DATA 581

Modeling and Simulation II

Lecture 4: Generalized Linear Model

What We Discuss Today

- **Linear Models**
- **Generalized Linear Model**
- **Examples**

Linear Models

- **A linear model is a predictive model where the expected value of the response variable can be expressed as a linear combination of predictor variables.**

- **Usually, linear models can be fit using least-squares methods, where a linear system of equations must be solved in order to find the parameter estimates.**

- **Multiple Regression models**

$$E[y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- **There are some conditions that need to be met in order to least squares be BLUE (Best Unbiased Linear Estimator)!**
  - **normal error**
  - **constant error variance**
- **Another assumption: response variable is continious/numeric.**

  In many real life data science applications, these assumptions are not valid.

- How about when y is a count data?
  - number of publications by a professor in a career
- How about when y is binary?
  - numerous studies in social science
    - Turnout - vote (1) /abstain (0).
    - Conflict - war (1)/ no war(1)
- what to use when error is not normal?
  - For count data and many other kinds of data, normality is not realistic.
- Hmmm! We want a more flexible model

How Generalized Linear Models Differ from Linear Models

Generalized Linear Models go beyond this in two major respects:

- **The response variable(s) can have a distribution other than normal — as long as it belongs to "exponential family of distributions".**
- **The relationship between the response and explanatory variables need not be simple! (Identity)**

A generalized linear model is a predictive model where the expected value of the response variable is a *function* of a linear combination of predictor variables.

$$g(Y) = \alpha + \beta x \rightarrow Y = g^{-1}(\alpha + \beta x)$$

Generalized linear models are usually fit using maximum likelihood estimation or a related method.

We can conisder three elements for Generalized Linear Models (GLM);

- **Random component : Distribution of response variables (Y)**
- **Systematic component : A linear predictors $\mathbf{X}\beta$**
- **Link : Specify a function that can describe the relationship between the expected value of the random component and the systematic component.**

Some notes on the Link function

The link function relates the linear predictor to some parameter of the distribution for Y (usually the mean; expected value).

$$g(\theta) = \mathbf{X}\beta$$

$$= g^{-1}(\mathbf{X}\beta)$$

where $E[y] = \theta$ and $g(.)$ is the link function.

Note that we usually use the inverse link function $g^{-1}(\mathbf{X}\beta)$ rather than the link function.

Some notes on the Link function

Some examples of link functions:

- **Indentity**
  - **Simple linear regression :** $g(Y) = I(Y) = \mathbf{X}\beta$
  - **Inverse Link : Identity** $Y = I(\mathbf{X}\beta) = \mathbf{X}\beta$
- **Inverse**
  - $g(Y) = Y^{-1} = \mathbf{X}\beta$
  - **Inverse Link :** $Y = \mathbf{X}\beta^{-1}$
- **Log**
  - **Poisson regression** $g(Y) = Ln(Y) = \mathbf{X}\beta$
  - **Inverse Link :** $Y = \exp(\mathbf{X}\beta)$
- **Logit**
  - **Logistic regression:** $g(Y) = Ln(\frac{y}{1-y}) = \mathbf{X}\beta$
  - **Inverse Link :** $Y = \frac{1}{1+\exp(-\mathbf{X}\beta)}$

Example: Wind Speed Modeling

Observations of noon hour wind speed taken at the Winnipeg International Airport are listed in the second column of the *MPV* object `windWin80`.

If we transform data by squaring these observations we can have a dataset for which the exponential distribution is a rough approximation.

```
library(MPV) # contains windWin80
windWin80sq <- windWin80$h12^2
```

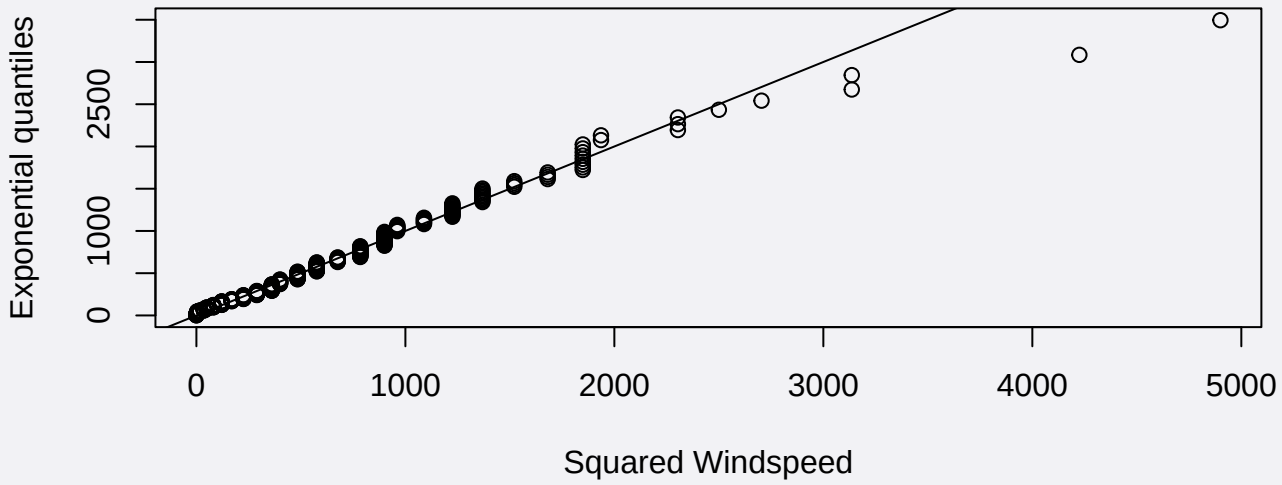
Exercise: Show that the MLE of the parameter for an exponential distribution is the reciprocal of the sample average.

```
lambdahat <- 1/mean(windWin80sq)
```

The estimated value of λ is 0.0016898.

Exponential QQ- plot of Squared Windspeeds

We can verify that this is a reasonable model with an exponential QQ-plot:



Exponential QQ-plot of squared Winnipeg wind speed observations for 1980.

Question? Do you think the wind speed rate parameter is always same?

In other word, do you think the wind rate parameter can be a function of the parameters i.e. the season, temperature, weather condition?

We will see that the usefulness of generalized linear models comes in large part from their ability to relate response variables which follow particular probability distributions to one or more predictor variables, through exploitation of the likelihood framework.

Predicting Noon Windspeed the Previous Midnight

The `windWin80` contains another column, called `h0`, containing the wind speed measurements at midnight, 12 hours before the noon wind speed measurements that we have been studying.

Can we predict the future noon wind speed from the previous midnight windspeed?

Identify three components for GLM

1. The transformed noon wind speed is exponentially distributed (stochastic component)
2. we build a linear combination of covariates (systematic component); the previous midnight's wind speed
3. a proper choice for link function for the exponential distribution is the inverse function

$$\lambda = \frac{1}{\beta_0 + \beta_1 x}$$

where x represents the previous midnight's wind speed.

Poisson Regression

The `cigbutts` data set (in the *MPV* package) gives counts of cigarette butts at locations along a sidewalk as a function of distance from a smoking gazebo.

We can use Poisson regression to model this count data as a function of distance.

Poisson Distribution

A popular choice of link function for the Poisson distribution model is the *log link*. A covariate x would be included in such a model through

$$\log(\lambda) = \beta_0 + \beta_1 x. \tag{1}$$

This link function is modelled in such a way that recognizes that λ must be positive, while the right hand side of the above equation can take any real value.

Likelihood for the cigarette butts data

If X_i is the number of butts observed at the ith location, its expected value would be λ_i and the corresponding distance might be denoted as d_i.

The model for the ith observation is really

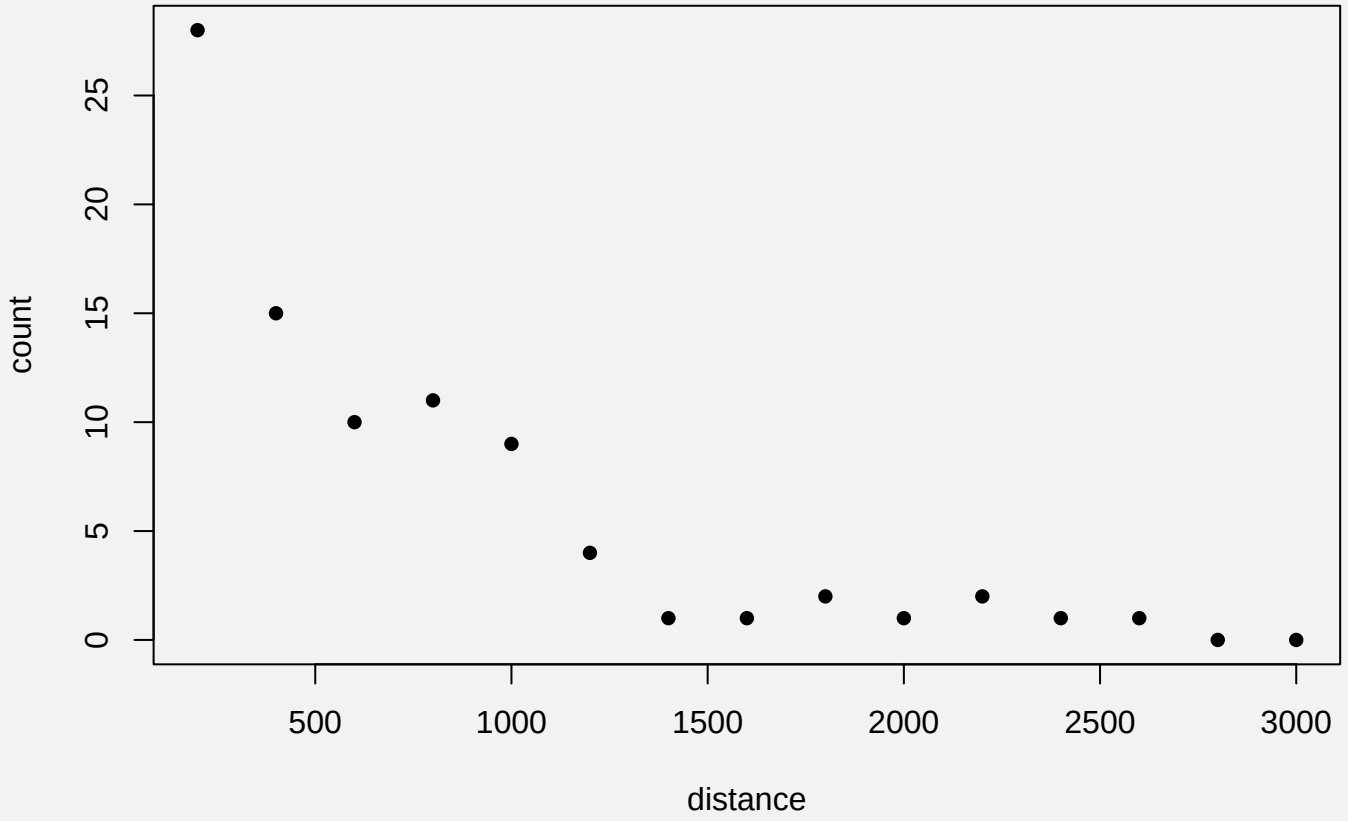
$$P(X_i = x_i) = \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}.$$

The likelihood function is the product of such probabilities

The log likelihood is then the sum of the logs of the probabilities:

$$\log L = -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} x_i \log \lambda_i - \sum_{i=1}^{n} \log x_i!.$$

Visualizing the Cigarette Butt Counts



Scatterplot of log of cigarette butt counts versus log of distance.

Predicting Counts Using Distance

We suspect a linear relation between the log of λ_i and d_i:

$$\log(\lambda_i) = \beta_0 + \beta_1 d_i$$

Let's suppose we know that $\beta_0 = 3.55$. Then we could say that

$$\log(\lambda_i) = 3.55 + \beta_1 d_i$$

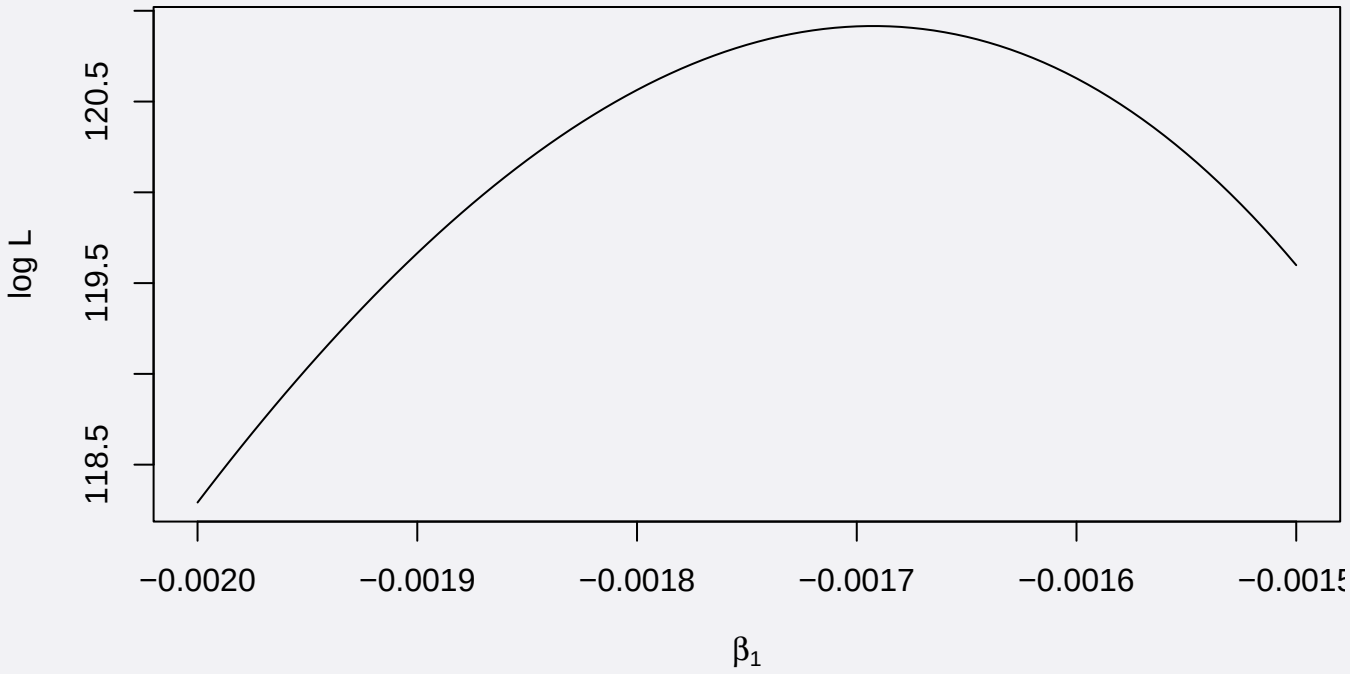
Plugging this into the log likelihood expression gives

$$\log L = -\sum_{i=1}^{n} e^{3.55+\beta_1 d_i} + \sum_{i=1}^{n} x_i(3.55 + \beta_1 d_i) - \sum_{i=1}^{n} \log x_i!.$$

The x_i's are the observed counts and the d_i's are the distances, so we can plot this as a function of β.

Predicting Counts Using Distance

Log likelihood function:



Log likelihood curve for the cigarette butts example. The maximizer is near -.0017.

A predictive model for log of the expected number of cigarette butts would be

$$\widehat{\log(\lambda)} = 3.55 - .0017d$$

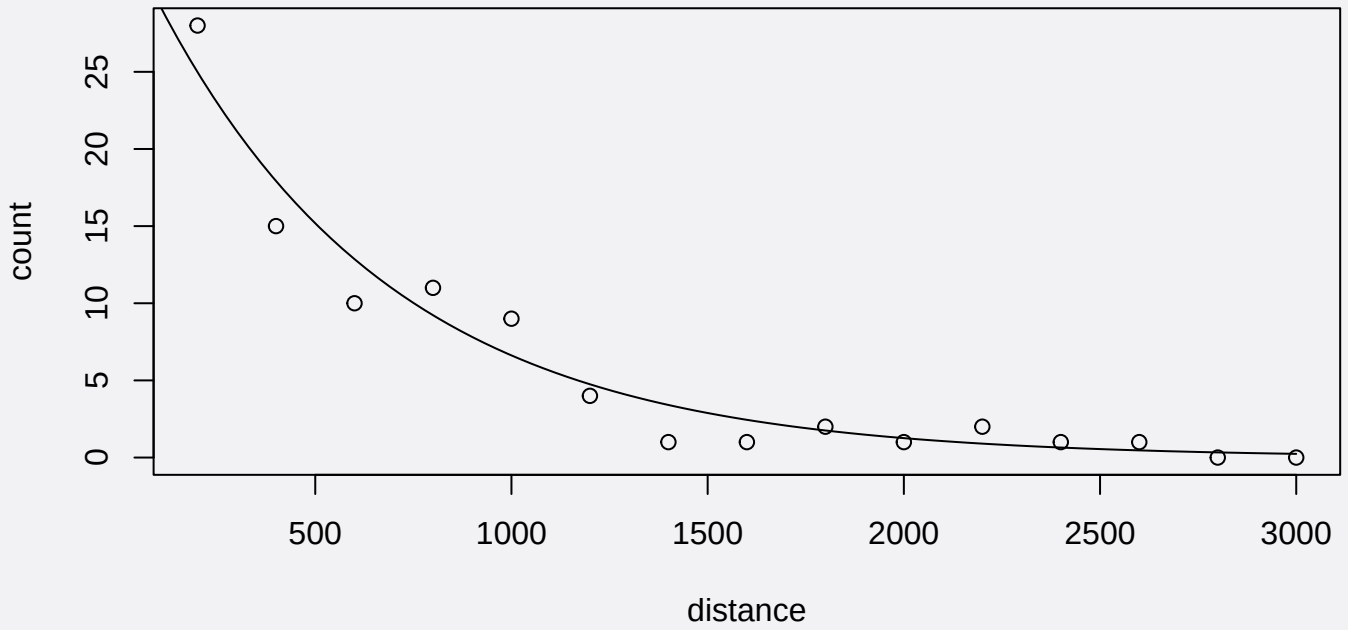
where d is distance from the gazebo.

Using Built-In Software

The `glm()` function fits Poisson regressions and logistic regressions fairly straightforwardly.

```
cig.glm <- glm(x ~ d, family = poisson)
coef(cig.glm)

##   (Intercept)             d
##  3.553514259 -0.001695636
```



Additional output:

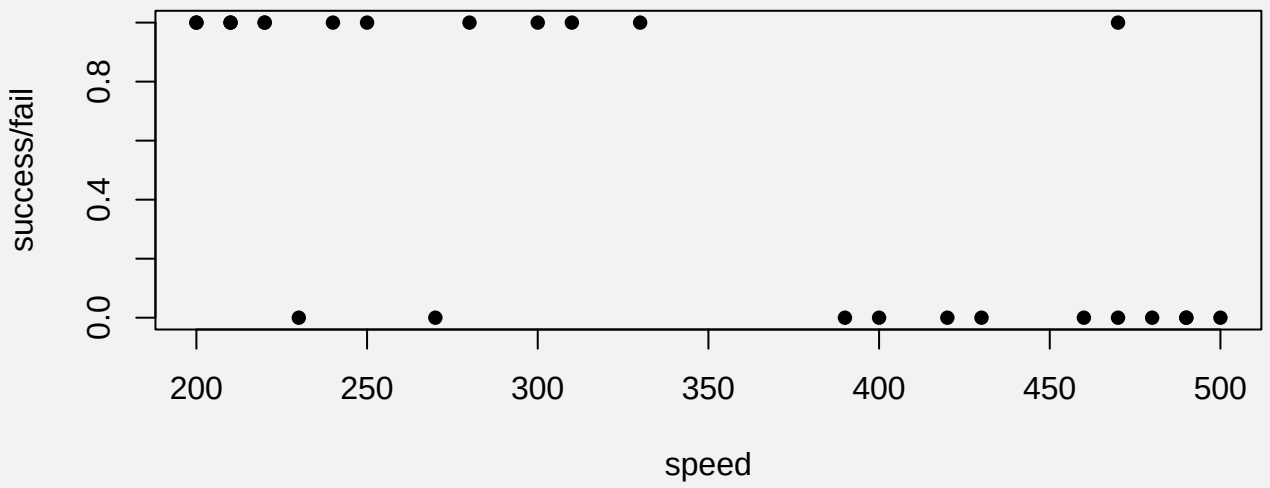
```
summary(cig.glm)$coefficients

##                 Estimate   Std. Error  z value
## (Intercept)  3.553514259 0.1735102026 20.48015
## d           -0.001695636 0.0002008638 -8.44172
##                 Pr(>|z|)
## (Intercept) 3.236877e-93
## d           3.127010e-17
```

Note, in particular, the standard errors of the parameter estimates.

The data in `p13.1` in the *MPV* package describes successes and failures of surface-to-air missiles as they relate to target speed.

```
library(MPV)
plot(p13.1, xlab = "speed", ylab = "success/fail", pch=16)
```



Surface-to-air missile successes (1) and failures (0) as they relate to target speed (in knots).

- **Fitting a straight line to such data makes no sense, since the plotted points do not at all scatter about such a line.**
- **if such a line were to be fit to the data, it would necessarily take values outside the interval $[0, 1]$ on subsets of the domain; interpretation of such values would be difficult.**
- **preferred interpretation of output arising from the fitting of models to such data is that of probability.**
- **useful models can provide answers to questions such as, "What is the probability of success at a given target speed?"**
- **Since probabilities must lie within the interval $[0, 1]$, we must consider models based on nonlinear functions.**

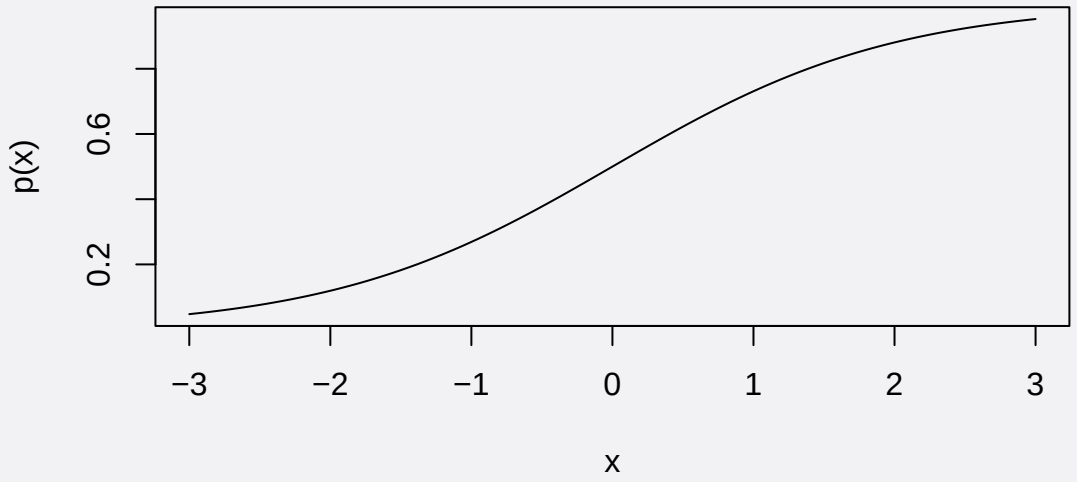
There are many functions which have values in $[0, 1]$.

For the current example, we might reasonably believe that the probability of success decreases as target speed increases.

Perhaps the most popular function for this purpose is the *logistic* function

$$p(x) = \frac{e^x}{e^x + 1}.$$

```
curve(exp(x)/(1 + exp(x)), from = -3, to = 3, ylab="p(x)")
```



The logistic function.

A bit of algebra allows us to express x in terms of p, yielding the *logit* function:

$$\ell(p) = \log\left(\frac{p}{1-p}\right).$$

While p is restricted to take values between 0 and 1, the logit function can take any possible value, so relating the logit function to a straight line or other linear combination is a possibility. For example,

$$\ell(p(x)) = \beta_0 + \beta_1 x$$

which means that we can express the probability of an event in terms of a covariate x, using a linear function, but the probability is related to the linear function through the logit.

The logit is an example of a *link function*, since it links the expected response, in this case the probability $p(x)$ to the linear function of the covariate(s).

To fit the logistic regression model to the missile success data, try

```
p13.glm <- glm(y ~ x, data = p13.1, family = binomial)
```

Note that we did not specify the link function; the default choice with the binomial family is the logit.

```
coef(p13.glm)

## (Intercept)           x
##   6.0708839  -0.0177047
```

The Coefficient part of the output tells us that the logit of the probability of success as a linear function of target speed has intercept 6.07 and slope -.0177.

More output:

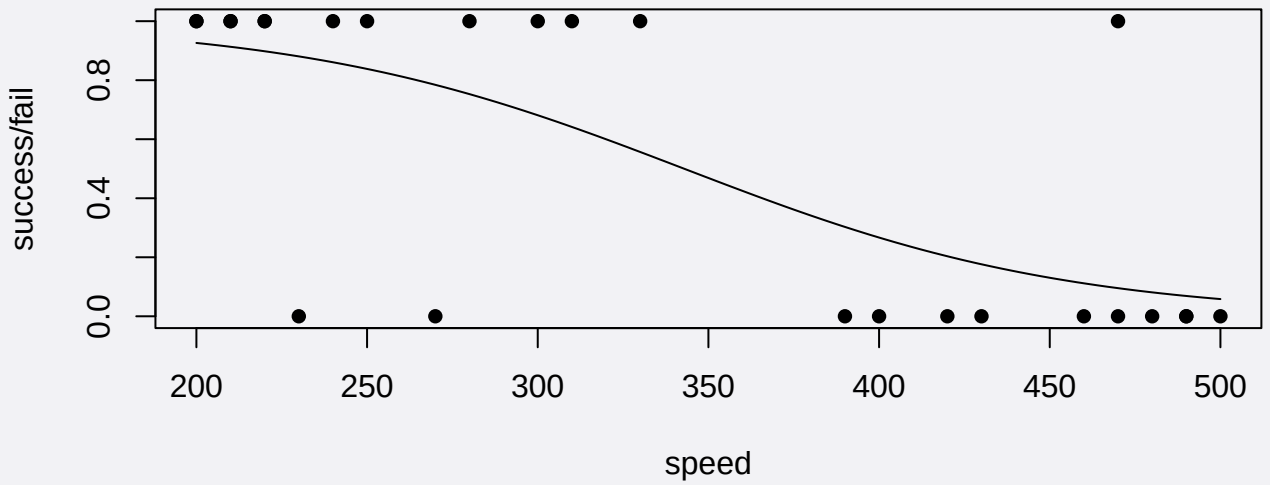
```
summary(p13.glm)$coefficients

##                Estimate  Std. Error   z value
## (Intercept)  6.0708839 2.108996265  2.878566
## x           -0.0177047 0.006075513 -2.914107
##                Pr(>|z|)
## (Intercept) 0.003994883
## x           0.003567073
```

Standard error estimates for these parameter estimates are supplied and indicate, in particular, that the slope is clearly negative.

Modelling Binary Responses - Visualizing the Model

```
plot(p13.1, xlab = "speed", ylab = "success/fail", pch=16)
newspeeds <- 200:500 # predict at these target speeds
lines(newspeeds, predict(p13.glm,
  newdata=data.frame(x = newspeeds),  type = "response"))
```



Surface-to-air missile successes (1) and failures (0) as they relate to target speed (in knots) with overlaid logistic curve.