Welcome to Week #12!

We have covered in an Intro to Machine Learning:

Linear Regression
Multiple Linear Regression

MLR: For multiple linear parameters

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

predicted y

intercept

"slopes along different parameters"

one response

many explanatory variables

Visualization in 3D for 2 explanatory variables: http://miabellaai.net/

SLR to MLR: R^2 to R_{adj}^2

The strength of the fit of a linear model is most commonly evaluated using R^2.

R^2 is calculated as the square of the correlation coefficient.

It tells us what percent of variability in the response variable is explained by the model.

n = # of data points

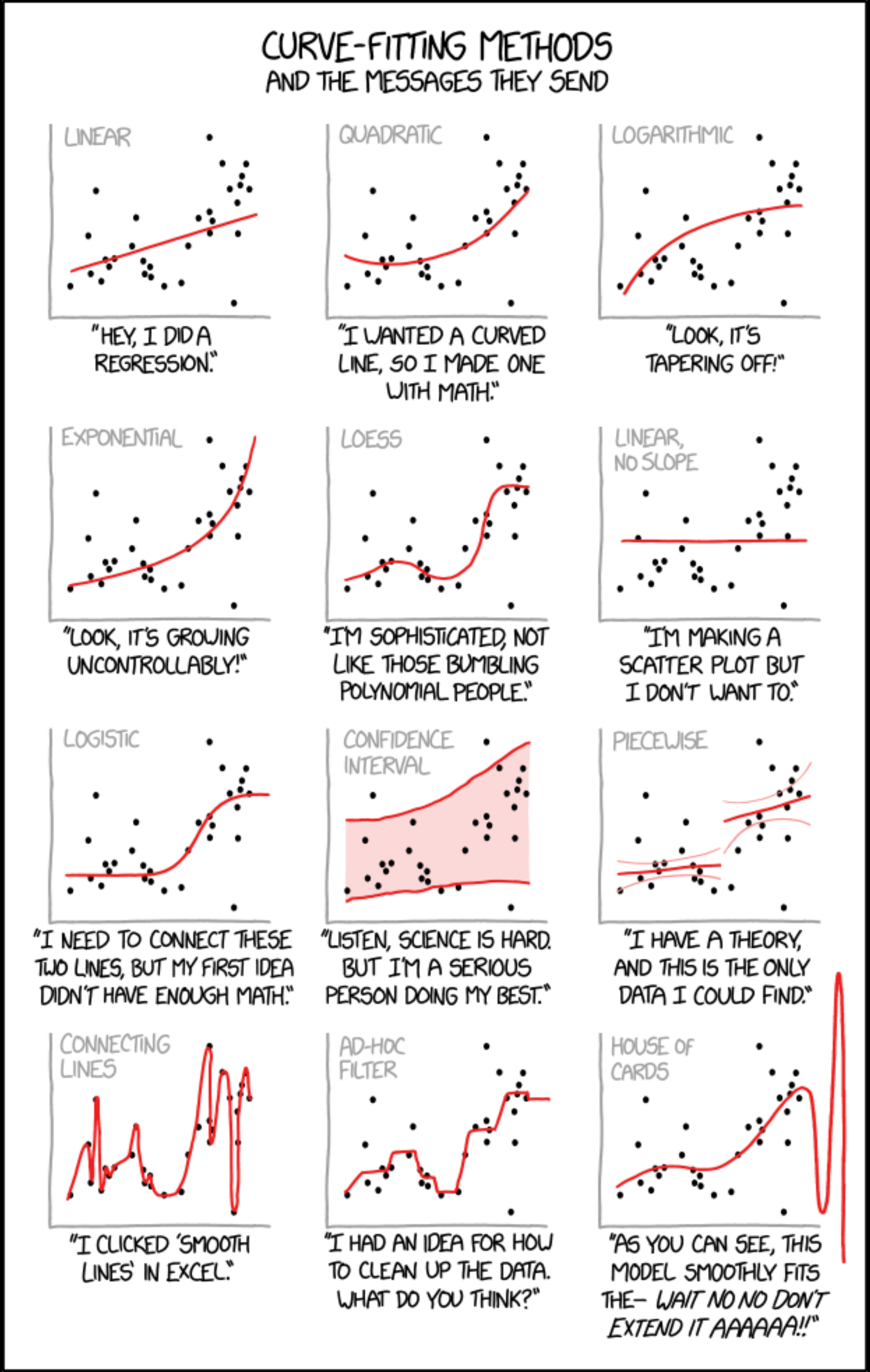
$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

k = # of parameters in our model

$$R_{adj}^2 = 1 - \frac{Var(e_i)/(n-k-1)}{Var(y_i)/(n-1)} = 1 - \frac{Var(e_i)}{Var(y_i)} \times \frac{n-1}{n-k-1}$$

So like R^2, but takes into account the number of degrees of freedom

More parameters (higher k) means worse R_{adj}^2 - adjusts for the fact that we can fit anything if we have a large enough number of parameters!



And in general: we should avoid making things overly complicated

xkcd.com

Intercepts in MLR don't usually make a lot of sense

```
Call:
lm(formula = Snow.Depth ~ Elevation + Min.Temp + Max.Temp, data = snotel2)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.508  -7.679  -3.139   9.627  26.394 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -10.506529  99.616286  -0.105   0.9170  
Elevation     0.012332   0.006536   1.887   0.0731 .
Min.Temp     -0.504970   2.042614  -0.247   0.8071  
Max.Temp     -0.561892   0.673219  -0.835   0.4133  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.6 on 21 degrees of freedom
Multiple R-squared:  0.6485,	Adjusted R-squared:  0.5983 
F-statistic: 12.91 on 3 and 21 DF,  p-value: 5.328e-05
```

"at 0 feet elevation, 0F minimum and maximum daily temperature the snowfall should be -10.5 inches"

Few notes on Collinearity & Model Selection

- Two predictor variables are said to be collinear when they are correlated, and this collinearity complicates model estimation.
  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. parsimonious model.
  It is OK to consider multiple reasons to select a model but it is dangerous to "shop" for a model across many possible models – a practice which is sometimes called ***data-dredging*** and leads to a high chance of spurious results from a single model that is usually reported based on this type of exploration.

- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

vif(our model) = $$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

Some “non-linear” linear models

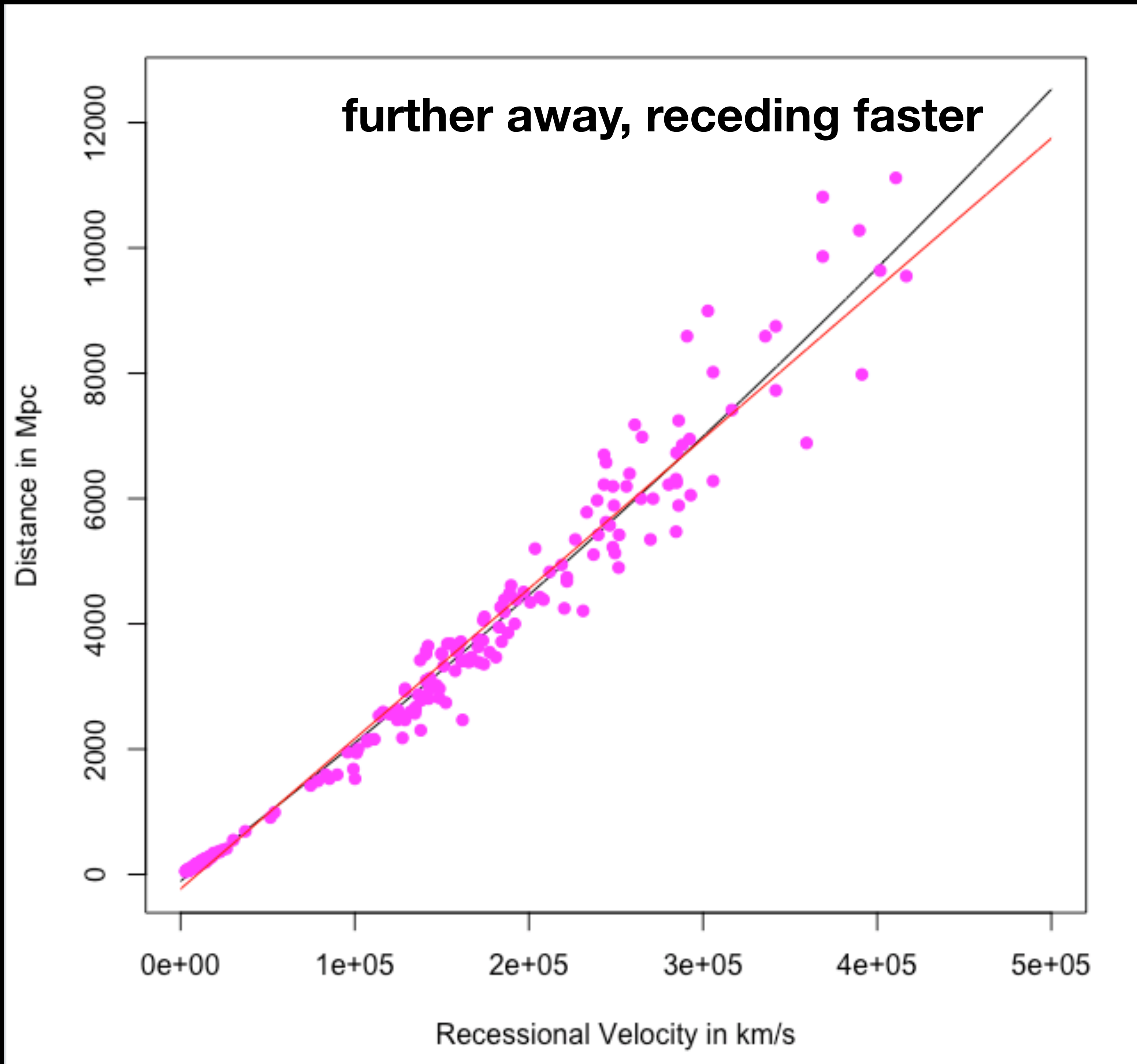
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n$$

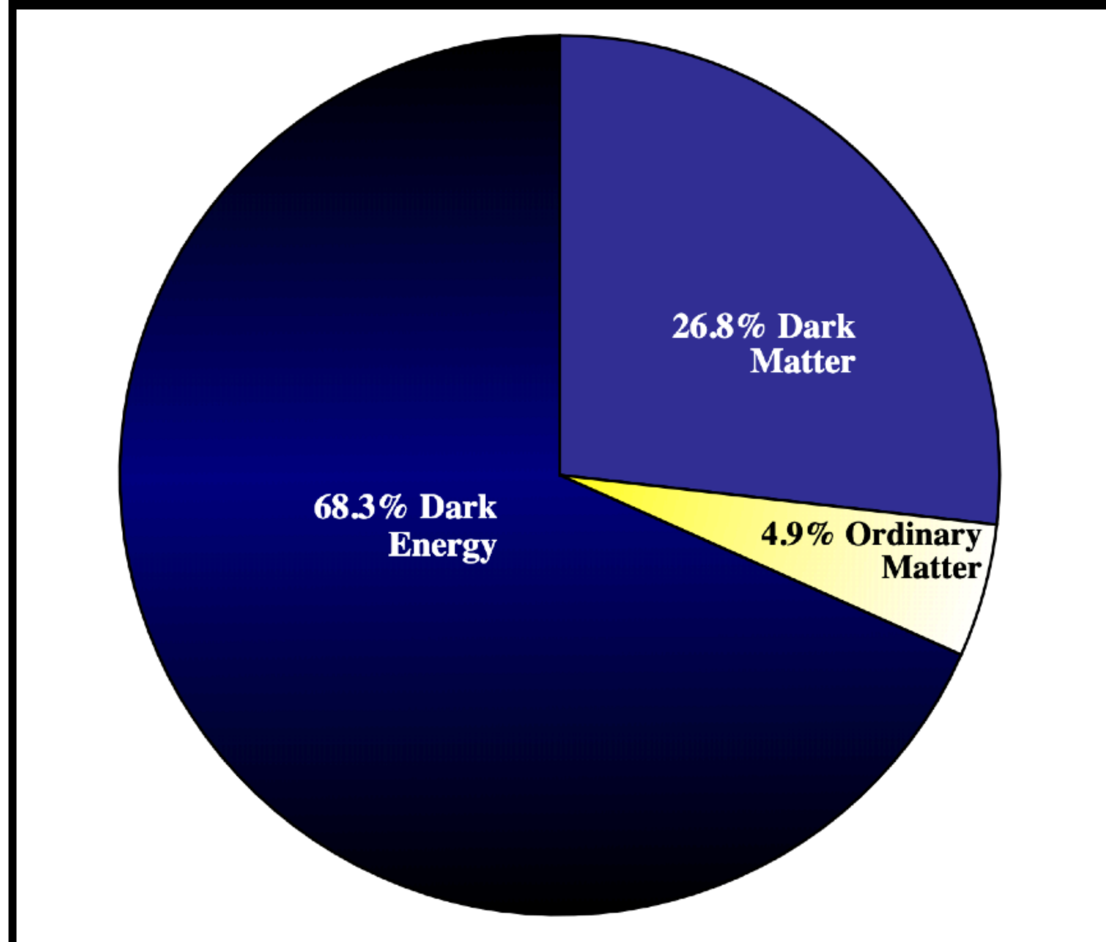
predicted y

intercept

“slopes along different parameters”

what can x_n be here? For example, I can say $x_4 = x_3^2$?





http://adamdempsey90.github.io/python/dark_energy/dark_energy.html

Issues with p-values

https://www.amstat.org//asa/files/pdfs/P-ValueStatement.pdf

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.* **SCIENCE!**

Logistic Regression:
Getting Numbers from Levels

Logistic Regression is a subset of Classification (more on that later)

Logistic Regression

At this point we have covered:

- Simple linear regression
  - Relationship between numerical response and a numerical or categorical predictor

- Multiple regression
  - Relationship between numerical response and multiple numerical and/or categorical predictors

What we haven’t seen is what to do when the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (categorical, count data, etc.)

Logistic Regression: A Morbid Example

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming.

There its leaders decided to attempt a new and untested rote to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake.

The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)

Let’s look at this data in R!

Logistic Regression: A Morbid Example

It seems clear that both age and gender have an effect on someone’s survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn’t something we can transform our way out of (there is no 0.5 “dead”) - we need something more.

One way to think about the problem - we can treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.

Logistic Regression: A Morbid Example

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable

2. A linear model
   - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$

3. A link function that relates the linear model to the parameter of the outcome distribution
   - $g(p) = \eta$ or $p = g^{-1}(\eta)$

g turns probability into a number, g^{-1} turns linear fit to probability

Logistic Regression: A Morbid Example

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model *p* the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to *p*. There are a variety of options but the most commonly used is the logit function.

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \le p \le 1$$

Logistic Regression: A Morbid Example

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

Logistic Regression: A Morbid Example

Ok, so what does the totality of our model look like?

$$y_i \sim \text{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{logit}(p) = \eta$$

From which we back out the probability of survival based on parameters 1-n, for the *i*th observation:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

Logistic Regression: A Morbid Example

Give me an example!

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

So, for example, the odds of survival of a newborn (age = 0):

Can I get an R example??

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

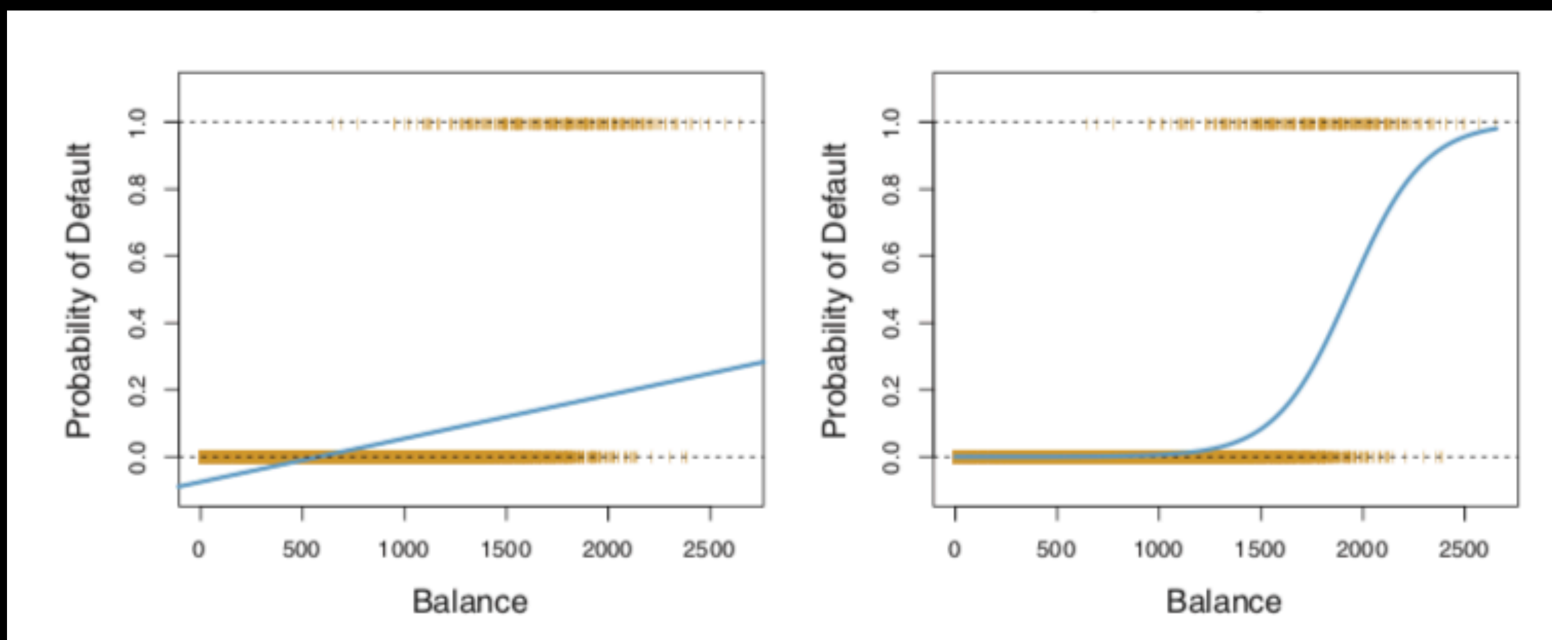
$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

A note about the logit function

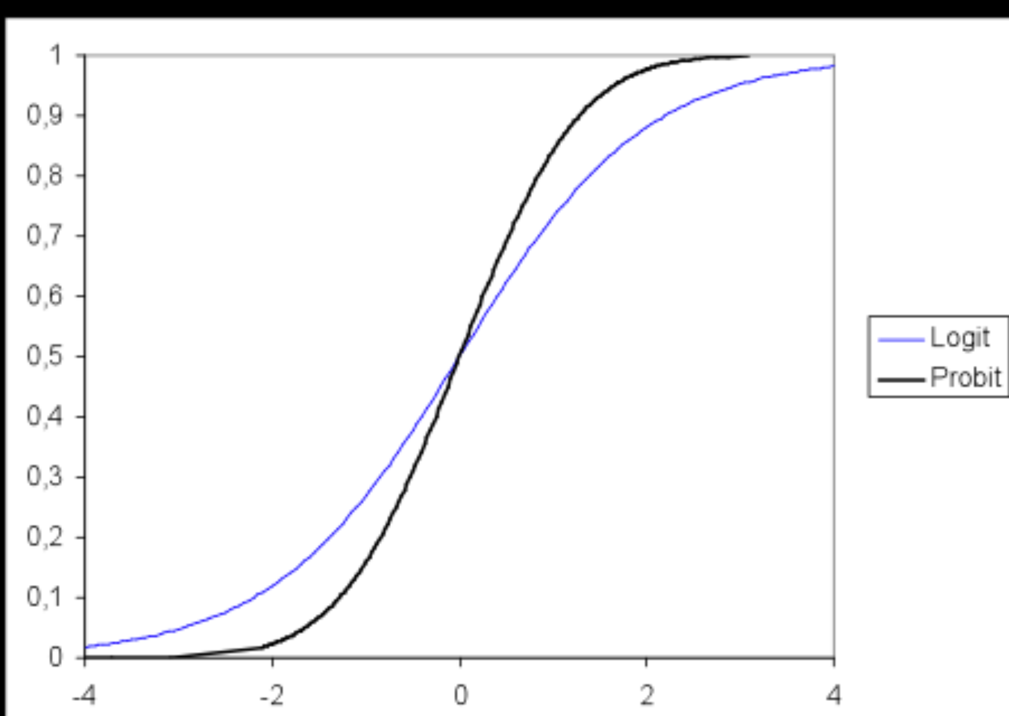
$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Weird mapping between a linear fit & probability of success



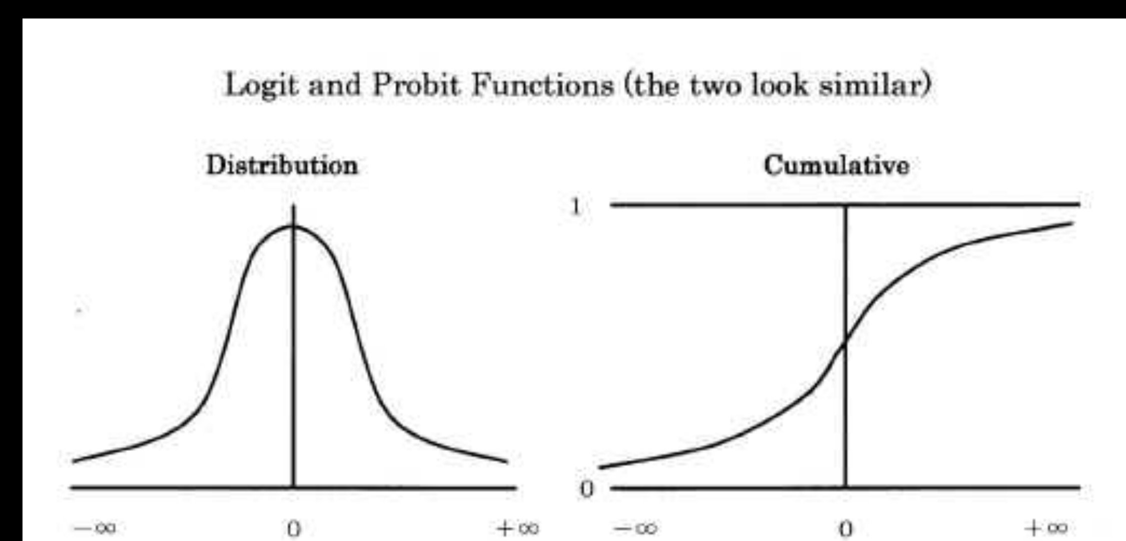
This is essentially to map observed successes and failures to probabilities.

Want to avoid probabilities < 0 or > 1



Other possible mappings however.

Logit the most (currently) popular.



Another note about the logit function

$$logit(p) = \log\left(\boxed{\frac{p}{1-p}}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

“the odds” or “odds”

Think horse racing:
1 in 20 odds means

1/20 = p/(1-p)

$$p = \frac{1/20}{1 + 1/20} = 0.048$$