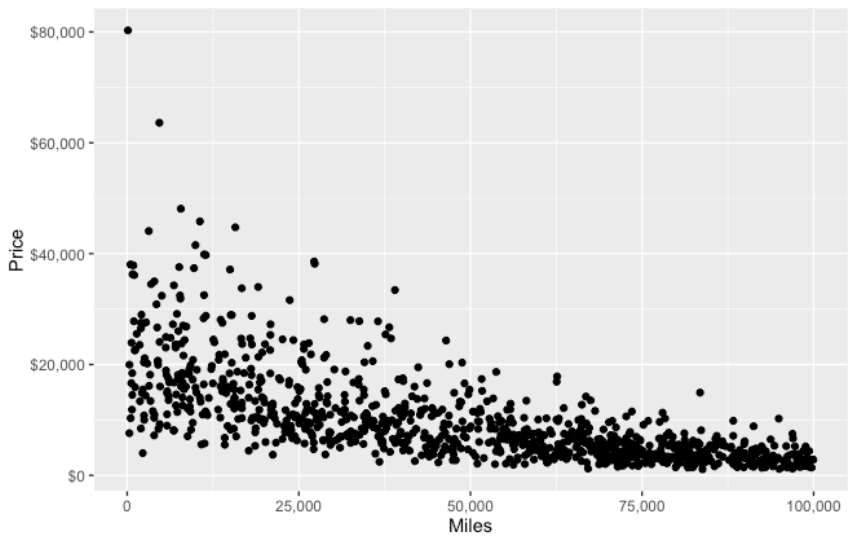
Logistic Regression

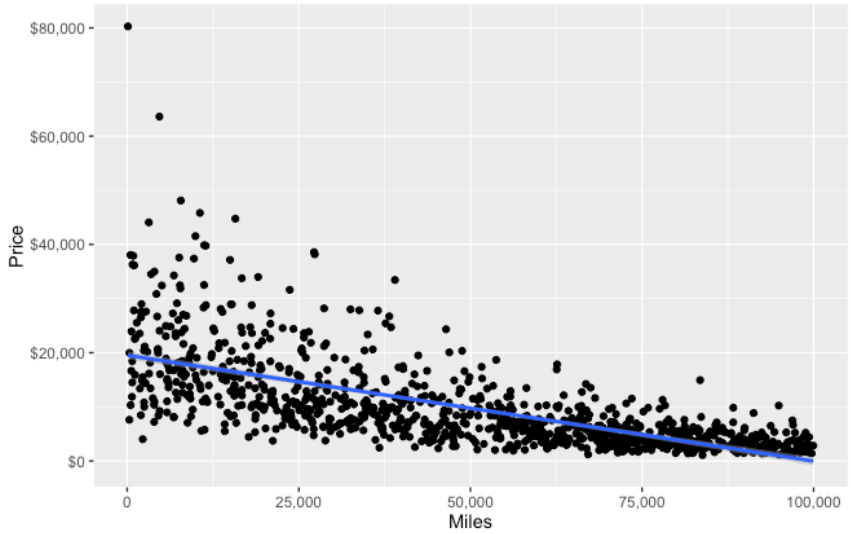
Why Logistic Regression?

- There are many important research topics for which the dependent variable is "limited."
- For example: voting, participation data is not continuous or distributed normally.
- Binary logistic regression is a type of regression analysis where the dependent variable (y) is a dummy variable: coded 0 (did not happen) or 1(did happen)

Car Prices (Revisited)



Car Prices (Cont)



Car Prices Model

- Fit a linear model to the data
  - $B_0 = 19{,}528$
  - $B_1 = -0.195$
- How much is a car with 100,000 miles worth?
  - $28
- How much is a car with 200,000 miles worth?
  - -$19,472

Transforming

- $Y = B_0 + B_1x$ - if you change x by 1, expect Y to change by B_1
- $\log(Y) = B_0 + B_1x$ - if you change x by 1, expect Y to change by $100 * B_1$ percent

Car Prices Model - Logged

- Logging our prices, we can refit our model

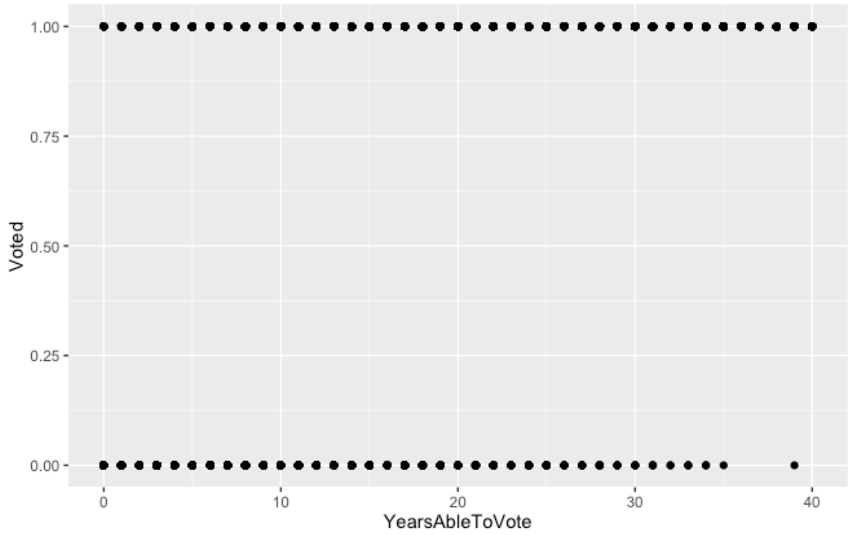
(i.e. **$\log(Y) = B_0 + B_1 * X$**) → **$Y = e^{(B0 + B1 * X)}$**

  - $B_0 = 9.839$
  - $B_1 = -0.00001929$
- How much is a car with 100,000 miles worth?
  - $2805
- How much is a car with 200,000 miles worth?
  - $419

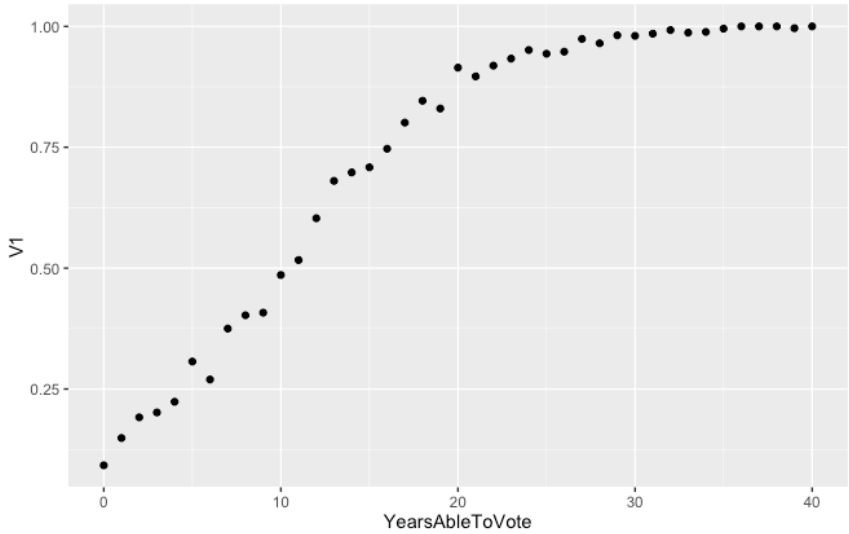
Probabilities (Odds)

- A typical case in which logistic regression helps is when our output variable is a probability (which is limited between 0 and 1).
- Transforming data can help model more accurately
- Suppose we have voting data. 1 for voted and 0 for non-voting. We also have age. How could we model the likelihood to vote given someone's age?

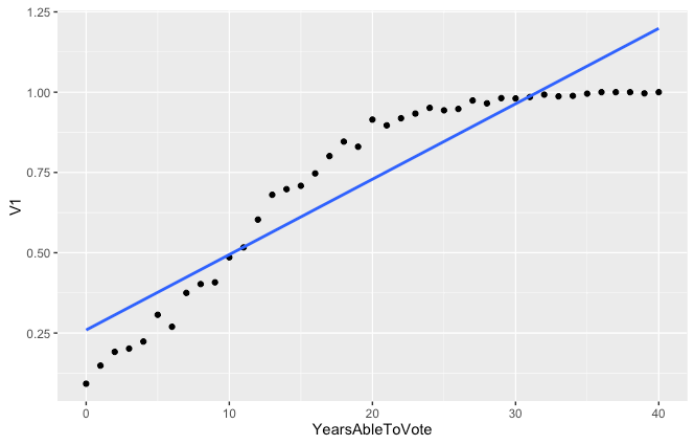
Voting by Age



Probability of voting by age



Fitting the probability with a line = ☹



Odds Ratio for the Logistic Transformation

- If there is a 75% chance that it will rain tomorrow, then 3 out of 4 times we say this it will rain. That means for every three times it rains once it will not. The odds of it raining tomorrow are 3 to 1. This can also be understood as (¾)/¼=3/1.
- If the odds that my pony will win the race is 1 to 3, that means for every 4 races it runs, it will win 1 and lose 3. Therefore I should be paid $3 for every dollar I bet.

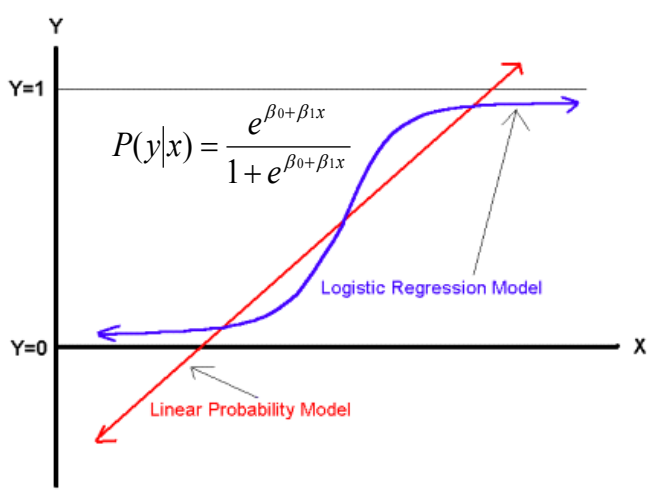
The Logistic Regression Model

The "logit" model solves these problems: we can model the relationship between the log of the odds ratio and the independent variable with a linear model

$\ln[p/(1-p)] = \beta_0 + \beta_1 X$

- p is the probability that the event Y occurs, p(Y=1)
  - [range=0 to 1]
- p/(1-p) is the "odds ratio"
  - [range=0 to ∞]
- ln[p/(1-p)]: log odds ratio, or "logit"
  - [range=-∞ to +∞]

Comparing the LP and Logit Models



Interpretation

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

$$p = 1/[1 + e^{(\beta 0 + \beta 1 X)}]$$

- if you let $\beta_0 + \beta_1 X = 0$, then p = .50
- as $\beta_0 + \beta_1 X$ gets really big, p approaches 1
- as $\beta_0 + \beta_1 X$ gets really small, p approaches 0

Interpretation

- More in general, we can use logistic regression to model relationship between a non-limited independent variable and a bounded **(limited)** dependent variable.
- It's effectively just a change of variable.