Advanced Data Analysis I

Generalized Linear Models and Logistic Regression

PA 541 Week 12

Michael D. Siciliano
Department of Public Administration
College of Urban Planning and Public Affairs

Admin Stuff

- HW 3 will be available tonight. Will not be due for two weeks (April 12th).
- Remember, all labs are also recorded if you are unable to attend.

Starter Question

The model below predicts the log of the price of the home (*lprice*), based off of the logged values of house size (*lsqrft*) and lot size (*llotsize*). The output for this model is below:

$\ln(\text{Price}_i) = \beta_0 + \beta_1 \ln(\text{homeSize}_i) + \beta_2 \ln(\text{lotSize}_i) + \varepsilon_i$

```
Call:
lm(formula = lprice ~ lsqrft + llotsize, data = house)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.6533 -0.1105 -0.0065  0.1182  0.6642 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.6401     0.6019   -2.72   0.0078 ** 
lsqrft        0.7624     0.0809    9.43  7.4e-15 ***
llotsize      0.1685     0.0385    4.38  3.4e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.185 on 85 degrees of freedom
Multiple R-squared: 0.635,	Adjusted R-squared: 0.627 
F-statistic:   74 on 2 and 85 DF,  p-value: <2e-16
```

Is this the best functional form from a theoretical standpoint? Would a semi-log model be better?

How do we interpret the coefficient on the log of house size?

How do we interpret the coefficient on the log of lot size?

Would it be possible to add the variable on the number of bedrooms to this model without it needing to be log transformed?

- **Linear:** No transformations
  - DV = Intercept + B1 * IV + Error
  - "One unit increase in IV is associated with a (B1) unit increase in DV."
- **Log-Linear**: Outcome transformed
  - log(DV) = Intercept + B1 * IV + Error
  - "One unit increase in IV is associated with a (B1 * 100) percent increase in DV."
- **Linear-Log**: Predictor transformed
  - DV = Intercept + B1 * log(IV) + Error
  - "One percent increase in IV is associated with a (B1 / 100) unit increase in DV."
- **Log-Log**: Outcome transformed and Predictor transformed
  - log(DV) = Intercept + B1 * log(IV) + Error
  - "One percent increase in IV is associated with a (B1) percent increase in DV."

GENERALIZED LINEAR MODELS

How do we model data when…

- The response variable is:
  - Count data expressed as proportions
  - Binary data indicating success/failure
  - Data on time to arrival or number of occurrences
- The error structure is:
  - Strictly bounded (as in proportions)
  - Unable to lead to negative fitted values (as with counts)

We will use a large class of statistical models known as generalized linear models (GLMs).

Generalized Linear Models

- All linear models are comprised of three components
  - **A random or stochastic component** that specifies the conditional distribution of the response variable, y, given the predictors (up until now we have assumed a normal distribution for y).
  - **A systematic component**; a linear function of the regressors, called the linear predictor.

$$\eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

  - And **a link function (g)** that translates from the scale of the response variable to the scale of the linear predictor. So that the conditional mean is equal to the linear combination of the regressors.

$$g[u(\boldsymbol{x})] = \eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

More on the link function

- For the OLS models, the link function is direct (and is referred to as the *identity* function).

$$u(\boldsymbol{x}) = \eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Because the direct link assumes that the expected value can take on any value from positive to negative infinity it is not appropriate for all models.
- For example, the mean of a binary outcome variable must be between 0 and 1. Therefore, we **need a link function** that allows us to translate from the scale of the response to the scale of the linear predictor.

Link function cont…

- For GLMs we apply a function (g) to the conditional mean of the response so the outcome can be modeled as a linear combination of the regressors.

$$\eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$$g[u(\boldsymbol{x})] = \eta(\boldsymbol{x})$$

- Reversing this relationship produces the inverse-link function and returns the outcome on the scale of the original response variable:

$$g^{-1}[\eta(\boldsymbol{x})] = u(\boldsymbol{x})$$

- Again, in linear regression, the transformation is the identify (that is, g(u) = u) and the data distribution is normal, with standard deviation σ, estimated from the data.
- In logistic regression, the transformation is the logit link, g(u) = log[u/(1-u)] and models the log odds. It is appropriate when the data distribution is defined by the probability for binary data and thus bounded between 0 and 1.

LINEAR PROBABILITY AND LOGISTIC REGRESSION MODELS

Linear Probability and Logistic Regression Models

- These models can be used to predict membership in a group or category of outcome for individual cases:
  - Disease or no disease
  - Graduate or drop out
  - Married or single
  - Drug use or no drug use
  - Took public transit today or did not
  - Company closed or open
  - Profits or no profits
  - War or no war
  - Employed or unemployed
- Two Questions we will look at today:
  - What is the probability of falling during a ski run based on the difficulty of the run and the season.
  - What is the likelihood of being rejected for a loan based on race, bankruptcy status, and loan amount?

Linear Probability Model

- The OLS regression model places **no restriction on the values of the independent variables** – we can use continuous, dichotomous, categorical, squared variables, etc…
- The **dependent variable, however, is assumed to be continuous**. Because our independent variables contain no restrictions on their values, Y is presumably free to vary between ± infinity.
- When our dependent variable is restricted to two values (zero or one) then the violation of this assumption warrants special attention. Which we will give it…but not just yet.

Linear Probability Model

- Typically, when we have a continuous dependent variable, we interpret β_j as the impact of a one unit change of x_j on y (holding everything else constant).
- What happens to our interpretation when the dependent variable is binary as y only changes from zero to one or from one to zero?

Linear Probability Model cont…

- In previous classes, we discussed conditional expectation and denoted it as:

$$E[y_i | \, x_1, x_2, \cdots x_k] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- When we have a binary dependent variable expectation changes to probability:

$$Prob[y_i = 1 | \, x_1, x_2, \cdots x_k] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Thus, Prob[y_i=1|x_1,x_2,…x_k] is the conditional probability of success. (We typically denote 1 as success and 0 as failure)

Linear Probability Model cont…

$$Prob[y_i = 1 | x_1, x_2, \cdots x_k] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Based on the previous slide, β_j can be interpreted as the impact of a one unit change in x_j on the probability that y equals one (holding all else constant).
- Given this definition, we are still able to apply the OLS regression models we have been using when we have binary dependent variables (though as we will see there are some problems with this).

Linear Probability Example

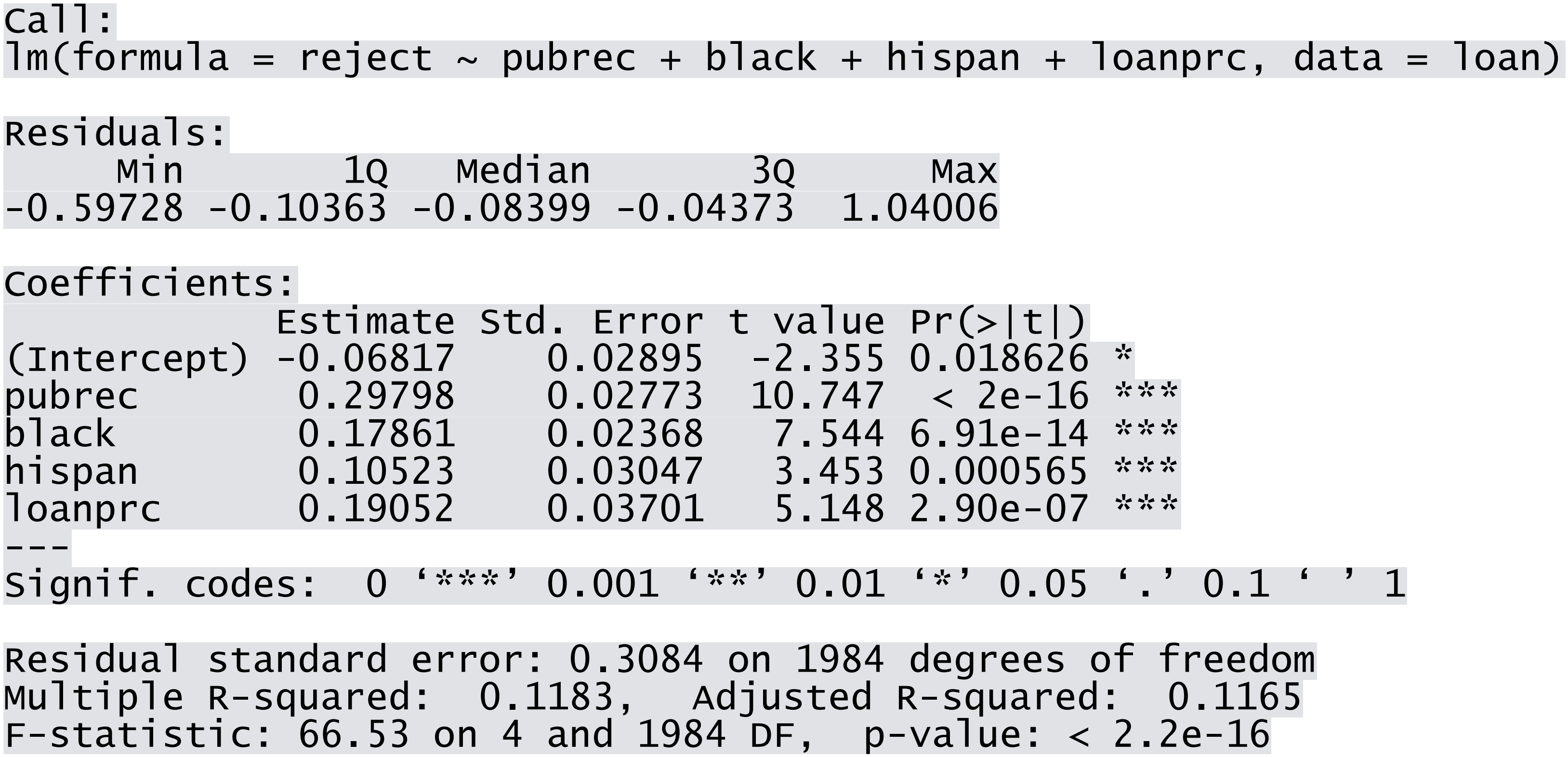
- Suppose we are interested in predicting loan rejections by banks. I run the following model:

$$reject_i = \beta_0 + \beta_1 pubrec_i + \beta_2 black_i + \beta_3 hisp_i \ + \beta_4 loanprc_i + \varepsilon$$

- Where
  - **Reject** = 1 if loan was rejected
  - **Pubrec** = 1 if person had previously filed for bankruptcy
  - **Black** = 1 if person is black
  - **Hisp** = 1 if person is Hispanic
  - **Loanprc** = is the amount of the loan divided by the price of the house
    - *(typically ranges between 0 and 1, where 1 means you are asking for a loan equivalent to the full price of the house)*

Example cont…

$$reject_i = \beta_0 + \beta_1 pubrec_i + \beta_2 black_i + \beta_3 hisp_i \ + \beta_4 loanprc_i + \varepsilon$$



```
Call:
lm(formula = reject ~ pubrec + black + hispan + loanprc, data = loan)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59728 -0.10363 -0.08399 -0.04373  1.04006 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.06817    0.02895  -2.355 0.018626 *  
pubrec       0.29798    0.02773  10.747  < 2e-16 ***
black        0.17861    0.02368   7.544 6.91e-14 ***
hispan       0.10523    0.03047   3.453 0.000565 ***
loanprc      0.19052    0.03701   5.148 2.90e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3084 on 1984 degrees of freedom
Multiple R-squared:  0.1183,	Adjusted R-squared:  0.1165 
F-statistic: 66.53 on 4 and 1984 DF,  p-value: < 2.2e-16
```

- How do we interpret each parameter estimate?
- For bankruptcy, the results indicate that someone who has filed for bankruptcy in the past has an increase in the probability of being rejected for a loan of .298.

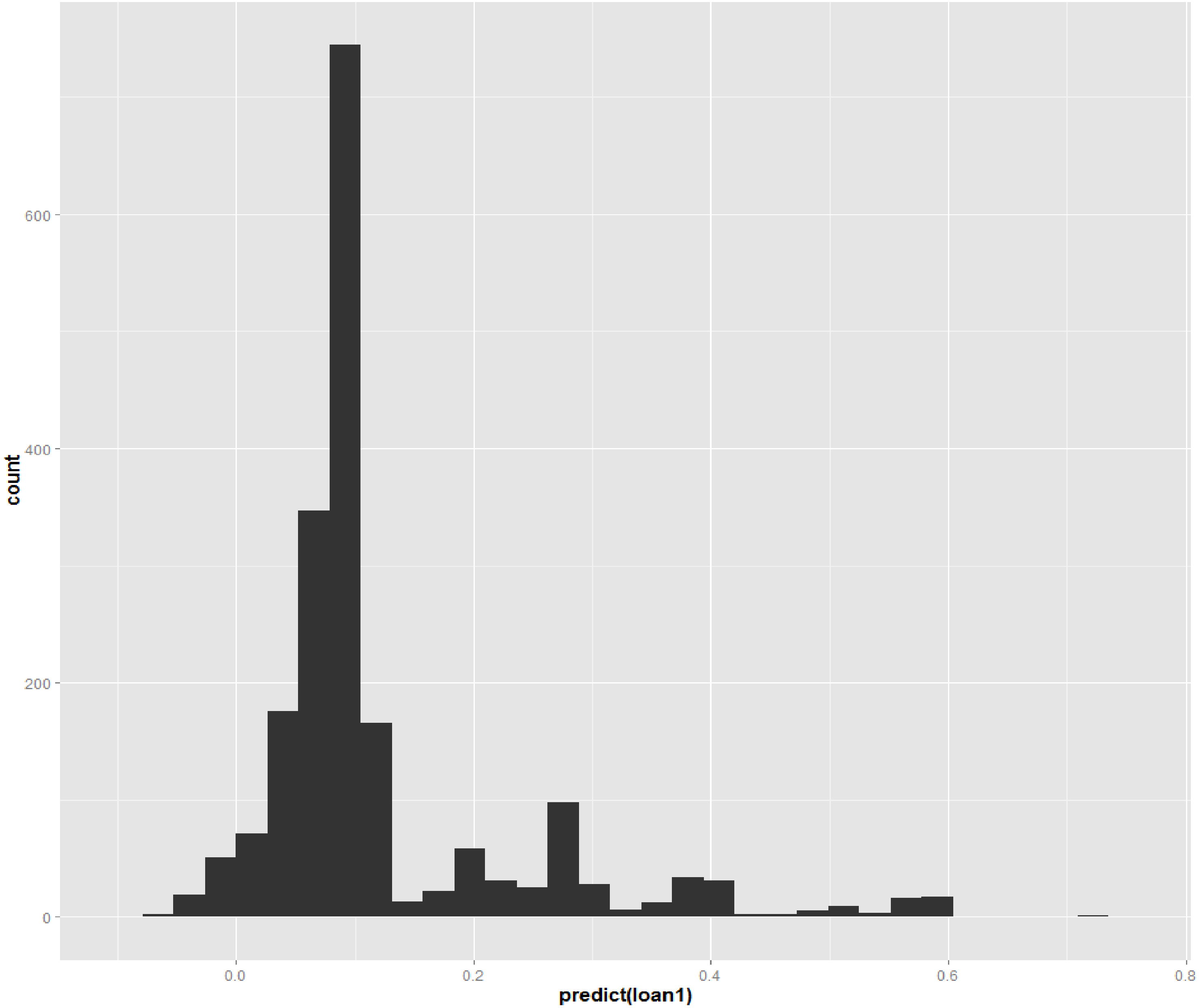
Example cont…

```
Call:
lm(formula = reject ~ pubrec + black + hispan + loanprc, data = loan)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59728 -0.10363 -0.08399 -0.04373  1.04006 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.06817    0.02895  -2.355 0.018626 *  
pubrec       0.29798    0.02773  10.747  < 2e-16 ***
black        0.17861    0.02368   7.544 6.91e-14 ***
hispan       0.10523    0.03047   3.453 0.000565 ***
loanprc      0.19052    0.03701   5.148 2.90e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3084 on 1984 degrees of freedom
Multiple R-squared:  0.1183,	Adjusted R-squared:  0.1165 
F-statistic: 66.53 on 4 and 1984 DF,  p-value: < 2.2e-16
```

- Since the coefficients are estimates of the impact of a one unit change in the regressors on the probability that y equals one, we can interpret the predicted outcome as the predicted probability that outcome equals one for the individual.
- What is the probability that a person who is (i) black, (ii) never filed for bankruptcy, and (ii) is asking for a loan amount that is twice the amount of the house, will be rejected?

Predicted Probabilities

Here is a histogram of our models predicted probability for each individual.



Issues with Linear Probability Models

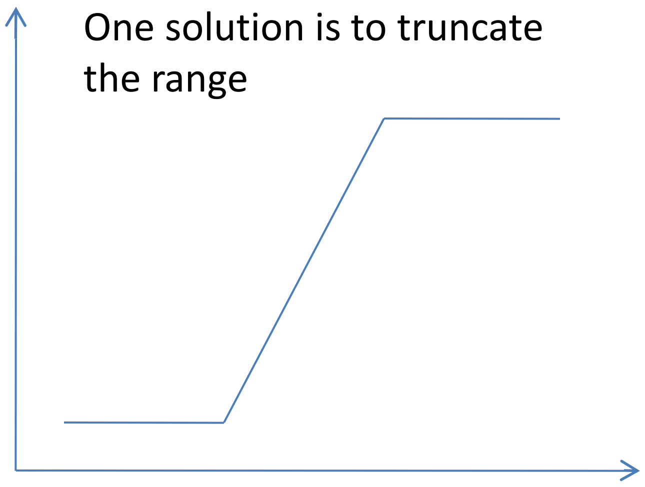
- Generally speaking, linear probability models do not constrain our predicted probabilities between zero and one.
- This aspect of linear probability models raises concern for some researchers. Hence, we will look at logistic regression models.
  - However, the linear probability model is still used by many researchers due to its simplicity and ease of interpretation.
- There are other statistical issues related with the linear probability model that we will discuss as well.

WHY LOGISTIC REGRESSION

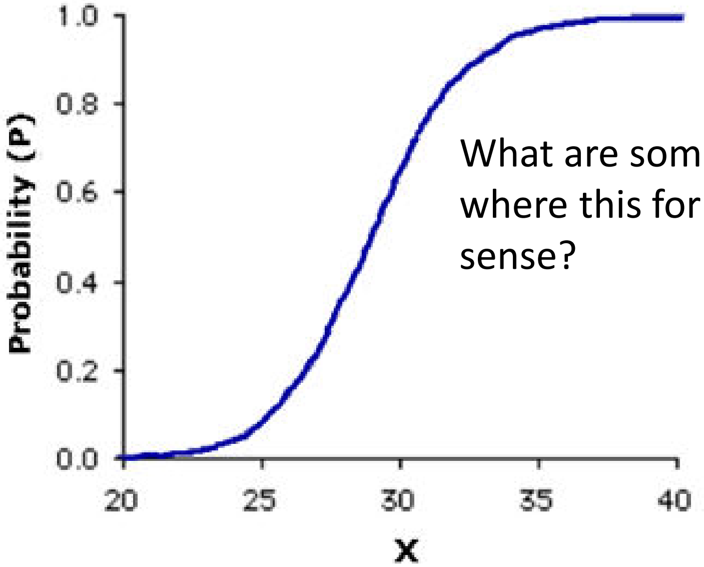
- First, it is important to note that there are other approaches to modeling binary data.
- These include:
  - Probit – link is based on the normal cumulative distribution function
  - Complementary log-log – whose link is log(-log(1-p))
- However, as argued by Faraway (2014), Hilbe (2016) and others, the logit link offers several advantages:
  - The mathematics are simpler
  - It is easier to interpret

Logistic Regression

- Problems with linear probability models (Following slides draw from Pampel 2000)
  - As we just saw, it does not restrict predicted values to fall between zero and one. This is a boundary problem.
  - Assumption of linearity: with a floor and ceiling, it seems logical that a one unit change in an IV on predicted probability of success would be smaller near the floor or ceiling



One solution is to truncate the range



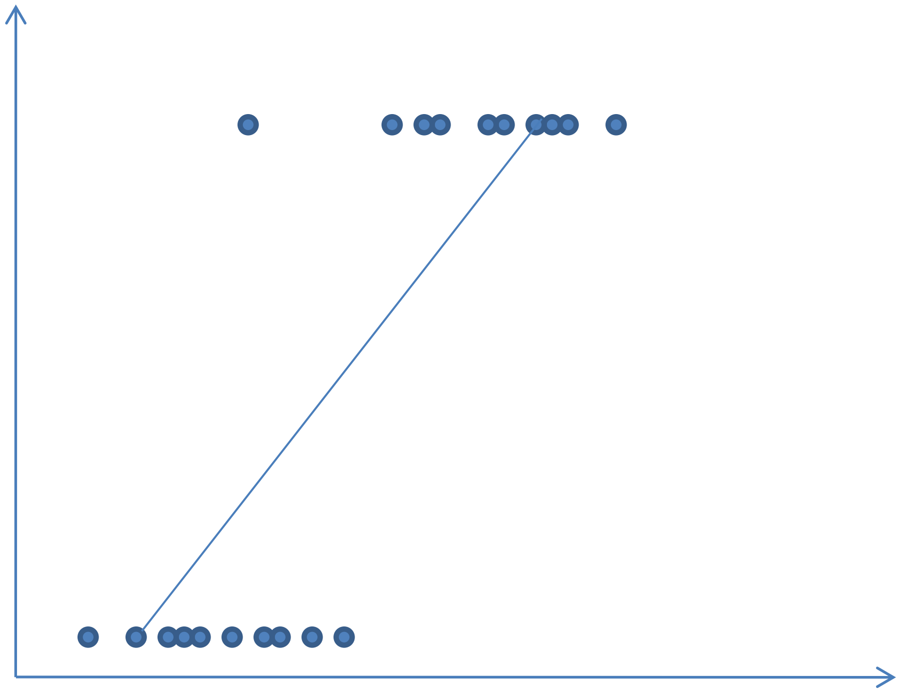
What are some examples where this form makes more sense?

Logistic Regression

- Problems with linear probability models cont…
  - Violation of the assumption of normality: Linear regression assumes a normal distribution of the error term around the predicted Y values associated with each X value. Because Y can only take on two values, the residual can only take on two values. To illustrate, the residuals take the value of:
    - $1-(\beta_0 + \beta_1 X_i)$ when Y_i equals 1, and
    - $0 - (\beta_0 + \beta_1 X_i)$ when Y_i equals 0
  - Even in the population, the distribution of errors for any X value cannot be normal when a distribution has only 2 potential values.

Logistic Regression

- Problems with linear probability models cont…
  - Violation of the assumption of homoskedasticity
    - Variances are not equal across all levels of the IV. The regression error term varies as a function of X.



Transforming Probabilities into Logits (Pampel 2000)

- Given the floor and ceiling problem of linear regression on binary dependent variables we need a transformation (i.e. a link function) to allow for **decreasing effects of X on Y** as the predicted value of Y approaches the floor or ceiling.
- There are many non-linear functions that could represent such a relationship – the logistic or logit transformation is used because of its desirable properties and relative simplicity.

Transforming Probabilities into Logits

- Each observation has a probability of experiencing some event defined as P_i. Since, we only observe the outcome as either 1 or 0, P_i is never actually observed and therefore must be estimated.
- Given this probability, the logit transformation involves two steps:
  - Take the ratio of P_i to $1-P_i$; which is the odds of experiencing the event.
  - Take the natural logarithm of the odds.

$L_i = \ln[P_i/(1-P_i)]$ this is the logged odds or logit

Why the Logit Transformation is Useful

- Odds express the likelihood of an occurrence relative to the likelihood of nonoccurence. Thus, like probabilities it has a lower limit of zero, but **has no upper bound or ceiling**. As the probability gets closer to 1, the odds become an increasingly large number. To illustrate the relationship between probabilities and odds:

P_i	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
$1-P_i$	0.99	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01
Odds	0.01	0.11	0.25	0.43	0.67	1.00	1.50	2.33	4.00	9.00	99.00

Why the Logit Transformation is Useful cont…

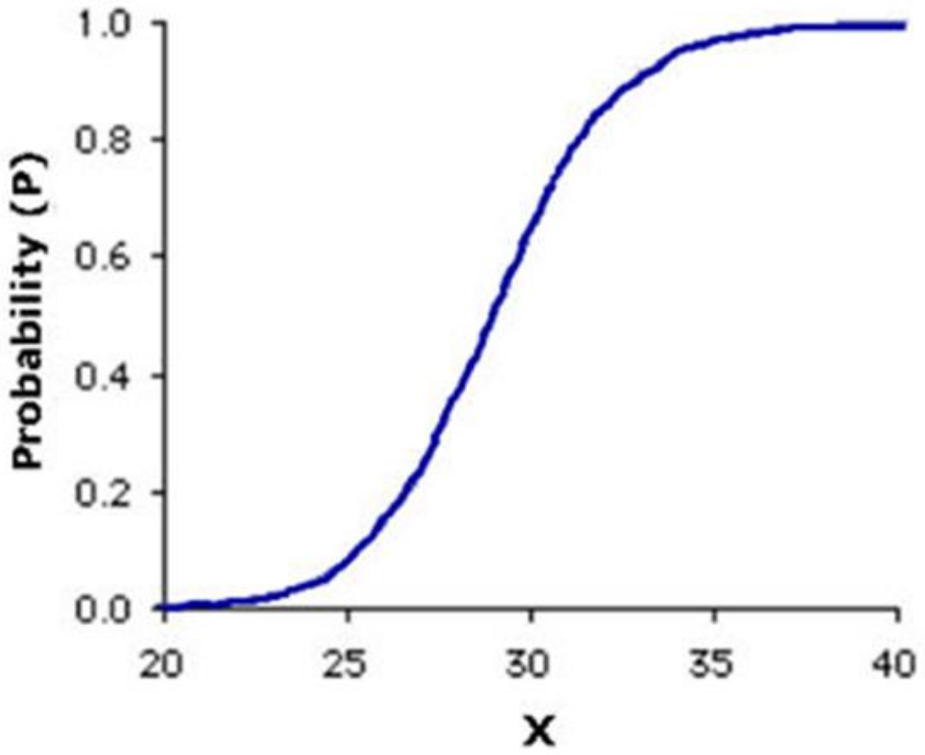
- Taking the natural log of the odds eliminates the floor (just as transforming probabilities into odds eliminating the ceiling). When using the natural log:
  - Odds above 0, but below 1, produce negative numbers
  - Odds equal to 1 produces 0; and
  - Odds above 1 produces positive numbers

Pi	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1-Pi	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Odds	0.11	0.25	0.43	0.67	1.00	1.50	2.33	4.00	9.00
Logit	-2.197	-1.386	-0.847	-0.405	0.000	0.405	0.847	1.386	2.197

- Properties of Logits:
  - No upper or lower boundary – the odds eliminate the upper boundary of probabilities, and logged odds eliminates the lower boundary of probabilities
  - The logit transformation is symmetric around the midpoint of probability of .5. The logit when P_i = .5 is zero. Probabilities below .5 produce negative logits (because the odds are less than one) and probabilities above .5 produce positive logits (because the odds exceed one)
  - The same change in probability results in different changes in logits. Thus, as P_i approaches 0 or 1, the same change in probability translates into greater change in the logged odds.

Pi	0.1	0.2	0.3	0.4	**0.5**	0.6	0.7	0.8	0.9
1-Pi	0.9	0.8	0.7	0.6	**0.5**	0.4	0.3	0.2	0.1
Odds	0.11	0.25	0.43	0.67	**1.00**	1.50	2.33	4.00	9.00
Logit	-2.197	-1.386	-0.847	-0.405	**0.000**	0.405	0.847	1.386	2.197

- Thus…the linear relationship between X and the logit implies a nonlinear relationship between X and the original probability.
- Such that a one unit change in the logit results in smaller differences in probabilities at high and low levels than at levels in the middle.



Logit	-3	-2	-1	0	1	2	3
Pi	0.047	0.119	0.269	0.5	0.731	0.881	0.953
Change	---	0.0720	0.1500	0.2310	0.2310	0.1500	0.0720

The Logistic Regression Equation

- The linear relationship between the predictors and the logit of the outcome implies non-linear relationships with the probabilities.
- The linear relationship between the predictors and the logit is defined as:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + \varepsilon$$

- Where P_i is just the predicted probability. Thus, this is the logged odds.

Obtaining Probability from Logits

- To express the probabilities rather than the logit as a function of X, first take the exponent of each side of the equation. Note that the logarithm of a number as an exponent equals the number itself: Exp (ln(X)) = X

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + \varepsilon \quad \longrightarrow \quad \frac{P_i}{1-\ P_i} = e^{\beta_0+\beta_1 x_1+\varepsilon}$$

- Solving for Pi gives the following formula:

$$P_i = \frac{\left(e^{\beta_0+\beta_1 x_1+\varepsilon}\right)}{\left(1+e^{\beta_0+\beta_1 x_1+\varepsilon}\right)}$$

Side Note

- The equation on the previous slide:

$$P_i = \frac{\left(e^{\beta_0+\beta_1 x_1+\varepsilon}\right)}{\left(1+e^{\beta_0+\beta_1 x_1+\varepsilon}\right)}$$

- Is often shown in the equivalent form of:

$$P_i = \frac{1}{\left(1+e^{-(\beta_0+\beta_1 x_1+\varepsilon)}\right)}$$

Logistic Transformation

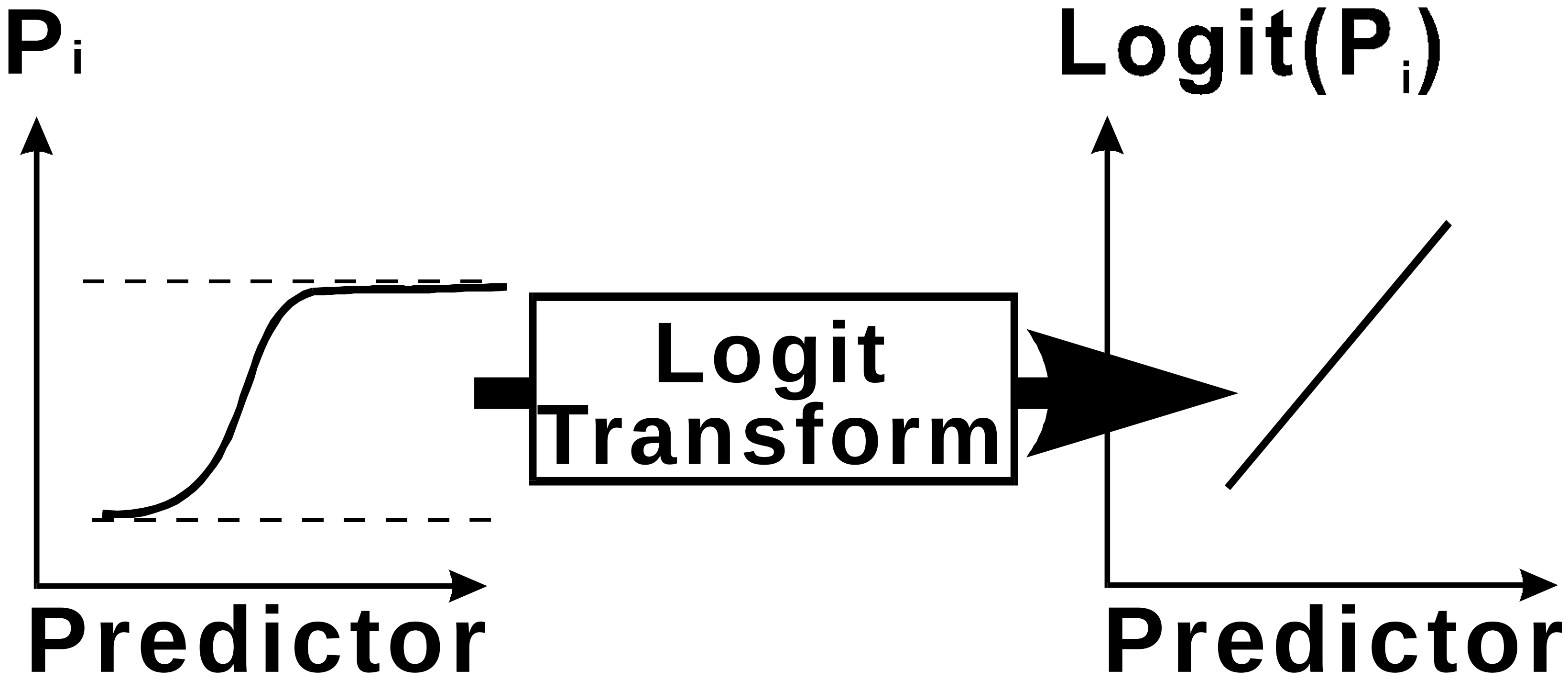


Image from: *www-personal.umich.edu/~kwelch/510/2007/handouts/LogisticRegression.ppt*

Logistic Regression Equation

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$$\hat{Y}_i = \frac{e^u}{1 + e^u} \qquad where\ u = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

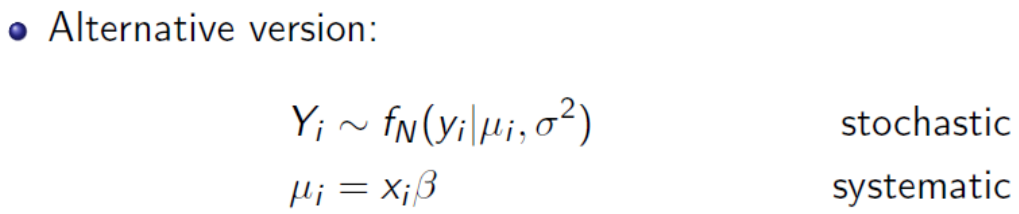
- The linear regression equation is the logit or log of the odds.
- The dependent variable is the natural log of the probability of being in one group divided by the probability of being in another.
- The procedure used for estimating the coefficients is **maximum likelihood**; the goal of this procedure is to find the best linear combination of predictors to maximize the likelihood of obtaining the observed frequencies.
- Stated differently, “maximum likelihood estimates are those parameter estimates that maximize the probability of finding the sample data that actually have been found” (Hox, 2002).
- I will have a brief discussion on MLE to start next week…

Let’s walk through a simple example

- Assume we want to predict the probability of **falling** on a ski run (0=not falling, 1=falling). The predictors in the equation will be **difficulty** of the run (ordered from 1 to 5 and will be treated as continuous) and a categorical variable for **season** (1=winter, and 0 for other).

But first…back to our definition of a statistical model

- A statistical model is a formal representation of the process by which a social system produces output.



- Alternative version:

$$Y_i \sim f_N(y_i|\mu_i, \sigma^2) \qquad \text{stochastic}$$
$$\mu_i = x_i\beta \qquad \text{systematic}$$

- Here we observe people falling or not falling on a ski run. We represent this process by a bernoulli distribution.

$$Y_i \sim f_{bern}\,(y_i|\pi_i)$$

$$\pi_i = g(x_i B)$$

Logistic Regression in R

- Logistic regression, as mentioned above, is a type of generalized linear model. These models are usually fit through the glm() function in R.
- The function follows a similar format to our lm() calls from last week:
  - glm(formula, family=family(link=function), data=), where the probability distribution (family) and corresponding default link function (function) are as follows:

Family	Default Link Function
Binomial	(link = “logit”)
Gaussian	(link = “identity”)
Gamma	(link = “inverse”)
Poisson	(link = “log”)
Quasipoisson	(link = “log”)
…and several others	

Our Model of Falling While Skiing

```
> ski$Fall = ifelse(ski$Fall=="falling", 1, 0)
> ski1=glm(Fall ~ Difficulty + Season, family=binomial(link="logit"),
data=ski)
> summary(ski1)

Call:
glm(formula = Fall ~ Difficulty + Season, family = binomial(link = "logit"),
    data = ski)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1696  -0.5409   0.2080   0.5031   2.2518

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.4990     1.5300  -2.941 0.003276 **
Difficulty     1.5688     0.4761   3.295 0.000984 ***
Seasonwinter   0.4773     1.0141   0.471 0.637861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48.263  on 34  degrees of freedom
Residual deviance: 26.470  on 32  degrees of freedom
AIC: 32.47

Number of Fisher Scoring iterations: 5
```

Interpreting the coefficients – Logged Odds

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -4.4990     1.5300  -2.941 0.003276 ** 
Difficulty     1.5688     0.4761   3.295 0.000984 ***
Seasonwinter   0.4773     1.0141   0.471 0.637861    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The logistic coefficients, β, for the predictors correspond to our OLS regression coefficients. They are often termed logistic regression coefficients, log odds-ratios, or logit coefficients. These coefficients can be used to construct prediction equations and hence generate predicted values (which we will do in the larger example below).
- We can say then, that a one unit increase on the predictor variable increases the log odds (or logit) of the dependent variable by β.
- For example, based on the output for our simple skiing data, increasing the difficulty of the ski run from 1 to 2 increases the log odds of falling by 1.569.
- The beta coefficients can vary between plus and minus infinity. A value of zero indicates that the given explanatory variable does not affect the logit.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -4.4990     1.5300  -2.941 0.003276 ** 
Difficulty     1.5688     0.4761   3.295 0.000984 ***
Seasonwinter   0.4773     1.0141   0.471 0.637861    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To help improve the interpretability of our results, we can exponentiate the coefficients.

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + \varepsilon \quad \longrightarrow \quad \frac{P_i}{1-P_i} = e^{\beta_0+\beta_1 x_1+\varepsilon}$$

```
> #to get the odds ratio of the coefficients, take the exponential
> exp(ski1$coef)
 (Intercept)   Difficulty Seasonwinter 
  0.01112051   4.80094881   1.61174752 
```

Interpreting the Coefficients - Odds

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -4.4990     1.5300  -2.941 0.003276 ** 
Difficulty     1.5688     0.4761   3.295 0.000984 ***
Seasonwinter   0.4773     1.0141   0.471 0.637861    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #to get the odds ratio of the coefficients, take the exponential
> exp(ski1$coef)
 (Intercept)   Difficulty Seasonwinter 
  0.01112051   4.80094881   1.61174752 
```

- The odds ratio, Exp(B) has a more straightforward interpretation. This value is simply the ratio of the odds for two groups where one group has a value on one of the predictor variables that is one unit larger than the other group.
- For example, β is 1.569 in our example, then the corresponding odds ratio (e^{β}) is 4.801. Hence, we can say that when the difficulty of the ski run increases by one unit, **the odds of falling are 4.8 times greater.**
- odds ratio = exp(β)   so 4.801 = exp(1.569)
  β = ln(odds ratio)   so 1.569 = ln(4.801)

More on the Odds Ratio

- What is an odds ratio?
  - From Tabachnick and Fidell p. 462: An odds ratio has a very clear intuitive understanding in a 2X2 table; it is the odds of an outcome for cases in a particular category of a predictor divided by the odds of that outcome for the other category of the predictor.

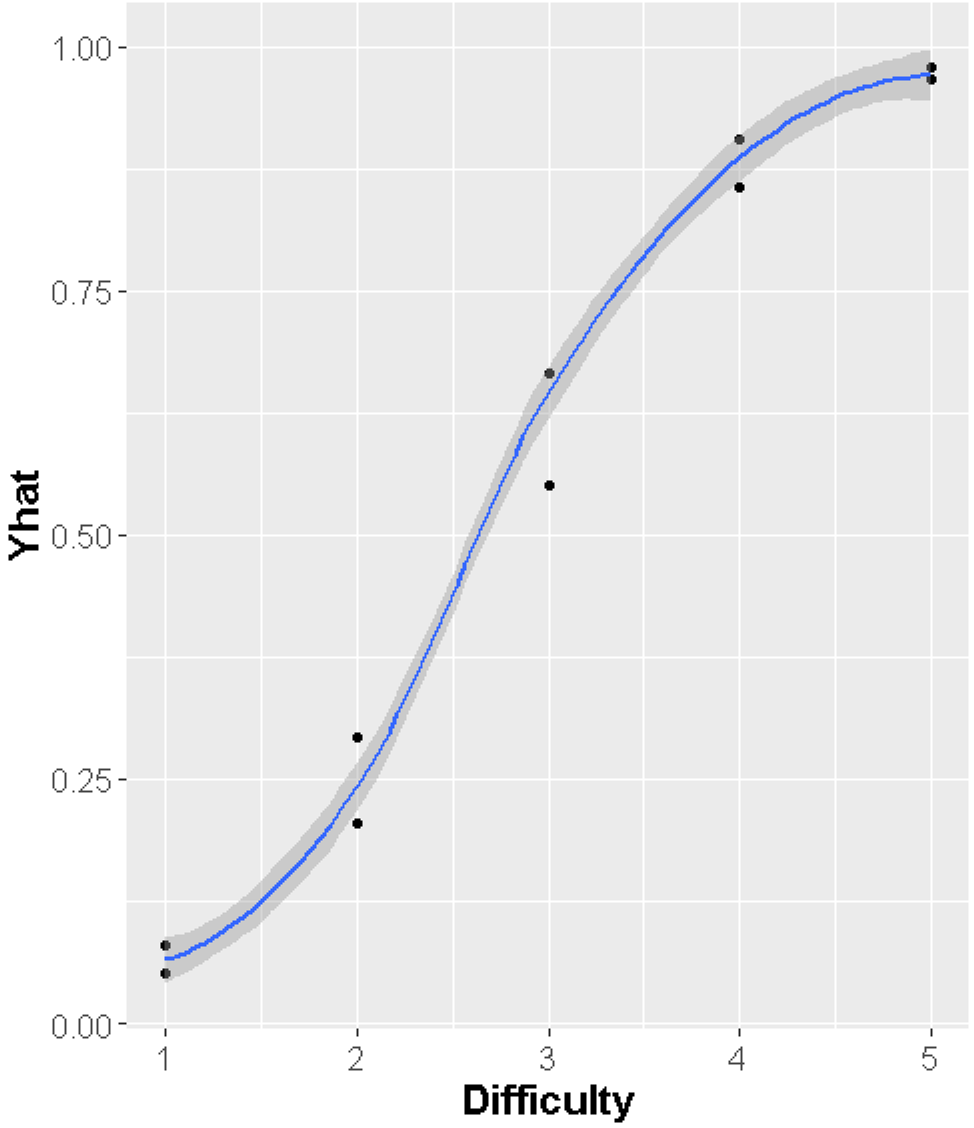
Odds Ratio Example

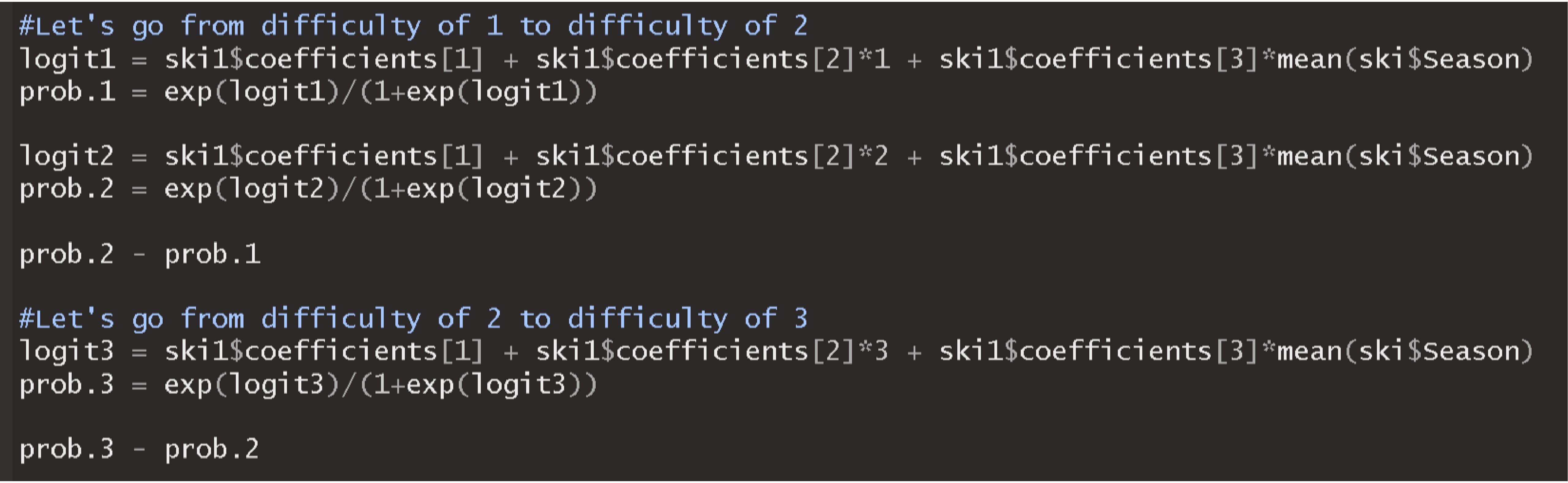
- Suppose that the outcome is hyperactivity in a child and the predictor is familial history of hyperactivity.
- Odds are 15:5 or 3:1 for hyperactivity in a child with familial history; odds are 9:150 (or 3:50) for a child without familial history. The odds ratio is just the ratio of these odds.
- Odds ratio is (15/5) / (9/150) = 50
- Thus, children with a familial history of hyperactivity are 50 times more likely to be hyperactive than those without a familial history.
- If we think of child hyperactivity as the DV, and familial history as the IV, going from no familial history (0) to familial history (1) increases the odds of hyperactivity in the child by 50 times.

		Familial History	
		yes	*no*
Child	*yes*	15	9
Hyperactivity	*no*	5	150

Probability interpretation

- We will talk more about this in our next example, but let me show you one option for linking a unit change to a change in probability. [Note, this relationship is nonlinear and so only holds at particular values of X]
- One approach for interpretability is to hold all other predictors at the mean and calculate and the change in probability if the variable of interest moves by 1.

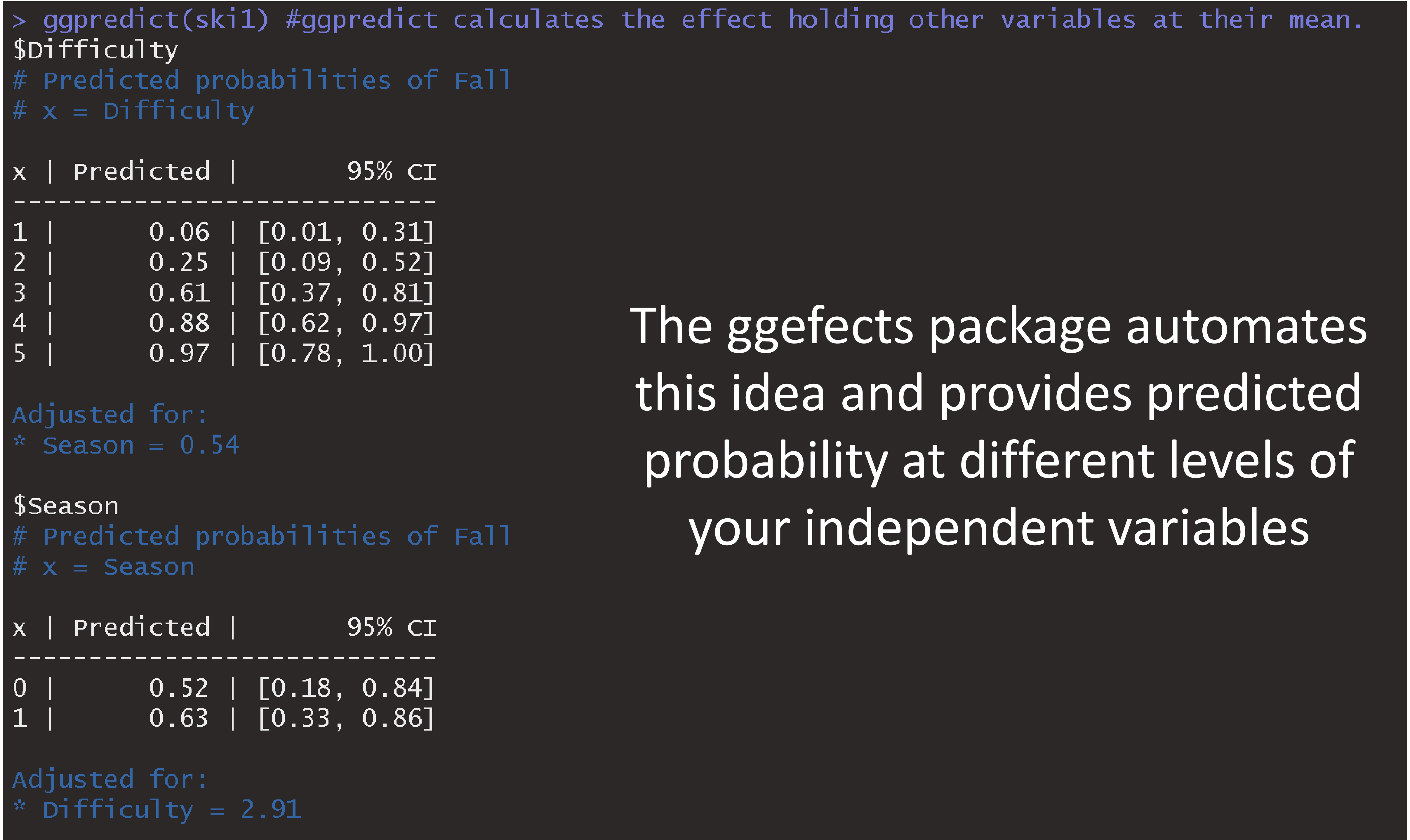




```
#Let's go from difficulty of 1 to difficulty of 2
logit1 = ski1$coefficients[1] + ski1$coefficients[2]*1 + ski1$coefficients[3]*mean(ski$Season)
prob.1 = exp(logit1)/(1+exp(logit1))

logit2 = ski1$coefficients[1] + ski1$coefficients[2]*2 + ski1$coefficients[3]*mean(ski$Season)
prob.2 = exp(logit2)/(1+exp(logit2))

prob.2 - prob.1

#Let's go from difficulty of 2 to difficulty of 3
logit3 = ski1$coefficients[1] + ski1$coefficients[2]*3 + ski1$coefficients[3]*mean(ski$Season)
prob.3 = exp(logit3)/(1+exp(logit3))

prob.3 - prob.2
```

- prob.2 – prob.1 = .184

- prob.3 – prob.2 = .365

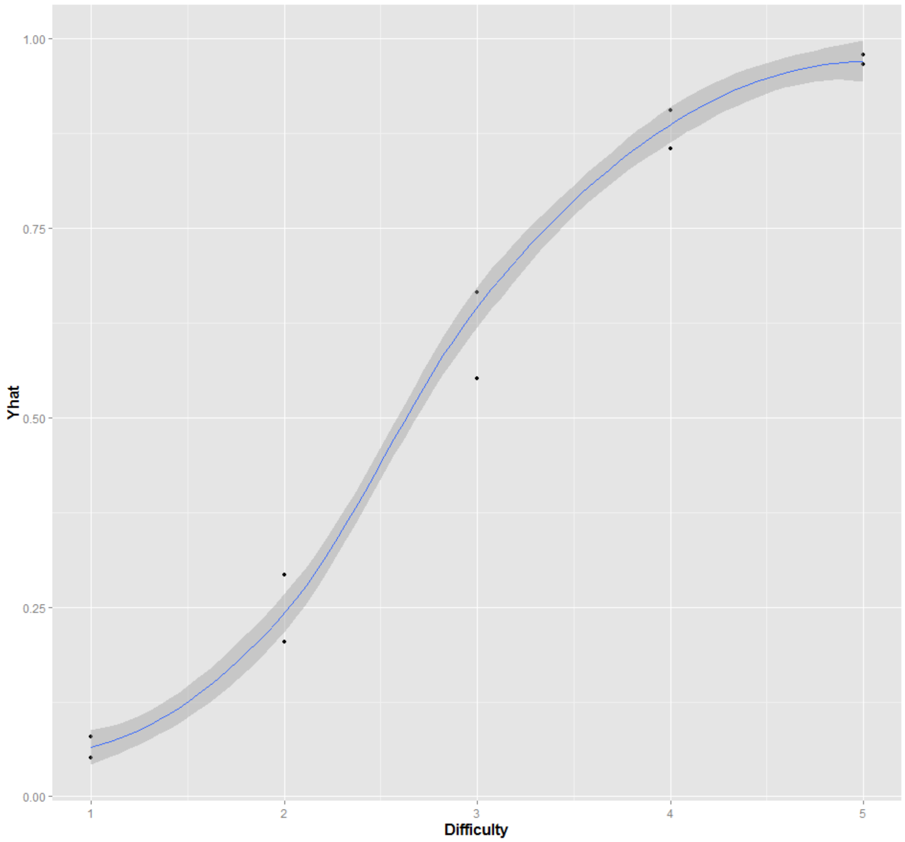


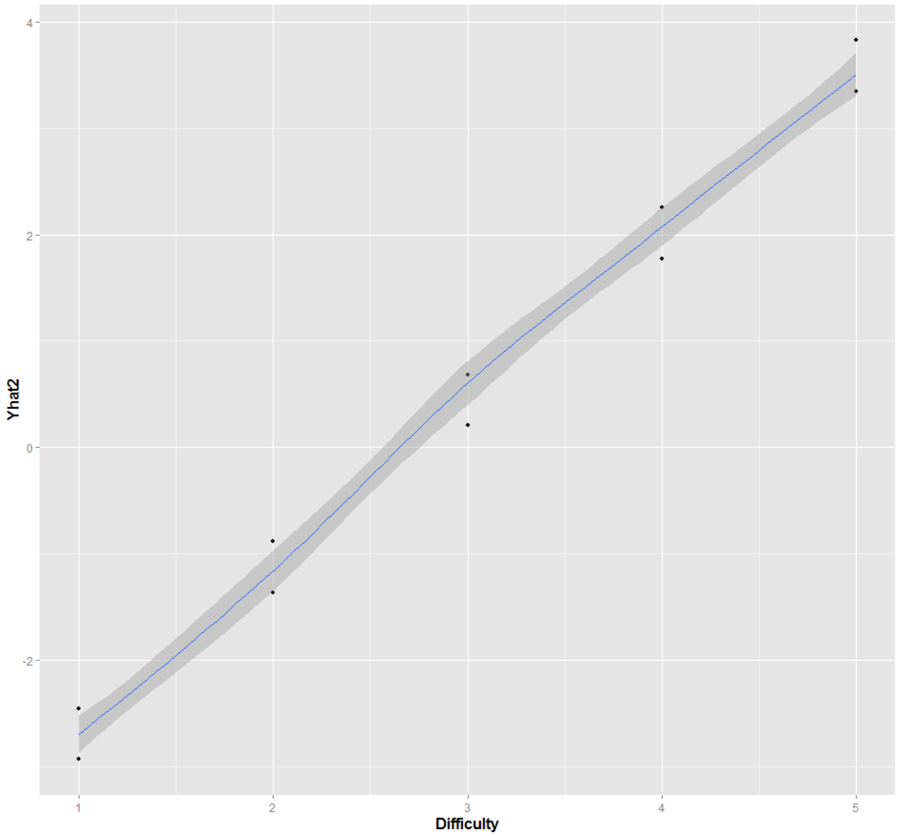
```
> ggpredict(ski1) #ggpredict calculates the effect holding other variables at their mean.
$Difficulty
# Predicted probabilities of Fall
# x = Difficulty

x | Predicted |       95% CI
----------------------------
1 |      0.06 | [0.01, 0.31]
2 |      0.25 | [0.09, 0.52]
3 |      0.61 | [0.37, 0.81]
4 |      0.88 | [0.62, 0.97]
5 |      0.97 | [0.78, 1.00]

Adjusted for:
* Season = 0.54

$Season
# Predicted probabilities of Fall
# x = Season

x | Predicted |       95% CI
----------------------------
0 |      0.52 | [0.18, 0.84]
1 |      0.63 | [0.33, 0.86]

Adjusted for:
* Difficulty = 2.91
```

The ggefects package automates this idea and provides predicted probability at different levels of your independent variables

A few visualizations

```
> #Let's plot the fitted values
> ggplot(data=ski, aes(y=Yhat, x=Difficulty)) + geom_point() + geom_smooth()
> #note if we plot the logits we get a straigh line
> Yhat2 = predict(ski1)
> ggplot(data=ski, aes(y=Yhat2, x=Difficulty)) + geom_point() + geom_smooth()
```





ANOTHER EXAMPLE OF LOGISTIC REGRESSION

Logistic Regression Example II

- Return to our loan example, where we were interested in predicting loan rejections by banks via the following model:

$$reject_i = \beta_0 + \beta_1 pubrec_i + \beta_2 black_i + \beta_3 hisp_i \ + \beta_4 loanprc_i + \varepsilon$$

- Where
  - **Reject** = 1 if loan was rejected
  - **Pubrec** = 1 if person had previously filed for bankruptcy
  - **Black** = 1 if person is black
  - **Hisp** = 1 if person is Hispanic
  - **Loanprc** = is the amount of the loan divided by the price of the house
- However, we are no longer running a linear probability model, and so we would write our logistic regression model as:

$$\ln\left(\frac{Y}{1-\hat{Y}}\right) = \beta_0 + \beta_1 pubrec_i + \beta_2 black_i + \beta_3 hisp_i \ + \beta_4 loanprc_i$$

Interpreting the Coefficients – Logged Odds

```
Call:
glm(formula = reject ~ pubrec + black + hispan + loanprc, family = binomial,
    data = loan)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2972  -0.4544  -0.4090  -0.3287   2.7762

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1480     0.3661 -11.332  < 2e-16 ***
pubrec        1.7297     0.1991   8.687  < 2e-16 ***
black         1.2444     0.1860   6.691 2.21e-11 ***
hispan        0.8436     0.2540   3.321 0.000895 ***
loanprc       2.1399     0.4375   4.892 9.99e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Logged Odds: Again, these coefficients have the exact same interpretation as in OLS regression except that the units of the DV are now in logged odds.
- Note that these Betas can be negative – but in our example all predicators are positively related with the DV, loan rejection.

Interpreting the Coefficients - Odds

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -4.1480     0.3661 -11.332  < 2e-16 ***
pubrec        1.7297     0.1991   8.687  < 2e-16 ***
black         1.2444     0.1860   6.691 2.21e-11 ***
hispan        0.8436     0.2540   3.321 0.000895 ***
loanprc       2.1399     0.4375   4.892 9.99e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #to get the odds ratio's for the estimates we can say
> exp(coef(loan2))
(Intercept)      pubrec       black      hispan     loanprc 
 0.01579567  5.63876579  3.47101232  2.32467818  8.49892805 
```

- Odds: As stated before, a coefficient of 1 leaves the odds unchanged (i.e. it has no effect). A coefficient greater than 1 increases the odds of occurrence and a coefficient less then 1 decreases the odds of occurrence. The greater the distance from one in either direction, the greater the impact of the predictor variable.

- So for pubrec, compared to an individual who never filed for bankruptcy, an individual with at least one filing has an increase in odds of loan rejection by 5.6 times.

Interpreting the Coefficients – Odds cont…

```
> #to get the odds ratio's for the estimates we can say
> exp(coef(loan2))
(Intercept)      pubrec       black      hispan     loanprc 
 0.01579567  5.63876579  3.47101232  2.32467818  8.49892805
```

- It is important to remember that the odds have a multiplicative effect. Lets assume a white person's odds of rejection based on a set of predictors is 3:1. Thus, if we took those same predictors for a black person, the odds of rejection would be 3*3.471 = 10.413:1.

- Based on this, when we divide the odds of someone who is white by someone who is black (as long as the other predictors are the same) then the result is just Exp(B). More specifically, 10.413/3 = 3.471. Thus, the coefficient shows the ratio of odds for a one unit increase in the independent variable.

- So, if you wanted to calculate the change in odds for increasing loanprc by one and going from 0 to 1 on pubrec, you need to multiply 8.499*5.639. So the odds increase by 47.9.

Interpreting the Coefficients – Odds cont…

```
> #to get the odds ratio's for the estimates we can say
> exp(coef(loan2))
(Intercept)      pubrec       black      hispan     loanprc 
 0.01579567  5.63876579  3.47101232  2.32467818  8.49892805 
```

- Let’s look at this one other way.
- Assume we have two people:
  - Person A: No public record, white, and asking for a loan of 75%.
  - Person B: Public record, white, and asking for a loan of 75%.

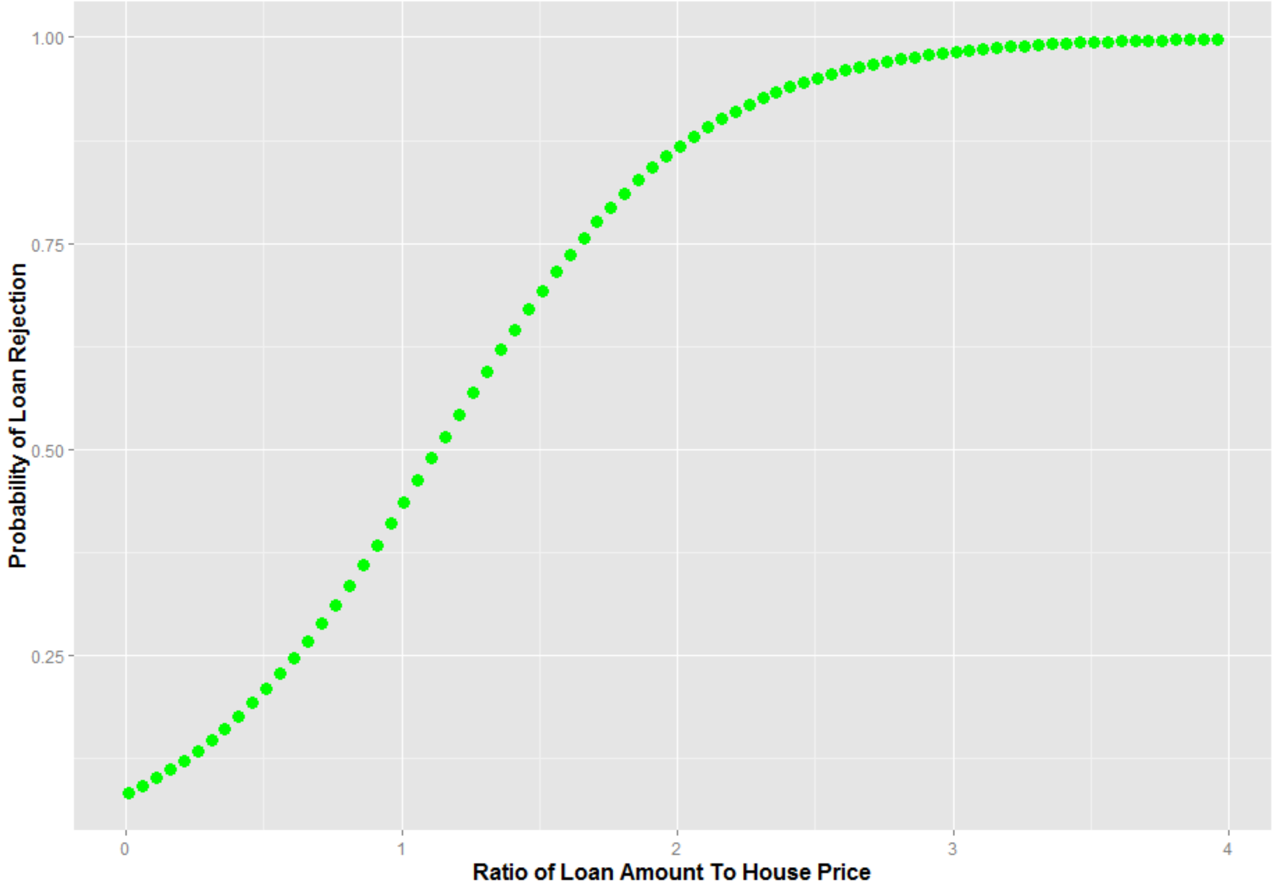
$$\frac{Odds_{x1=1,x2=0,x3=0,x4=.75}}{Odds_{x1=0,x2=0,x3=0,x4=.75}} = \frac{\exp(\beta_0 + \ \beta_1 + \ \beta_4 * .75)}{\exp(\beta_0 + \ \beta_4 * .75)} = \exp(\beta_1)$$

Probability Interpretations

- Logistic coefficients are most often interpreted in terms of odds (as we have been doing).
- However, it is possible to convert logits back to probabilities. We can calculate the predicted probability for any observation using the output.
- To do so, think back to the equations used to transform probabilities into logged odds. We now need to take the inverse to get probabilities again. [recall our discussion on GLMs]

- **Take a look at the to excel file on Blackboard!!**

A helpful tool for interpretation/ presentation of results



```
sampdat2 = expand.grid(pubrec = 1,
                       black = 0,
                       hispan=0,
                       loanprc=seq(from=.01, to=4, by=.05))

predsamp2=(predict(loan2, new=sampdat2, type="response"))

#ggplot
ggplot(data=sampdat2, aes(x=loanprc, y=predsamp2)) +
geom_point(colour="green", size=4) +
  xlab("Ratio of Loan Amount To House Price") +
  ylab("Probability of Loan Rejection") +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
```

- Create a new dataset and only vary one of the variables of interest, say loan amount. Use that dataset to produce new predicted values and then plot those predicted values against the predictor you varied.

IN CLASS EXERCISE

In class exercise

- Load the ski data from blackboard
- Run a logistic regression predicting falling based on difficulty and season. (Note, consider difficulty as a continuous variable).
- Answer the following:
  - Calculate the increase in **odds** for falling on a slope in winter of difficulty 1 versus difficulty 2.
  - Calculate the increase in **odds** for falling on a slope in winter of difficulty 1 versus difficulty 3. Calculate the increase in odds for falling in a season other than winter of difficulty 1 versus difficulty 3.
  - Calculate the **predicted probability** for falling in winter on a slope of difficulty 2 and difficulty 5.