Regression is a versatile and useful tool that predicts continuous values by fitting a straight line (or a flat plane/hyperplane) that describes the relationship between a predict**ed** variable and one or more predict**or** variables (predictor variables are sometimes called features.)

Logistic regression is a simple extension of the linear regression framework. This family of models is often referred to as Generalized Linear Models (GLMs) because they are extensions of linear models such as regression. The ultimate goal of logistic regression is often *classification* (predicting categories). For example, logistic regression could be used to predict whether or not a cat will develop joint issues late in life, or whether a customer is likely to upgrade their subscription plan.

THE MATH

Despite the differences in output, logistic regression is essentially linear regression with some added algebra. Let's say that we want to predict whether a customer is likely to upgrade their subscription plan to a monthly fashion box. We can collect data and code the people who did upgrade with a 1, and people who did not with a 0.

Age	Income (k)	...	Upgraded?
29	150.0	...	1
23	45.2	...	0
44	124.3	...	1

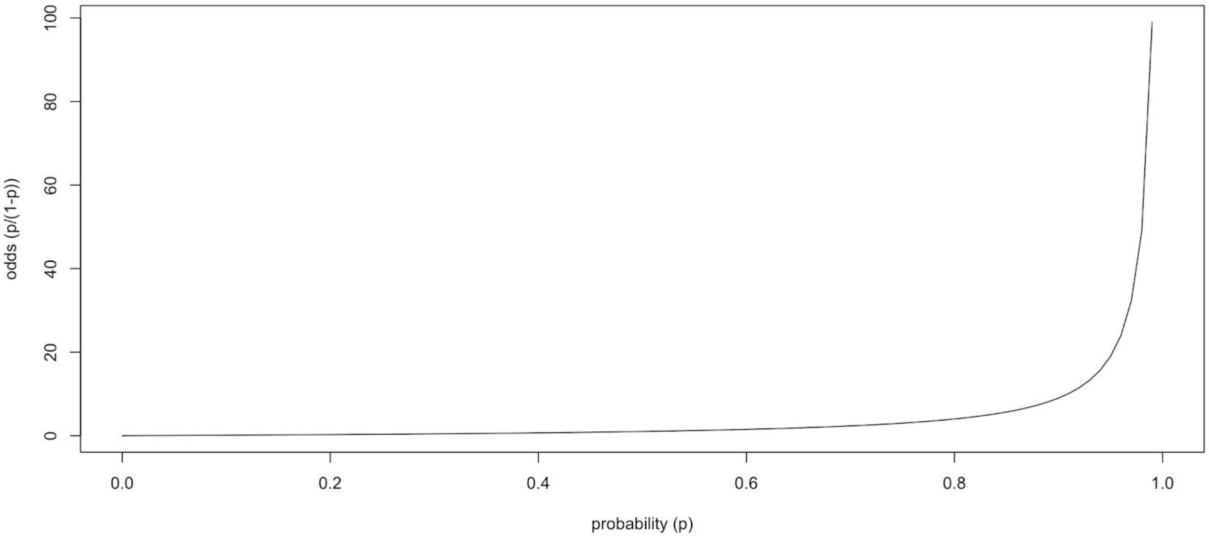
Instead of asking the logistic regression model to strictly predict 0's or 1's, we will ask it to predict something continuous: the probability (0-1) of upgrading. Asking for a continuous value is one step closer to fitting our classification problem into the linear regression framework, but it's only the first step.

Probabilities have a range between 0 and 1. Linear Regression typically predicts values that could theoretically range from -∞ to ∞. To get from a range of 0-1 to a range of -∞ to ∞ we take two steps:

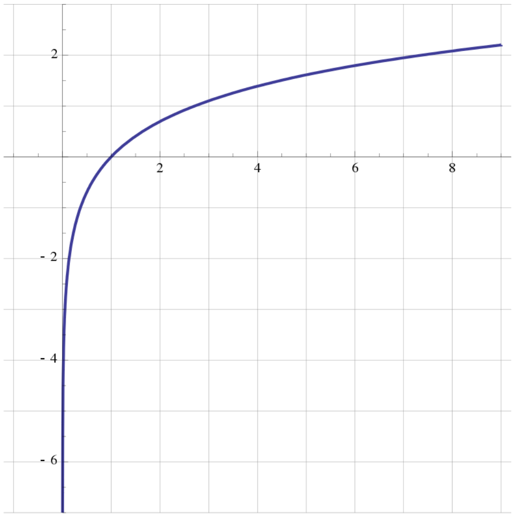
1. Instead of a regular probability, we can transform the probability, *p* into the odds of *p*. Odds are a ratio of the probability of an event occurring, and the probability of it not occurring. If the probability of a soccer player scoring a goal during a match is 0.6, the odds of them scoring a goal is 0.6/(1-0.6) = 1.5. In other words, it is 1.5 times more likely that the player will score a goal than not. The general formula for odds is:

$$\frac{p}{1-p}$$

Odds can range from 0 (when *p* is 0) to Infinity (when *p* is infinitely close to 1), so we are one step closer to the regular -∞ to ∞ range that linear regression typically works with.



2. To turn odds--which have a range from 0 to ∞--into something that has a range from -∞ to ∞, we can take the natural logarithm of the odds to get something called the **log odds**. The natural logarithm takes odds between 0 and 1 and transforms them into negative numbers between -∞ and 0. On the other hand, it takes odds between 1 and ∞ and transforms them into positive numbers between 0 and ∞.



So, we took the probability, turned it into odds, and took the natural logarithm to create log-odds which range from -∞ to ∞.

p	Odds (p/1-p)	Log Odds log((p/1-p))
0.1	0.1111	-2.1972
0.5	1	0
0.9	9	2.1972

Log odds are useful because they help us create the range from -∞ to ∞. The transformation we do-- $log(\frac{p}{1-p})$ --is called the logit function and it is one of many **link functions** that we can use to expand linear regression.

THE MODEL

The general form for linear regression is

$$y = mx + b$$

Where *y* is the predict**ed** (or outcome) variable, and *m* represents one or more predict**or** variables (or features).

Logistic regression expands this formula by applying the logit link function, resulting in this:

$$log(\frac{y}{1-y} = mx + b$$

Instead of directly predicting the probability, we predict the log odds based on one or more predict**or** variables using linear regression. Then, we can apply our logit function to recover our probabilities. Once we have our probabilities, we can either keep them, or we can dichotomize them. If we want categorical predictions (either someone did or did not upgrade their fashion subscription box), we can choose a **threshold** and predict that anyone with a predicted probability above the threshold will upgrade, and anyone below the threshold will not.). 0.5 is a common threshold, but depending on the specifics of your situation, you could choose almost any threshold. For example, in our fashion subscription example, we might want to be conservative in our predictions since we would rather underestimate than overestimate our projected income. In that case, we could use a threshold of 0.7; someone would need a predicted probability of 0.7 or higher before we will predict that they're going to upgrade.

AN EXAMPLE AND COEFFICIENT INTERPRETATION

Logistic regression is often used as a predictive machine learning model, but at its core, it is linear regression and can be interpreted as such. Let's look at more data for the fashion subscription boxes. Here’s a sample of the data (the full set can be found at the bottom of this page):

Upgrade?	Age	Income	Months Subbed
1	22	108.99	35
1	32	58.63	46
1	38	62.89	35
1	14	74.39	19
0	25	72.67	11

Our logistic regression model will look like this:

$$log(\frac{p_{upgrade}}{1-p_{upgrade}}) = Age + Income + MonthsSubbed$$

Or, with a little algebra:

$$p_{upgrade} = \frac{e^{Age+Income+MonthsSubbed}}{1+e^{Age+Income+MonthsSubbed}}$$

The output of a logistic regression model will look similar to the output of a linear regression model, but the interpretation of the coefficients is a little trickier. Statistical Significance can be interpreted in the same way: variables with significant p-values are still considered statistically significant predictors.

Here is the output for our fashion subscription model.

	Estimate	Std. Error	z	p
Intercept	-2.977707	2.780547	-1.071	0.2842
age	0.144545	0.073131	1.977	0.0481
income	-0.006551	0.016506	-0.397	0.6914
months subbed	0.001459	0.016298	0.089	0.9287

Interpreting Raw Coefficients.

In general, the sign of the coefficients will tell you whether something is associated with an increase or decrease in the probability of your outcome (here, fashion subscription upgrade). Here, we see that increased values for age and month's subbed are both associated with an increased probability of upgrading since their coefficients (0.145 and 0.001) are both positive. Income, on the other hand, has a negative coefficient which means that in our sample, increased income is associated with a decreased probability of upgrading.

Technically speaking, the coefficients from the model represent the change in log odds associated with a one-unit increase in our variable. For example, this model tells us that for each increase of 1 year in age, we expect an increase of about 0.145 in our log odds of upgrading. But log odds can feel unintuitive to interpret by themselves.

Interpreting Exponentiated Coefficients (Odds).

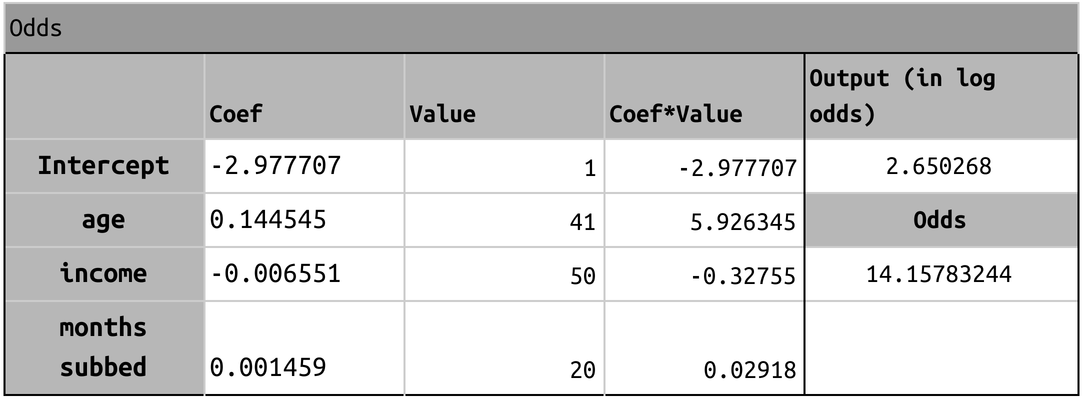
$$e^{coefficient}$$

Often, people will exponentiate the coefficients (i.e. take the constant *e* to the power of the coefficients) so that the results are in terms of odds instead of log-odds. This transformation preserves the linear relationship between your predicted variable (now the odds of upgrading) and the predictors, so the interpretation of exponentiated coefficients is similar to that of the raw coefficients. In our fashion subscription model--with all other variables held constant--an increase of 1 year in age is associated with an exp(0.145) = 1.156 increase in the odds. In other words, we expect a 15.6% increase in the odds for each additional year of age. Exponentiated coefficients larger than 1 represent an *increase* in the odds of your outcome (here, upgrading your subscription), and exponentiated coefficients smaller than 1 represent a *decrease* in the odds of your outcome.

For example, using our model and converting to odds (by exponentiating our prediction), say we take a subscriber who is 40, makes 50K a year, and has been subscribed for 20 months. Our model predicts that their odds of upgrading is 12.25.

Odds				
	Coef	**Value**	**Coef*Value**	**Output (in log odds)**
Intercept	-2.977707	1	-2.977707	2.505723
age	0.144545	40	5.7818	**Odds**
income	-0.006551	50	-0.32755	12.25241426
months subbed	0.001459	20	0.02918	

If we increased their age by 1, keeping all else equal, their odds of upgrading is now 14.16. An increase of 1 year in age increased the odds about 1.16x (or a 16% increase).



Odds				
	Coef	Value	Coef*Value	Output (in log odds)
Intercept	-2.977707	1	-2.977707	2.650268
age	0.144545	41	5.926345	Odds
income	-0.006551	50	-0.32755	14.15783244
months subbed	0.001459	20	0.02918	

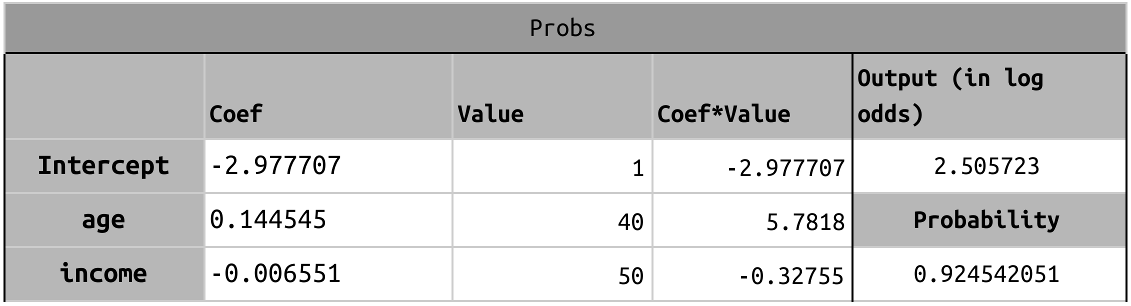
If we look at another person who is 24, makes 50K a year, and has been subscribed for 20 months. And increase their age by 1 (holding all other variables constant), their odds increase from 1.21 to 1.40. Again, a 1-year increase in age is associated with a 1.16x (or 16% increase) change in the odds. Because converting from log odds to odds preserves the linear relationship, the amount is constant. No matter what age you plug in, increasing it by one year (while keeping everything else equal) will increase the predicted odds by 1.16x (or 16%).

Interpreting Probabilities (with caution).

$$\frac{e^{coefficient}}{1+e^{coefficient}}$$

Even though odds can be easier to understand than log-odds, they can still be tough to wrap your head around. Converting odds to probabilities gives you the ability to present your results in a way that can feel more intuitive both to yourself and to others. However, the transformation between odds and probabilities does *not* preserve the linear relationship between your predicted variable (now the probability of upgrading) and the predictors, so proceed with caution. When converting to probabilities, a 1 unit increase in one of your predictors may cause a differently sized increase in the probability of your event (here, upgrading) depending on what value you're starting with.

For example, using our model and converting to probabilities, say we take a subscriber who is 40, makes 50K a year, and has been subscribed for 20 months. Our model predicts that their probability of upgrading is 92.5%



Probs				
	Coef	Value	Coef*Value	Output (in log odds)
Intercept	-2.977707	1	-2.977707	2.505723
age	0.144545	40	5.7818	Probability
income	-0.006551	50	-0.32755	0.924542051

months subbed	0.001459	20	0.02918	

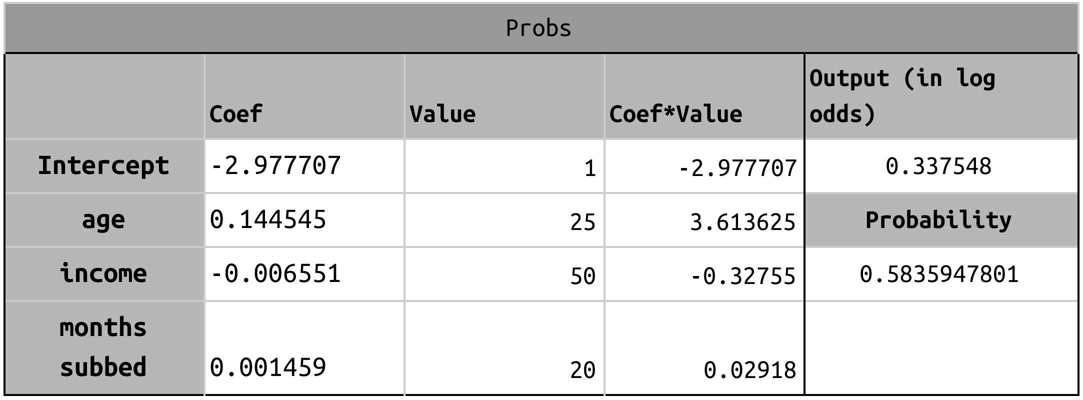
Keeping the other variables constant, an increase in age from 40 to 41 increases the predicted probability from 92.5 to 93.4%. Adding one year of age increased the predicted probability by 1.01x (or a 1% increase).

Probs				
	Coef	**Value**	**Coef*Value**	**Output (in log odds)**
Intercept	-2.977707	1	-2.977707	2.650268
age	0.144545	41	5.926345	**Probability**
income	-0.006551	50	-0.32755	0.9340275066
months subbed	0.001459	20	0.02918	

However, let's look at another person who is 24, makes 50K a year, and has been subscribed for 20 months. They have a predicted probability of upgrading of 54.8%.

Probs				
	Coef	**Value**	**Coef*Value**	**Output (in log odds)**
Intercept	-2.977707	1	-2.977707	0.193003
age	0.144545	24	3.46908	**Probability**
income	-0.006551	50	-0.32755	0.5481015268
months subbed	0.001459	20	0.02918	

Keeping other variables constant and increasing their age by 1 results in a predicted probability of 58.4%. In this case, increasing age by 1 year increased the predicted probability by 1.06x (or a 6% increase). This is higher than the 1.01x increase we saw before. This is evidence that converting from odds to probabilities does *not* preserve the linear relationship between the predictor variables and the predicted value. Proceed with caution when transforming your output to probabilities. When you do, make sure you're clear about the values of the predictor variables you're using to calculate the probability. As we saw above, they can greatly affect the probability values you get (note: these issues are less important when working with only categorical predictors).

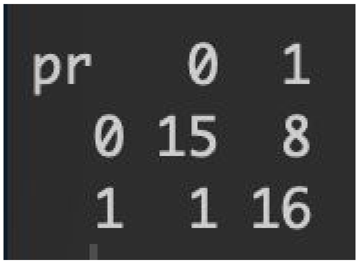


Probs				
	Coef	Value	Coef*Value	Output (in log odds)
Intercept	-2.977707	1	-2.977707	0.337548
age	0.144545	25	3.613625	Probability
income	-0.006551	50	-0.32755	0.5835947801
months subbed	0.001459	20	0.02918	

LOGISTIC REGRESSION AS A PREDICTIVE MODEL

Logistic regression is also often used as a purely predictive model, in which case we may care less about the coefficients and more about the accuracy of the model's predictions. In this case, we can take a set of data and run it through our logistic regression model to get their predicted probabilities. Then, after choosing a threshold (like 0.5) we can calculate our predictions: 0 if we predict they will not upgrade, and 1 if we predict they will.

Then, we can compare our predictions to the actual upgrade values and create something called a **confusion matrix** which shows how our predicted values matched the actual values.



```
pr   0  1
  0 15  8
  1  1 16
```

Values on the main diagonal represent predictions we got right (15, 16). Values off the diagonal (1,8) represent predictions we got wrong. We can also use the confusion matrix to calculate our model's accuracy by taking the number of items we got correct (15+16) divided by the total number of items (15 + 16 + 8 + 1). Here, our model had 77.5% accuracy; not bad!

Spreadsheet to explore LR coefficients:
https://docs.google.com/spreadsheets/d/14UtMFLP6JCe7eidFY3aQqUXvDfnFxS2oEYQqx-1gdQs/edit?usp=sharing