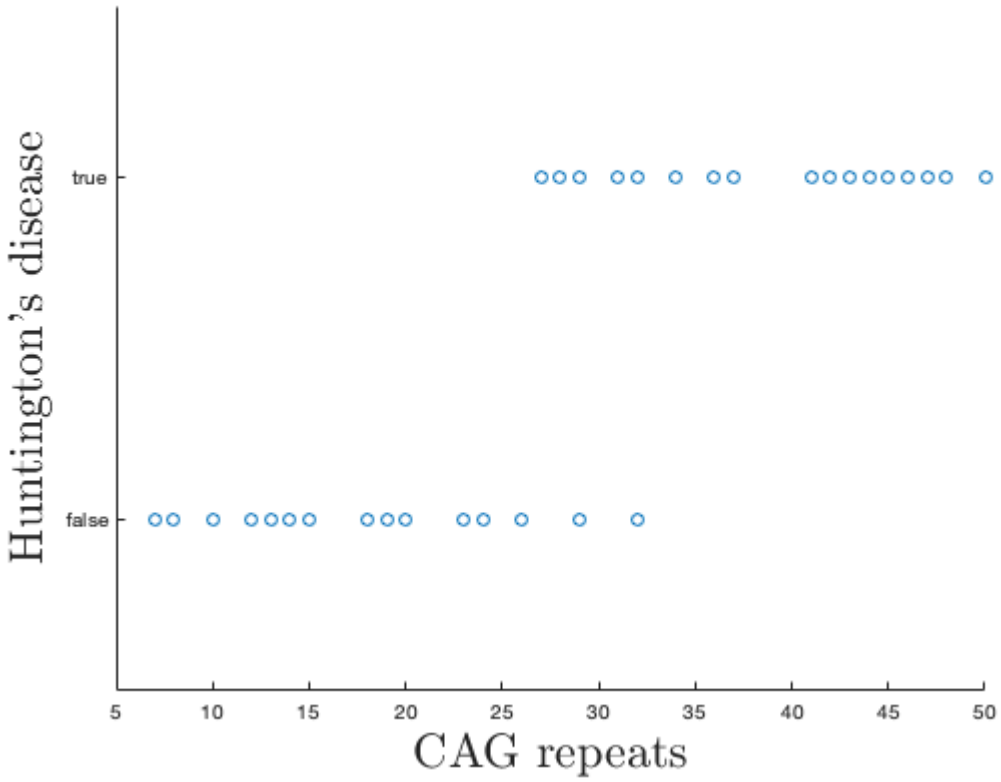
ctc aag tcc ttc cag cag ... cag cag caa cag ccg

# of CAG Repeats	Disease Outcome
< 28	Not affected.
28-35	Increased risk.
36-40	Affected; some offspring affected.
> 40	Affected; all offspring affected.

Source: Walker FO. Huntington's disease. *The Lancet*. 2007: **369**, (9557), 218–228

Let's build a model to predict the probability of developing Huntington's based on the number of CAG repeats. The response variable is binary (Huntington's disease or not) and the predictor variable is continuous (the number of CAG repeats in the *HTT* gene). To train the model, we counted the number of CAG repeats in 50 individuals with and without the disease.

```
load huntington.mat
scatter(hunt.CAGs,categorical(hunt.disease))
xlabel('CAG repeats',axargs{:})
ylabel("Huntington's disease",axargs{:})
```



We see from these data that predicting disease status with low (<25) or high (>35) CAG repeats is straightforward. However, there is a region between 25 and 35 CAG repeats where disease status can be ambiguous. Let's build a logistic regression model to predict Huntington's status. We use the Matlab function `fitglm`, for "fit generalized linear model". The `fitglm` function is similar to `fitlm`; the first argument is a table of data, and the second argument is a formula describing the model. However, `fitglm` can use a wide range of link functions and datatypes when fitting linear models. For logistic regression using binary responses we need to specify the logit link function and a binomial distribution.

```
model = fitglm(hunt,'disease ~ CAGs','link','logit','Distribution','binomial')
```

```
model =
Generalized linear regression model:
    logit(disease) ~ 1 + CAGs
    Distribution = Binomial

Estimated Coefficients:
                   Estimate      SE        tStat      pValue
                   ________    ________    _______    _________

    (Intercept)     -14.032      5.7832    -2.4263    0.015252
    CAGs            0.50558     0.20395     2.4789    0.013179


50 observations, 48 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 55, p-value = 1.18e-13
```

Remember that the model we're fitting is

$$\log\left(\text{odds}(y)\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

so the parameters are not directly interpretable as probabilities. We can rearrange the model to find a formula for the odds of the response:

$$\text{odds}(y) = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}$$

Since the odds of y is equal to $P(y=1)/(1-P(y=1))$, we can also solve for the probability that $y=1$.

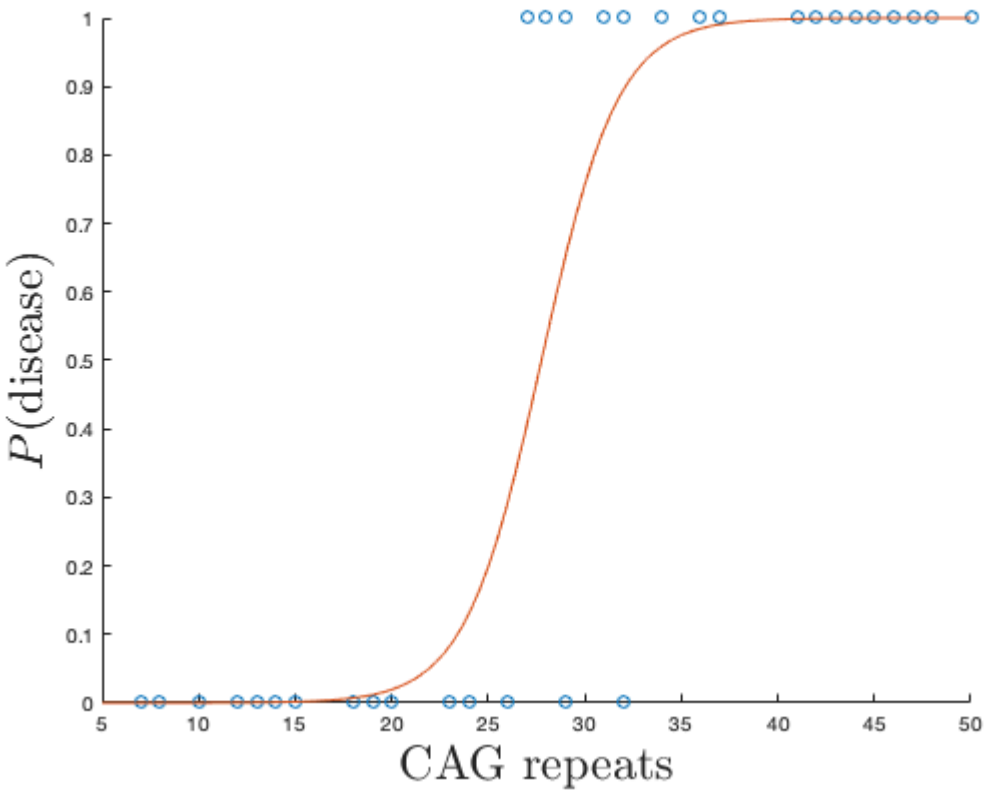
$$P(y=1) = \frac{1}{1+e^{-t}}, \quad t = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

For our Huntington's example

$$P(\text{disease}) = \frac{1}{1+e^{-(\beta_0+\beta_1[\text{CAGs}])}} = \frac{1}{1+e^{14.03-0.51[\text{CAGs}]}}$$

```
scatter(hunt.CAGs,hunt.disease)
hold on
cag_range = linspace(5,50,100);
beta = model.Coefficients.Estimate;
plot(cag_range, 1./(1+exp(-(beta(1)+beta(2)*cag_range))))
hold off
xlabel('CAG repeats',axargs{:});
```

```
ylabel('$$P(\mathrm{disease})$$',axargs{:});
```



We are often interested in when $P(\text{disease}) = 1/2$, as this is the threshold number of CAG repeats where a person is equally likely to have or not have Huntington's. We could solve the above equation, or we could recall that the logistic function reaches its midpoint when the linear model moves from negative to positive. Thus we can simply solve for when $\beta_0 + \beta_1[\text{CAGs}] = 0$.

$$-14.03 + 0.51[\text{CAGs}] = 0 \Rightarrow [\text{CAGs}] = 14.03/0.51 \approx 28 \text{ CAG repeats}$$

Interpreting Individual Coefficients

The entire linear model ($\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$) is useful for predicting the probability of the reponse variable. To interpret individual coefficients we can calculate the *odds ratio*, or the change in odds per unit change in a single predictor variable. For our Huntington's example, let's calculate the odds ratio for adding one more CAG repeat.

$$\text{odds ratio([CAGs])} = \frac{\text{odds}([\text{CAGs}]+1)}{\text{odds}([\text{CAGs}])} = \frac{e^{\beta_0+\beta_1([\text{CAGs}]+1)}}{e^{\beta_0+\beta_1[\text{CAGs}]}} = \frac{e^{\beta_0}e^{\beta_1[\text{CAGs}]}e^{\beta_1}}{e^{\beta_0}e^{\beta_1[\text{CAGs}]}} = e^{\beta_1}$$

Since $\beta_1 = 0.51$, having one more CAG repeat increases the odds of developing Huntington's by $e^{0.51} = 1.67$ fold. For any logistic regression model, the odds ratio for the ith predictor variable is the exponential of the ith coefficient:

$$\text{odds ratio}(x_i) = \frac{\text{odds}(x_i+1)}{\text{odds}(x_i)} = e^{\beta_i}$$