Jonathan M Brown
Updated 2022-07-03

Let's chat about…Logistic Regression!

(NBA 2014-15 data from Basketball-Reference.com)

I. What is "logistic regression" and why do we need it?

Logistic regression is just like regression with a number of differences:

	Linear Regression	Logistic Regression
Nature of response variable	Quantitative	Binary: 0, 1
What the model predicts	Predicts values for the dependent variable.	Predicts the logarithm of the odds, or "log-odds" of something occurring (of the response variable = 1).
How the model is found (i.e, "estimation method')	<u>OLS: Ordinary Least Squares.</u> Minimizing the error, or more specifically, the sum of the squared residuals.	<u>ML: Maximum Likelihood.</u> Maximizes the probability of the data being true, given the parameters.
Measure of "goodness of fit"	R^2	Residual deviance
What the regression line looks like	Before any data transformations: a straight line, or a linear "surface" in multiple regression. After data transformations: a variety of curved lines or curved "surfaces".	An S-curve

OK, that's a bunch of stuff, can we simply what's happening here?

Yes, let's start with the basics and then give an example.

We use logistic regression whenever we have a response variable that takes on ONLY two values, generally, 0 or 1, which can stand for a number of things. As a response variable, we might think of 0 as "failure" or "something NOT happening". We might think of 1 as "success" or "something happening".

Do you have some examples of this?

Examples of binary response variables that use 0 or 1:

- Admission into college:
    - 0 = for not admitted
    - 1 = for admitted
- Medical drug response:
    - 0 = for not recovered from illness
    - 1 = recovered from illness
- Employment status:
    - 0 = for unemployed
    - 1 = employed

That seems simple enough, why can't we run a simple linear regression on this?

We can, but we run into a problem as you'll see!

To show this, let's use an example with the NBA data. Suppose we want to predict whether an NBA team will make the playoffs. This is our **PLAYOFF** variable:

- If **PLAYOFF** = 0, then the team did not make the playoffs
- If **PLAYOFF** = 1, then the team made the playoffs

Let's say that we want to explain the **PLAYOFF** variable using, average points per game, **PTS.G**. That is, we think a team's average points per game might help us predict whether or not that team makes the playoffs.

Thus, this would be the population model:

$$PLAYOFF = \beta_0 + \beta_1 PTS.G + \varepsilon$$

Also, here's what several teams in the data look like:

	Team	PLAYOFF	PTS.G
1	Atlanta Hawks	1	102.5
2	Boston Celtics	1	101.4
3	Brooklyn Nets	1	98.0
4	Charlotte Hornets	0	94.2
5	Chicago Bulls	1	100.8
6	Cleveland Cavaliers	1	103.1
7	Dallas Mavericks	1	105.2
8	Denver Nuggets	0	101.5

When you say "predict" whether or not a team will make the playoffs, what does that mean?

Good question!

Remember what our simple linear regression model predicts? It predicts the AVERAGE VALUE of the response variable at a given value of the explanatory variable!

The same is true here: We want to predict the "average" value for the **PLAYOFF** variable.

a. Average value of a 0,1 variable

What does "average" mean when the variable is 0,1?

The short answer is that for a 0, 1 variable, the "average" value is the probability of an individual case being a 1 (and whatever a "1" represents!).

Let's see some examples.

Pretend I only have 10 NBA teams and that none of them made the playoffs. What would that data look like?

```
   Team Playoff
1     A       0
2     B       0
3     C       0
4     D       0
5     E       0
6     F       0
7     G       0
8     H       0
9     I       0
10    J       0
```

What's the average value for **Playoff**?

$$Playoff\ Average = \frac{0+0+0+0+0+0+0+0+0+0}{10} = 0$$

When no teams make the playoffs, the average of **Playoff** = 0.

Put another way: What proportion of teams made the playoffs? 0. Thus, the probability of a team in this sample making the playoffs is 0 (or 0%).

But what if ALL teams made the playoffs?

What would all teams making the playoffs look like in the data?

```
   Team Playoff
1     A       1
2     B       1
3     C       1
4     D       1
5     E       1
6     F       1
7     G       1
8     H       1
9     I       1
10    J       1
```

Now, what's the new average value for **Playoff**?

$$Playoff\ Average = \frac{1+1+1+1+1+1+1+1+1+1}{10} = \frac{10}{10} = 1$$

Put another way: When, all teams made the playoffs, what proportion of teams made the playoffs? 1. Thus, the probability of a team in this sample making the playoffs is 1 (or 100%).

OK, but what about when some teams made the playoffs but some didn't?

Let's say that only FOUR teams made the playoffs. What might that data look like?

```
   Team Playoff
1     A       0
2     B       1
3     C       0
4     D       0
5     E       1
6     F       1
7     G       0
8     H       0
9     I       0
10    J       1
```

Again, what's the average value for **Playoff**?

$$Playoff\ Average = \frac{0+1+0+0+1+1+0+0+0+1}{10} = \frac{4}{10} = 0.4$$

Aha! The proportion of teams making the playoffs = 0.4. Thus, the average value of 0.4 indicates that the probability of a team having made the playoffs is 0.4 or 40%.

That's it?

Yes: **The average value for a 0,1 variable is the probability that a single case within the dataset is a 1 or whatever being a 1 represents, e.g., "success".**

But wait, regression predicts average values of the response variable, so what does this tell us?

Exactly! **By using a 0,1 response variable, linear regression predicts the probability of a case being a 1, however a 1 is defined. If a 1 indicates "success", then linear regression predicts the probability of "success".**

So, with our actual NBA data, let's find the average of the **PLAYOFF** variable:

```
> mean(nba15$PLAYOFF)
[1] 0.5333333
```

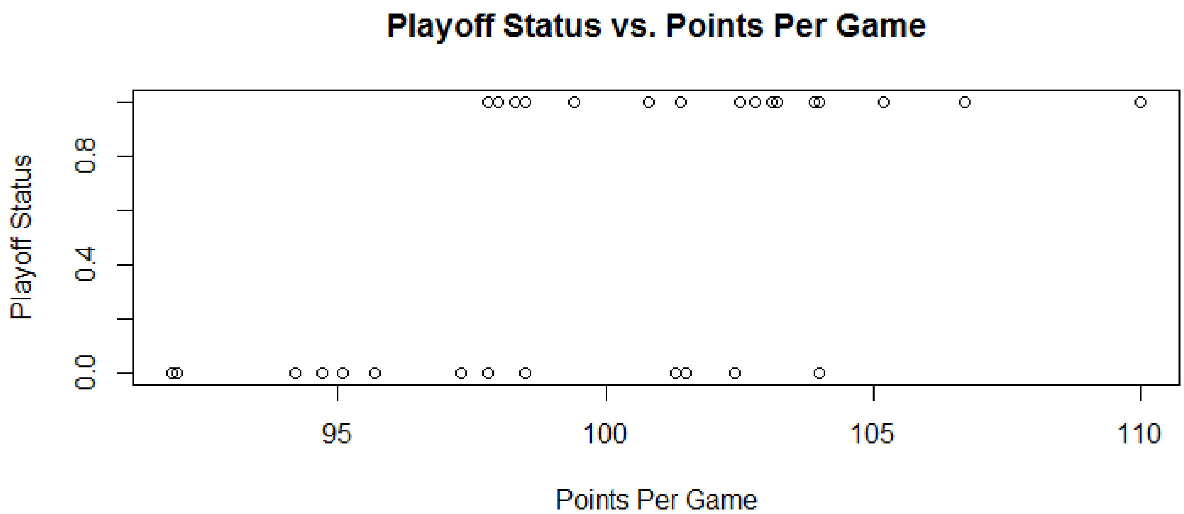
Thus, we see that since the proportion of teams making the playoffs = 0.53. Therefore, if we randomly were to choose one team, then we have a 53% chance of choosing a playoff team!

But as we know, this probability may depend on a number of factors! Just as housing price might depend on square feet, age, and other factors, the probability of making the playoffs may depend on other factors.

Let's model it!

b. Plotting a 0,1 variable

Our first step in regression has been to make a scatterplot of the response vs. explanatory variables:



Huh, is that how it's supposed to look?

Yes! Since we ONLY have 0s and 1s, we will only see two horizontal rows of dots.

That doesn't look linear, isn't that a problem?

In general, yes. And that's where logistic regression comes in, but let's keep going anyway.

c. A simple linear regression with a 0,1 response variable

Reminder of population model:

$$PLAYOFF = \beta_0 + \beta_1 PTS.G + \varepsilon$$

After estimating the above model with the data, we obtain these results:

```
> summary(lm(PLAYOFF ~ 1 + PTS.G, data = nba15))

Call:
lm(formula = PLAYOFF ~ 1 + PTS.G, data = nba15)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.79880 -0.32563  0.01443  0.29617  0.61439 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -6.13216    1.85762  -3.301  0.00263 **
PTS.G        0.06664    0.01856   3.591  0.00124 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.4273 on 28 degrees of freedom
Multiple R-squared:  0.3154,	Adjusted R-squared:  0.2909 
F-statistic:  12.9 on 1 and 28 DF,  p-value: 0.001242
```

This gives us the following regression line:

$$\widehat{PLAYOFF} = -6.13 + 0.067 * PTS.G$$

Since we're predicting probabilities, another symbol is sometimes used:

$$\hat{\pi}_{Playoff} = -6.13 + 0.067 * PTS.G$$

Where $\hat{\pi}$ is just the predicted probability of a team being in the playoffs.

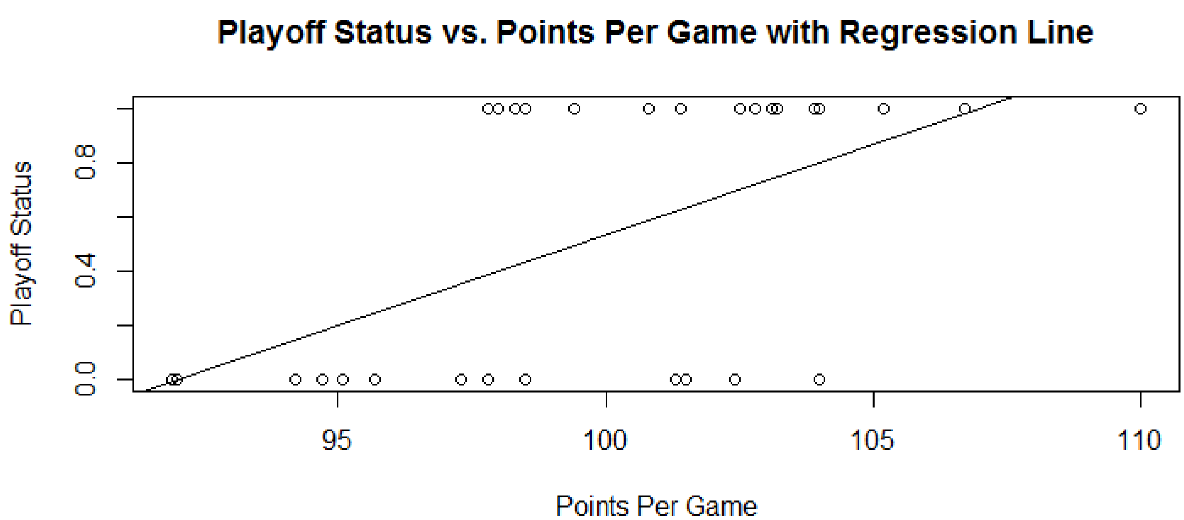
How do we interpret our coefficients?

This model is called the **linear probability model** because it uses a straight line to predict probabilities. Our coefficient interpretations reflect this:

- $\widehat{\beta_0} = -6.13$ For a team that scores 0 points-per-game, the probability of making the playoffs is -613%. That doesn't seem possible!! For several reasons, we wouldn't interpret this intercept. The primary reason is that we don't have teams that scored 0 points per game.
- $\widehat{\beta_1} = 0.067$ For a 1 point-per-game increase in a team's average points-per-game, there is an associated 6.7 percentage-point increase in the probability of making the playoffs. The higher the points per game a team had, the higher their chances of making it into the playoffs.

BUT, the story doesn't end here!

If we were to plot our regression line on our original data, it would look like this:



This looks kind of strange, but maybe it's OK. Right?

Here's the problem:

Remember the nature of our dependent variable: It is either a 0 for "failure" or a 1 for "success". That allows us to interpret things as probabilities. <u>But what if we get a predicted value SMALLER than 0 or GREATER than 1??</u>

That is, what does a negative probability mean?

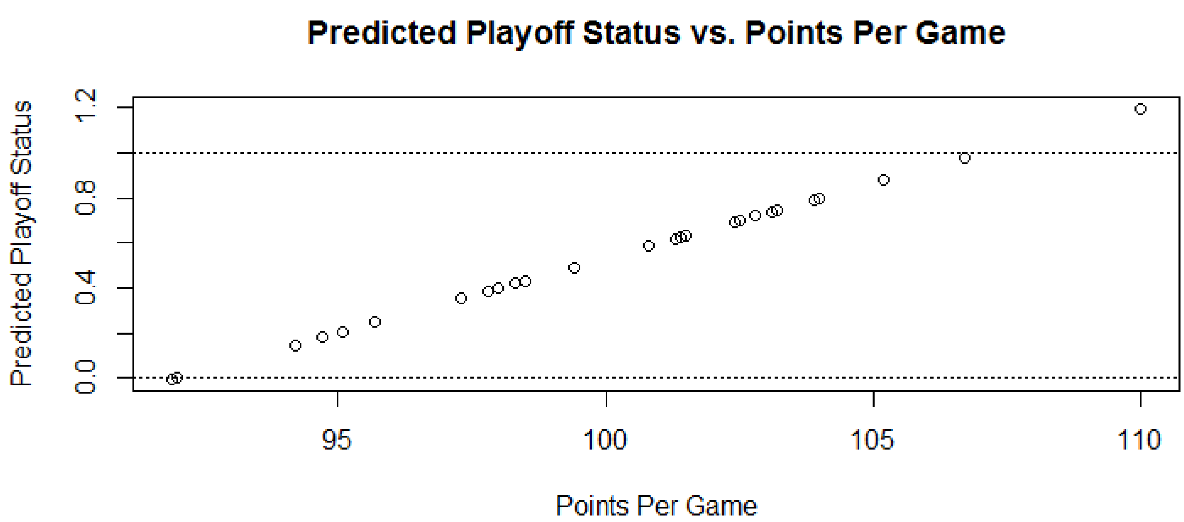
What does a probability greater than 100% mean?

<u>Both are meaningless. They are flaws from the fact that we are trying to fit a straight line on a variable that only takes on TWO values!</u>

Thus, we may get a regression line that makes IMPOSSIBLE predictions.

Can we see an example of an "impossible" prediction?

To see this, let's plot the predictions of our model with a reference line at 0 (meaning 0% probability of occurring or "failure") and a reference line at 1 (meaning 100% probability of occurring or "success").



We see that we have one prediction WELL ABOVE a probability 1 (or 100% chance), and we have a couple of predictions just lower than 0. These predictions came from points-per-game values that are within the range of the data but are still giving nonsensical predictions.

For a 0,1 variable we MUST make predictions that are BETWEEN 0 and 1!

Therefore, while our linear probability model is easy to interpret, it does NOT do a good job at modeling 0,1 variables realistically. We must look elsewhere.

OK, so what else can we do to prevent predictions from falling out the range of 0 to 1?

Short answer: Fit a model that is constrained to make predictions between 0 and 1!

You're telling me that there are models that are specifically bounded by 0 and 1?

YES, and THAT is where logistic regression helps us.

II. The logistic regression model

While linear regression models things using a LINE, logistic regression models things using an S-shaped curve that is BOUNDED by 0 and 1. It's that simple!

Wait, what does that actually look like in terms of the model?

This is where I'll skip most of the math and jump to the conclusion.

If we go back to our notation where $\boldsymbol{\pi}$ is the probability of making it into the playoffs, the logistic model that looks like an S-curve and is constrained to predict values between 0 and 1 looks like this:

Logistic Model

$$\pi_{Playoff} = \frac{e^{\beta_0+\beta_1 PTS.G}}{1+\ e^{\beta_0+\beta_1 PTS.G}} = \frac{1}{1+\ e^{-(\beta_0+\beta_1 PTS.G)}}$$

For comparison, recall the linear probability model we just saw:

Linear Model

$$\pi_{Playoff} = \beta_0 + \beta_1 PTS.G + \varepsilon$$

Whoa! That logistic model looks a little crazy.

Yes, it can be a mess. But that logistic model indeed gives predicted probabilities that are constrained between 0 and 1. The *e* is the exponential function, like you've hopefully seen before. Also, the error term goes away because we are predicting a probability, which is already considered random!

Is there a nicer way to write this model?

To make things more interpretable, that whole mess of an equation can be rearranged to this:

The logistic model as typically used:

$$\ln\left(\frac{\pi_{Playoff}}{1-\pi_{Playoff}}\right) = \beta_0 + \beta_1 PTS.G + \varepsilon$$

OK, that looks a little better, but what are we predicting now in this form?

Instead of predicting direct probabilities, we predict what's called the **log-odds** of success (of being a 1).

What are log-odds?

Quick probability detour:

If the probability of success = π, then we define the odds of success as the probability of success divided by the probability of failure. If we take the natural logarithm of the odds, then we have the log-odds!

In summary:

	Formula	Explanation
Probability	π	The probability of "success" (of being a 1). Values are between 0 and 1.
Odds	$\frac{\pi}{1-\pi}$	The odds of success (of being a 1). This is the probability of "success" divided by the probability of "failure". Values are between 0 and infinity.
Log-odds	$\ln\left(\frac{\pi}{1-\pi}\right)$	This is simply taking the natural logarithm of the odds. Values are between negative infinity and positive infinity.

Wait, we now predict log-odds instead of probability? How do we interpret log-odds?

For the value of the log-odds we predict, we can simply use the formulas above to transform the log-odds prediction to an odds prediction. Then, we can convert the odds value to a probability value.

Can we see an example of how that works?

Suppose my model predicts a log-odds of "success" to be 1.7, as shown:

$$\log \text{odds} = \ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 1.7$$

To find the odds of success, we undo the natural logarithm with the exponential function:

$$\text{odds} = e^{\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)} = \frac{\hat{\pi}}{1-\hat{\pi}}$$

Let's carry this out! $e^{1.7} = 5.47$

Thus, the odds of success are 5.47, or the probability of success is 5.47 times that of the probability of failure.

To convert this to a predicted probability, we use the following math:

$$\text{Probablity } = \hat{\pi} = \frac{odds}{1+odds} = \frac{5.47}{1+5.47} = 0.85$$

Thus, the probability of making the playoffs = 85%.

Is there a way to summarize these relationships?

	Predicted Probability	Predicted Odds	Predicted Log-odds
Probability Success > Probability Failure	$\hat{\pi} > 0.5$	Odds > 1 $\frac{\hat{\pi}}{1-\hat{\pi}} > 1$	Log-odds > 0 $\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) > 0$
Probability Success = Probability Failure	$\hat{\pi} = 0.5$	Odds = 1 $\frac{\hat{\pi}}{1-\hat{\pi}} = 1$	Log-odds = 0 $\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 0$
Probability Success < Probability Failure	$\hat{\pi} < 0.5$	Odds < 1 $\frac{\hat{\pi}}{1-\hat{\pi}} < 1$	Log-odds < 0 $\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) < 0$

OK, can we actually run the logistic model now? How do we do that?

Fortunately, with the right code, R does all of the work for us!

To run a logistic regression, we use the glm() function. To make sure we get lot-odds, we specify “logit” as highlighted below:

```
> glm.1 = glm(PLAYOFF ~ 1 + PTS.G, data = nba15, family = binomial(link = "lo
git"))
```

We view our results like usual:

```
> summary(glm.1)

Call:
glm(formula = PLAYOFF ~ 1 + PTS.G, family = binomial(link = "logit"),
    data = nba15)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9425  -0.7916   0.2688   0.7342   1.4693

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -38.3098    14.6018  -2.624  0.00870 **
PTS.G         0.3849     0.1462   2.633  0.00847 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 30.229  on 28  degrees of freedom
AIC: 34.229

Number of Fisher Scoring iterations: 4
```

This gives us an estimated model, in terms of log-odds of making the playoffs:

$$\ln\left(\frac{\hat{\pi}_{PLAYOFF}}{1-\hat{\pi}_{PLAYOFF}}\right) = -38.3 + 0.38 * PTS.G$$

III. Interpreting the logistic regression model

How do we interpret things?

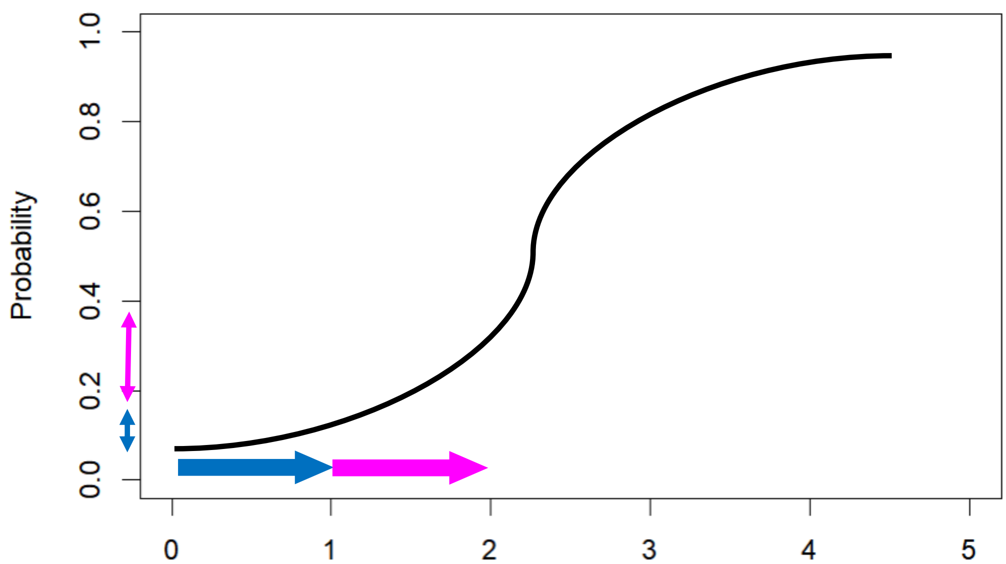
First, notice that we are given p-values like regular (but now Z-scores instead of t-scores; we use Z-scores because we are predicting probabilities, which changes things). Therefore, we find evidence that **PTS.G** is a potentially useful predictor of **PLAYOFF** status, as p = 0.008.

Next, let’s interpret the coefficients:

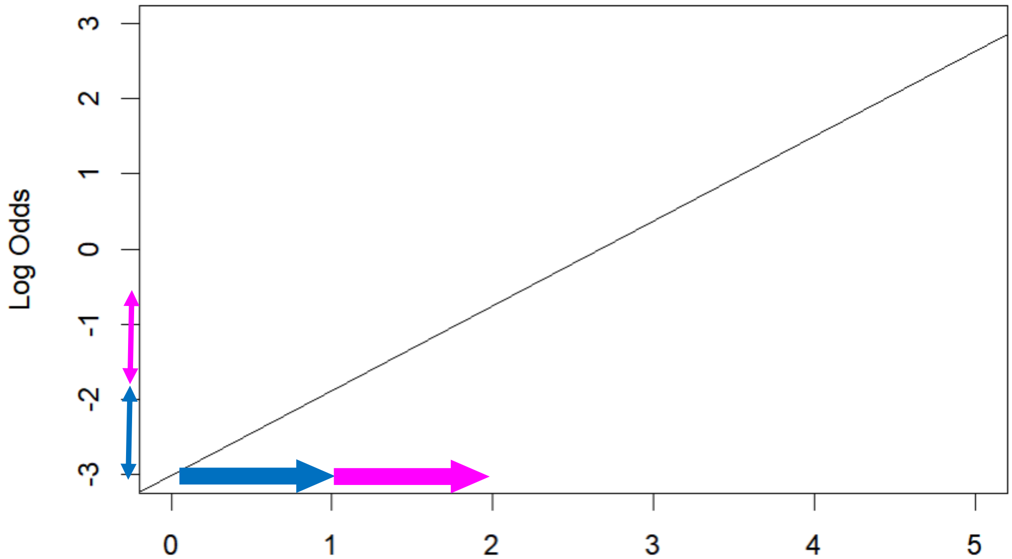
- $\widehat{\beta_0} = -38.3$ For teams that scored 0 points-per-game, the log-odds of making the playoffs is -38.3. The intercept still doesn't make sense, as we don't have teams that score 0 points-per-game.
- $\widehat{\beta_1} = 0.38$ Each additional point-per-game that a team scored is associated with a 0.38 increase in the log-odds of making the playoffs.

Since we're in log-odds, that's about as much as we can say without more mathematical work! Why? We can convert this slope of 0.38 to a probability, but since our model is CURVED, the change in probabilities is NOT constant over the range of our data. That is, the change in probabilities from one part of the curve is DIFFERENT than the change in probabilities for another part of the curve, as we look at changing values of points-per-game.

This is shown with an example S-curve below. Look at the change along the y-axis when moving from 0-to-1 vs. 1-to-2 along the x-axis. The change in probability is BIGGER for the latter case (moving 1-to-2).



However, if we stay in log-odds, then the general shape is a line. Thus, the change in y vs. x is a constant slope, as shown by the example below. (The change in y is the same for each one-unit change in x.)



In summary:

- We can talk about a constant change log-odds.
- We can plug in different values for points-per-game, convert the predicted log-odds to odds, convert from odds to probability, and then see how the probability changes between different teams.

Is there a way to interpret the slope in some way OTHER than log-odds or probability?

Yes! If you apply the exponential, e^x, to the slope, then you transform the slope into something called the ODDS RATIO!

What is an "odds ratio"?

It's exactly as it sounds: It's the RATIO of ODDS of success occurring (for two different events).

For example, if the odds of the Milwaukee Bucks making the playoffs were 0.5 and the odds of the Minnesota Timberwolves making the playoffs were 0.3, then the ODDS RATIO would be 0.5 / 0.3 = 1.67

In other words, the odds of the Bucks making the playoffs were 1.67 times HIGHER than the Timberwolves.

Put another way: There is a 1.67-fold increase in the odds of making the playoffs.

*So, how do we interpret the slope of **PTS.G**?*

Let's do an example. First, take e^x of the slope of **PTS.G**, which was 0.38:

$$e^{1.38} = 1.46$$

Thus, we interpret the 1.46 as an odds-ratio. BUT, a ratio implies TWO things. So, what two things are we comparing?

We are comparing TWO teams where the DIFFERENCE between them is a ONE-unit difference in the explanatory variable (e.g., a ONE-point difference in **PTS.G**).

To interpret: For a one-point increase in **PTS.G**, there is a 1.46-fold increase in the odds of a team making the playoffs!

Can you show an example of how this odds-ratio works for the slope?

Suppose we have one team that scored 91 points per game vs. another team that scored 90 points per game. Thus, the odds that the first team makes the playoffs is 1.46 times higher than the odds that the second team makes the playoffs.

To get the odds that the first team makes the playoffs, plug in 91 for **PTS.G**, obtain the predicted log-odds from the model, and then use e^x on the result:

$$log\ odds\ (playoffs) = -38.3 + 0.38 * PTS.G$$

$$-38.3 + 0.38(91) = -3.72$$

$$e^{-3.72} = 0.024$$

Thus, the odds of the first team making the playoffs = 0.024

Next, repeat the above for a team scoring 90 points-per-game:

$$log\ odds\ (playoffs) = -38.3 + 0.38 * PTS.G$$

$$-38.3 + 0.38(90) = -4.1$$

$$e^{-4.1} = 0.017$$

Thus, the odds of the second team making the playoffs = 0.017

So, what's the RATIO of these two odds?

$$Odds\ Ratio = \frac{Odds\ for\ PTS.G = 91}{Odds\ for\ PTS.G = 90} = \frac{0.024}{0.017} = \mathbf{1.41}$$

Thus, the odds of first team (PTS.G = 91) making the playoffs is 1.41 times higher than the odds of the second team (PTS.G = 90).

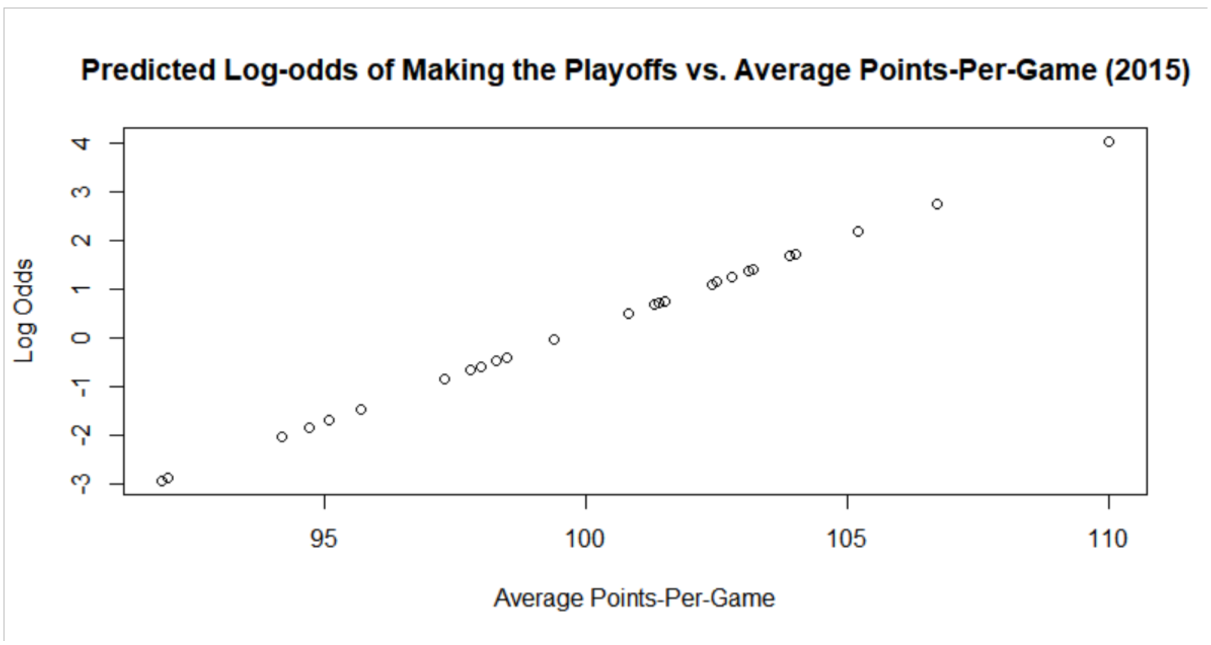
Other than rounding issues, this odds ratio of 1.41 matches the original slope when it's converted to an odds ratio of 1.46.

Ta-dah!

Couldn't we see this with some graphing?

YES! In fact, to show you how this works:

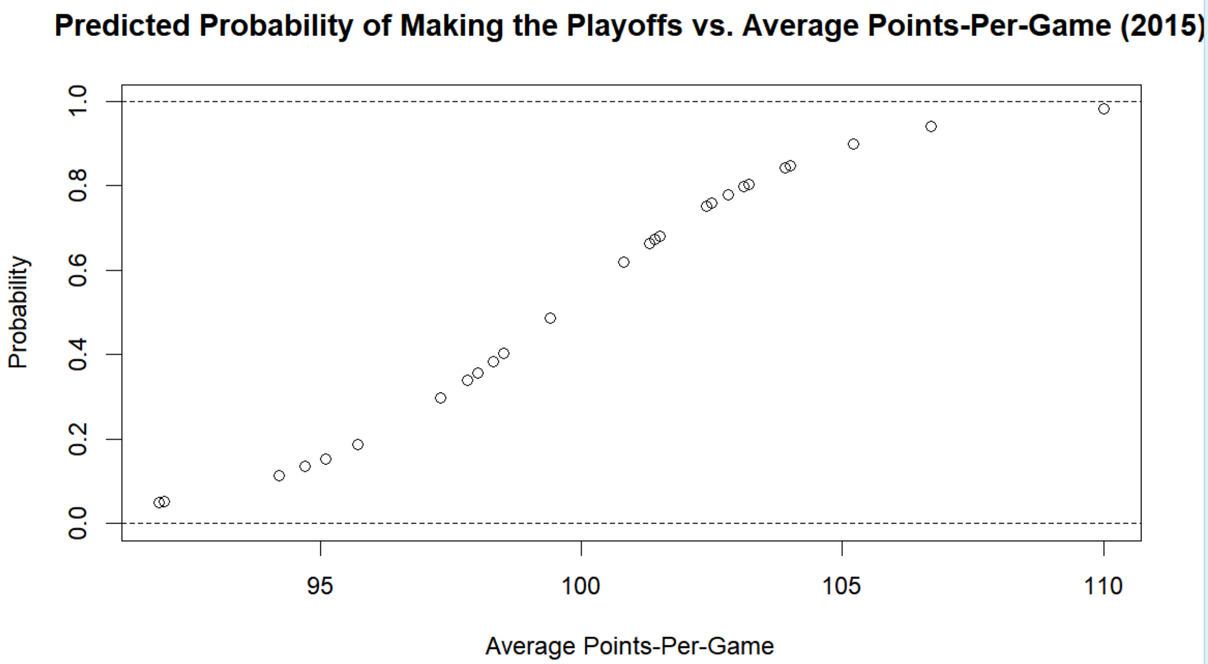
Here's a graph of the predicted log-odds. This will be a straight line, because of right-hand side of the regression equation looks just like our usual linear regression:



This looks just like a regular regression line, except now our y-axis is log-odds of success, instead of probability of success.

What if we convert everything to probabilities, what will that look like?

If we convert everything to probabilities, it will look like this:



Wait, is that an S-curve?

Ta-dah! YES! Remember, our ORIGINAL logistic model wanted to constrain our predictions of probabilities to be between 0 and 1. By using logistic regression, we use an S-curve as our model. Therefore, when we convert our predicted log-odds back to probabilities, it should look like the S-curve that we want!

I added a dotted lines at y = 0 and y = 1 to show this. As you can see, all predictions fall between 0 and 1 and are in an "S shape".

IV. Quality of the logistic regression model

One last question, how do we measure error or "goodness of fit"?

This gets complicated.

The short answer is that we can use the "residual deviance". Let me re-paste the output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept) -38.3098    14.6018  -2.624  0.00870 **
PTS.G         0.3849     0.1462   2.633  0.00847 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 30.229  on 28  degrees of freedom
AIC: 34.229

Number of Fisher Scoring iterations: 4
```

The residual deviance is a way of measuring "error", but it's a whole separate topic to explain what that is. Residual deviance is a special kind of error. We want the residual deviance as small as possible. By comparison the "null deviance" is the "error" if we just used the average value of the response variable as our model, that is, if we only included the intercept in our model and NO explanatory variables. As you can see, including **PTS.G** reduced our "deviance" from 41 to 30, which means **PTS.G** did help us with our predictions.

Lastly, the "AIC" or Akaike Information Criterion is another way to measure model fit. It only has meaning when compared to other models. If you run a bunch of models, the smaller the AIC, the better the fit. Without getting too technical: AIC is a way to measure how much "information" a model loses compared to using the true best model (which we may never know). Thus, the smaller the information loss, the better. Therefore, the smaller AIC, the better, when comparing many models.

How is this model actually estimated?

In "Ordinary Least Squares" from multiple regression, we minimize the sum of the squared residuals.

Logistic regression uses something called "Maximum Likelihood". This approach finds the values for the coefficients that MAXIMIZE the probability that the model is the "truest" model, given whatever data we have. It is a much different way of thinking about modeling.

What about assumptions?

There are assumptions, but there are varying perspectives on this.

<u>The MOST IMPORTANT assumption is the Independence Assumption.</u>

This assumption assumes the residuals or errors from the model are NOT RELATED to each other (knowing something about one error should NOT tell you anything about another error).

Working backward, this essentially assumes the OBSERVATIONS of the RESPONSE VARIABLE are independent from one another (unless we account for the explanatory variables in the model that explain how the response variable observations are related to each other).

This is assumption is essential, because the coefficients for the model are found by computing the JOINT PROBABILITY of a set of things from the data. Joint probability is REALLY

MESSY to calculate if your things are related to each other. Thus, the underlying math is MUCH SIMPLER if we just assume our errors are unrelated: the joint probability calculation that helps the computer estimate the intercept, slope, etc., is way easier.

Loosely, the degree to which the Independence Assumption is violated is the degree to which you cannot trust the results of your model.

There are other assumptions, but this is just the start.

Logistic regression is cool and powerful, but it certainly is in a different world from linear regression. Thus, it takes a bit more work to use and understand it!