Download the notebook for this section from the CS109 repo or here: http://bit.ly/109_S6

Linear Regression

$$Y=\alpha+\beta1X1+...+\beta n+Xn+\epsilon$$

Four Assumptions of Linear Regression:

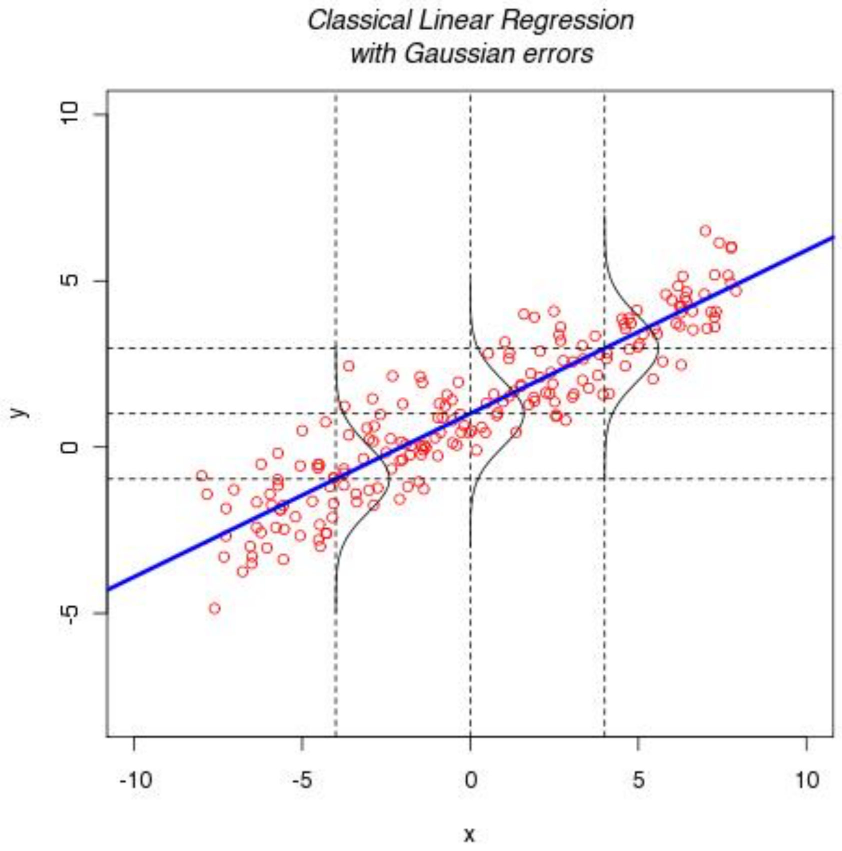
Linear Regression

$$Y=\alpha+\beta1X1+...+\beta n+Xn+\epsilon$$

Four Assumptions of Linear Regression:

1. **Linearity**: Our dependent variable Y is a linear combination of the explanatory variables X (and the error terms)
2. Observations are **independent** of one another
3. I.I.D error terms that are **Normally Distributed** ~ N(0,σ^2)
4. Design matrix X is **Full Rank**. That is:
    1. We don't have more predictors than we have observations (aka, our model is not “overdetermined”)
    2. We can’t have an exact linear relationship between two of our predictors (multicollinearity)

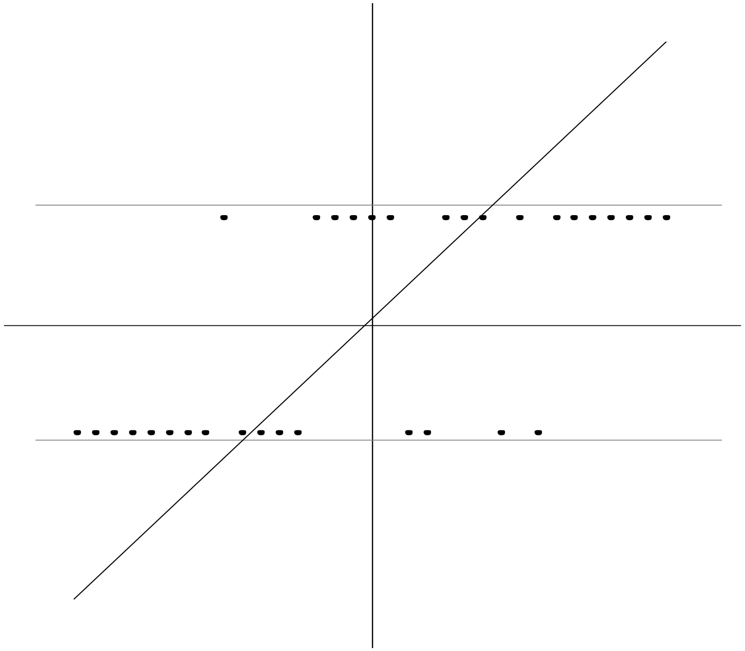
Linear Regression



Linear models presume that the **only** stochastic part of the model is the normally-distributed noise ϵ around the predicted mean.

Linear Regression

Suppose we have a binary outcome variable. Can we use Linear Regression?



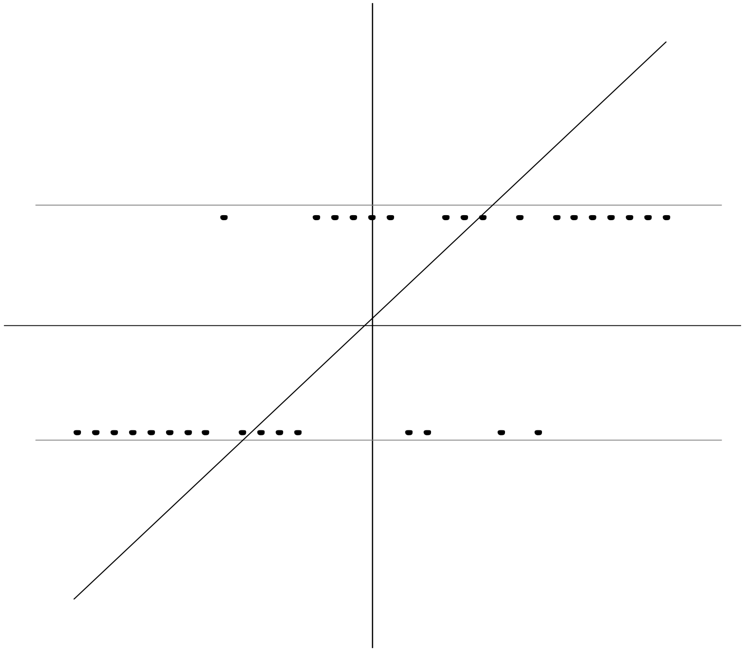
Linear Regression for binary outcomes?

If our OLS regression is of the form:

$Y = \beta_0 + \beta_1 X + \epsilon$; where Y = (0, 1)

Then we will have the following problems:

- The error terms are **heteroskedastic**
- ϵ is **not normally distributed** because Y takes on only two values
- The predicted probabilities can be greater than 1 or less than 0



More generally, just not a very useful model!

Datasets where linear regression is problematic

Linear models presume that the only stochastic part of the model is the normally-distributed noise ϵ around the predicted mean. However, there are many data sets where this is not the case such as:

- **Binary response** data where there are only two outcomes (yes/no, 0/1, etc.)
- **Categorical or Ordinal Data** of any type, where the outcome is one of a number of discrete (possibly ordered) classes
- **Count data** in which the outcome is restricted to non-negative integers
- Continuous data in which the noise is **not normally distributed**

Generalized Linear Models (GLMs), of which Logistic regression is a specific type, allow us to model and predict these types of datasets without violating the assumptions of linear regression. Logistic regression is most useful for binary response and categorical data.

Odds & Odds Ratios

Recall the definitions of an **odds**: $odds = \frac{p}{1-p}$

The odds has a range of 0 to ∞ with values greater than 1 associated with an event being more likely to occur than to not occur and values less than 1 associated with an event that is less likely to occur than not occur.

The **logit** is defined as the log of the odds:

$$\ln(odds) = \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p)$$

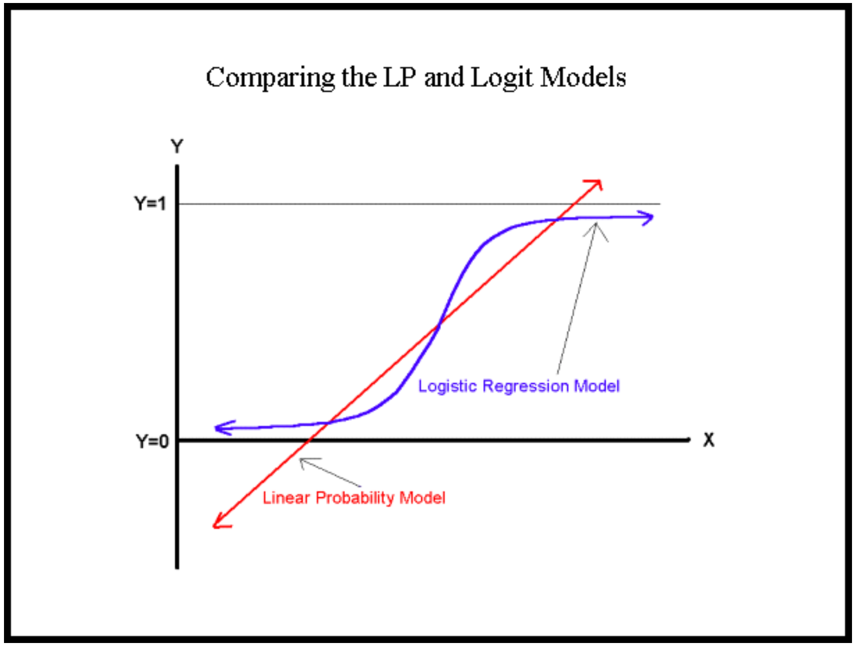
This transformation is useful because it creates a variable with a range from -∞ to +∞. Hence, this transformation solves the problem we encountered in fitting a linear model to probabilities. Because probabilities (the dependent variable) only range from 0 to 1, we can get linear predictions that are outside of this range. If we transform our probabilities to logits, then we do not have this problem because the range of the logit is not restricted. In addition, the interpretation of logits is simple—take the exponential of the logit and you have the odds for the two groups in question.

Logistic Regression

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

$$P(y|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

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


Comparing the LP and Logit Models
Y
Y=1
Y=0
X
Logistic Regression Model
Linear Probability Model