An Introduction to Logistic Regression

Emily Hector

University of Michigan

June 19, 2019

Modeling Data

- Types of outcomes
  - Continuous, binary, counts, ...
- Dependence structure of outcomes
  - Independent observations
  - Correlated observations, repeated measures
- Number of covariates, potential confounders
  - Controlling for confounders that could lead to spurious results
- Sample size

These factors will determine the appropriate statistical model to use

What is logistic regression?

- **Linear regression** is the type of regression we use for a continuous, normally distributed response variable
- **Logistic regression** is the type of regression we use for a binary response variable that follows a Bernoulli distribution

Let us review:

- Bernoulli Distribution
- Linear Regression

Review of Bernoulli Distribution

- $Y \sim Bernoulli(p)$ takes values in $\{0, 1\}$,
  - e.g. a coin toss
- $Y = 1$ for a success, $Y = 0$ for failure,
- $p =$ probability of success, i.e. $p = P(Y = 1)$,
  - e.g. $p = \frac{1}{2} = P(\text{heads})$
- Mean is p, Variance is $p(1 - p)$.

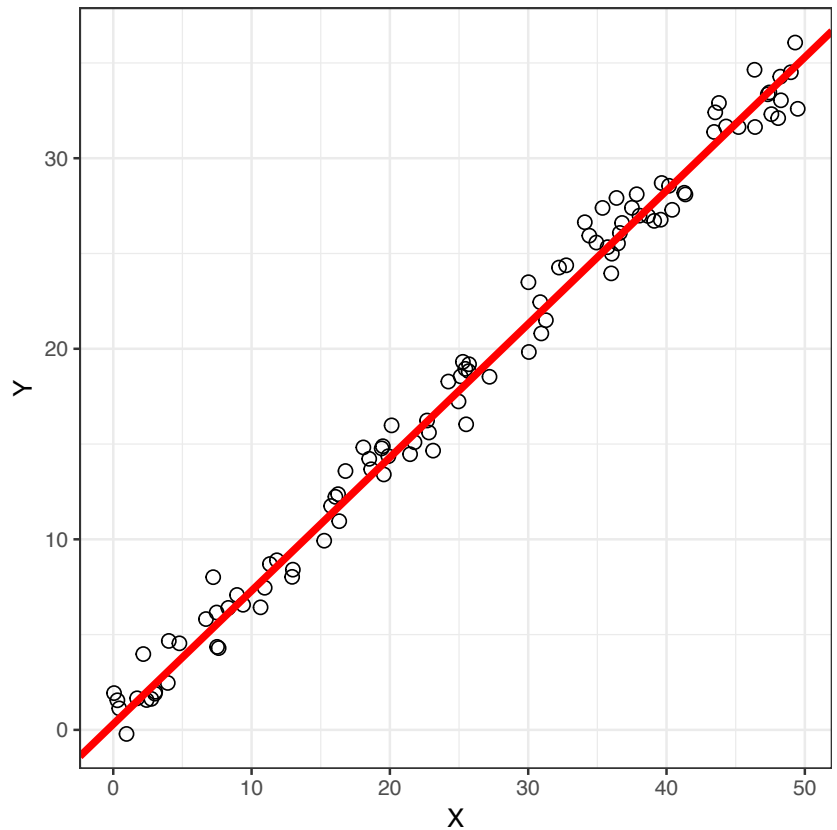
Bernoulli probability density function (pdf):

$$f(y; p) = \begin{cases} 1 - p & \text{for } y = 0 \\ p & \text{for } y = 1 \end{cases}$$
$$= p^y (1 - p)^{1-y}, \quad y \in \{0, 1\}$$

Review of Linear Regression

- When do we use linear regression?

1. Linear relationship between outcome and variable
2. Independence of outcomes
3. Constant Normally distributed errors (Homoscedasticity)



Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$,
$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
Then $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$,
$Var(Y_i) = \sigma^2$.

- How can this model break down?

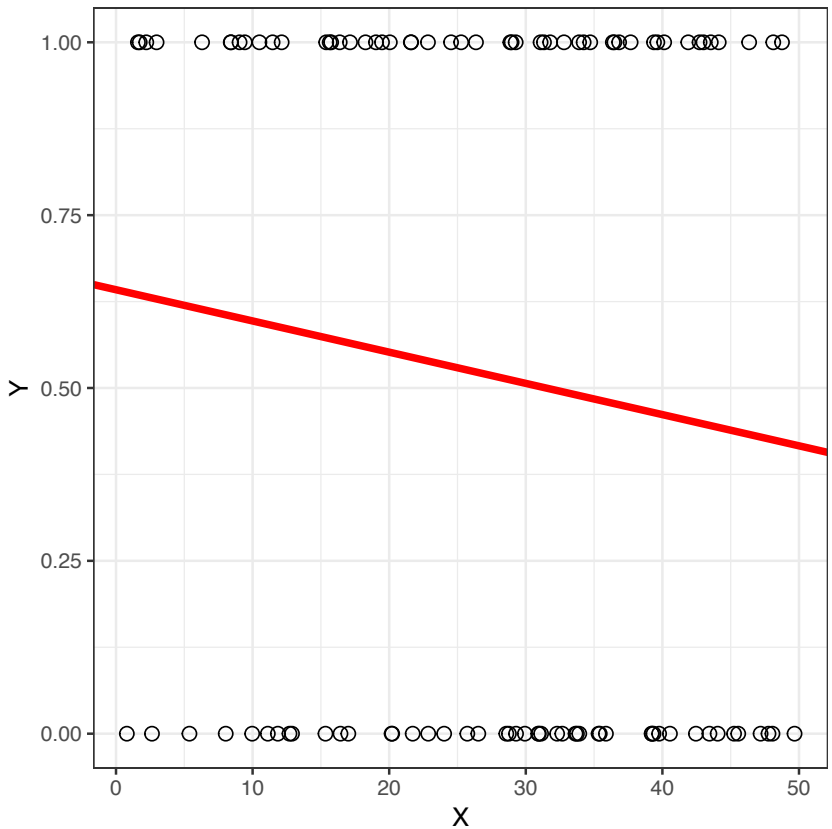
Modeling binary outcomes with linear regression

Fitting a linear regression model on a binary outcome Y:

- $Y_i|X_i \sim Bernoulli(p_{X_i})$,
- $E(Y_i) = \beta_0 + \beta_1 X_i = \widehat{p}_{X_i}$.

Problems?

- Linear relationship between X and Y?
- Normally distributed errors?
- Constant variance of Y?
- Is $\widehat{p}$ guaranteed to be in $[0, 1]$?

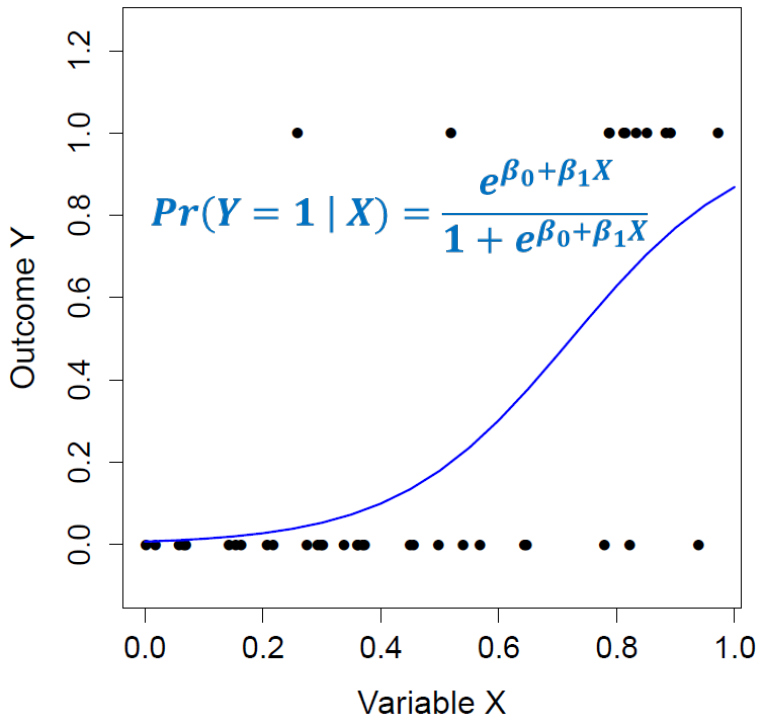


Why can't we use linear regression for binary outcomes?

- The relationship between X and Y is not linear.
- The response Y is not normally distributed.
- The variance of a Bernoulli random variable depends on its expected value p_X.
- Fitted value of Y may not be 0 or 1, since linear models produce fitted values in $(-\infty, +\infty)$

A regression model for binary data

- Instead of modeling Y, model $P(Y = 1|X)$, i.e. probability that $Y = 1$ conditional on covariates.
- Use a function that constrains probabilities between 0 and 1.



Logistic regression model

- Let Y be a binary outcome and X a covariate/predictor.
- We are interested in modeling $p_x = P(Y = 1|X = x)$, i.e. the probability of a success for the covariate value of $X = x$.

Define the **logistic regression model** as

$$logit(p_X) = \log\left(\frac{p_X}{1 - p_X}\right) = \beta_0 + \beta_1 X$$

- $\log\left(\frac{p_X}{1-p_X}\right)$ is called the **logit** function
- $p_X = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$
- $\lim_{x\to-\infty} \frac{e^x}{1+e^x} = 0$ and $\lim_{x\to\infty} \frac{e^x}{1+e^x} = 1$, so $0 \leq p_x \leq 1$.

Likelihood equations for logistic regression

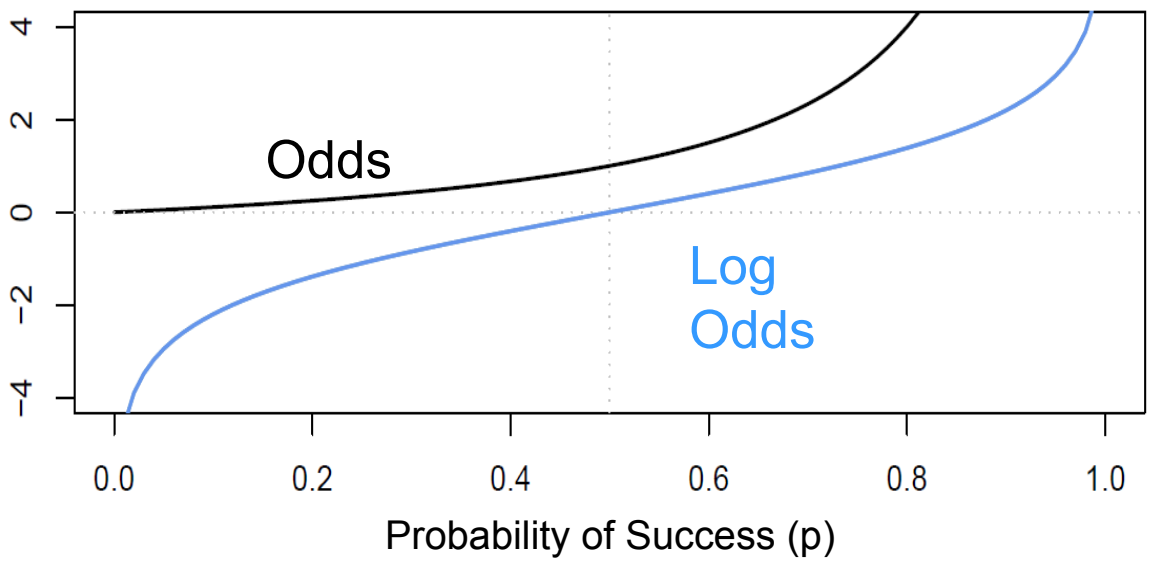
- Assume $Y_i|X_i \sim Bernoulli(p_{X_i})$ and $f(y_i|p_{x_i}) = p_{x_i}^{y_i} \times (1 - p_{x_i})^{1-y_i}$
- Binomial likelihood: $\mathcal{L}(p_x|Y, X) = \prod_{i=1}^{N} p_{x_i}^{y_i}(1 - p_{x_i})^{1-y_i}$
- Binomial log-likelihood:
$$\ell(p_x|Y, X) = \sum_{i=1}^{N} \left\{ y_i \log\left(\frac{p_{x_i}}{1-p_{x_i}}\right) + \log(1 - p_{x_i}) \right\}$$
- Logistic regression log-likelihood:
$$\ell(\beta|X, Y) = \sum_{i=1}^{N} \left\{ y_i(\beta_0 + \beta_1 x_i) - \log(1 + e^{\beta_0+\beta_1 x_i}) \right\}$$

- No closed form solution for Maximum Likelihood Estimates of β values.
- Numerical maximization techniques required.

Logistic regression terminology

Let p be the probability of success. Recall that $logit(p_X) = \log\left(\frac{p_X}{1-p_X}\right) = \beta_0 + \beta_1 X$.

- Then $\frac{p_X}{1-p_X}$ is called the **odds** of success,
- $\log\left(\frac{p_X}{1-p_X}\right)$ is called the **log odds** of success.



Another motivation for logistic regression

- Since $p \in [0, 1]$, the log odds is $\log[p/(1-p)] \in (-\infty, \infty)$.
- So while linear regression estimates anything in $(-\infty, +\infty)$,
- logistic regression estimates a proportion in $[0, 1]$.

Review of probabilities and odds

Measure	Min	Max	Name
$P(Y=1)$	0	1	"probability"
$\frac{P(Y=1)}{1-P(Y=1)}$	0	∞	"odds"
$\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right]$	$-\infty$	∞	"log-odds" or "logit"

- The odds of an event are defined as

$$\text{odds}(Y=1) = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)} = \frac{p}{1-p}$$

$$\Rightarrow p = \frac{\text{odds}(Y=1)}{1+\text{odds}(Y=1)}.$$

Review of odds ratio

		Outcome status +	Outcome status −
Exposure status	+	a	b
	−	c	d

$$\text{OR} = \frac{\text{Odds of being a case given exposed}}{\text{Odds of being a case given unexposed}}$$

$$= \frac{\frac{a}{a+b} / \frac{b}{a+b}}{\frac{c}{c+d} / \frac{d}{c+d}} = \frac{a/c}{b/d} = \frac{ad}{bc}.$$

Review of odds ratio

- Odds Ratios (OR) can be useful for comparisons.
- Suppose we have a trial to see if an intervention T reduces mortality, compared to a placebo, in patients with high cholesterol. The odds ratio is

$$OR = \frac{\text{odds(death|intervention T)}}{\text{odds(death|placebo)}}$$

- The OR describes the benefits of intervention T:
  - OR<1: the intervention is better than the placebo since odds(death|intervention T) $<$ odds(death|placebo)
  - OR$=1$: there is no difference between the intervention and the placebo
  - OR>1: the intervention is worse than the placebo since odds(death|intervention T) $>$ odds(death|placebo)

Interpretation of logistic regression parameters

$$\log\left(\frac{p_X}{1-p_X}\right) = \beta_0 + \beta_1 X$$

- β_0 is the log of the odds of success at zero values for all covariates.
- $\frac{e^{\beta_0}}{1+e^{\beta_0}}$ is the probability of success at zero values for all covariates
- Interpretation of $\frac{e^{\beta_0}}{1+e^{\beta_0}}$ depends on the sampling of the dataset
  - Population cohort: disease prevalence at $X = x$
  - Case-control: ratio of cases to controls at $X = x$

Interpretation of logistic regression parameters

Slope β_1 is the increase in the log odds ratio associated with a one-unit increase in X:

$$\begin{aligned}\beta_1 &= (\beta_0 + \beta_1(X+1)) - (\beta_0 + \beta_1 X) \\ &= \log\left(\frac{p_{X+1}}{1+p_{X+1}}\right) - \log\left(\frac{p_X}{1-p_X}\right) = \log\left\{\frac{\left(\frac{p_{X+1}}{1-p_{X+1}}\right)}{\left(\frac{p_X}{1-p_X}\right)}\right\}\end{aligned}$$

and e^{β_1}=OR!.

- If $\beta_1 = 0$, there is no association between changes in X and changes in success probability (OR= 1).
- If $\beta_1 > 0$, there is a positive association between X and p (OR> 1).
- If $\beta_1 < 0$, there is a negative association between X and p (OR< 1).

Interpretation of slope β_1 is the same regardless of sampling.

Interpretation odds ratios in logistic regression

- OR>1: positive relationship: as X increases, the probability of Y increases; exposure $(X = 1)$ associated with higher odds of outcome.
- OR<1: negative relationship: as X increases, probability of Y decreases; exposure $(X = 1)$ associated with lower odds of outcome.
- OR$=1$: no association; exposure $(X = 1)$ does not affect odds of outcome.

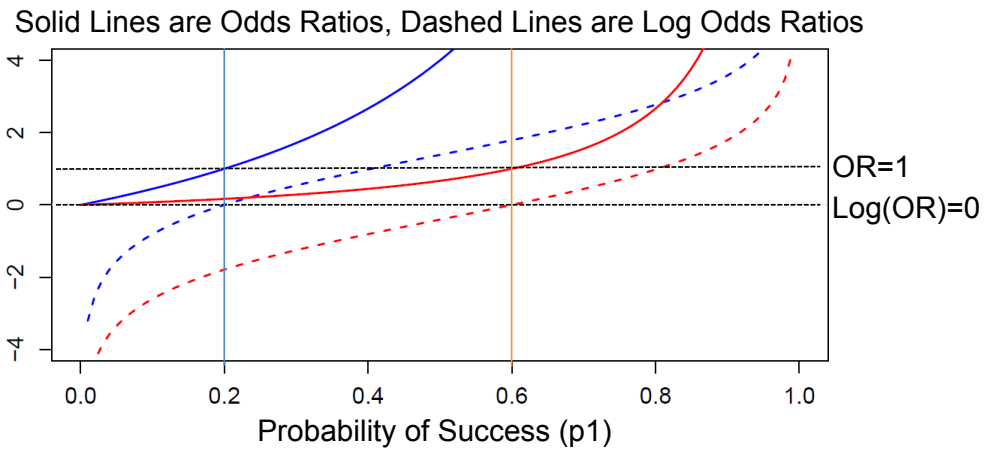
In logistic regression, we test null hypotheses of the form $H_0 : \beta_1 = 0$ which corresponds to OR$=1$.

Logistic regression terminology

- OR is the ratio of the odds for difference success probabilities:

$$\frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{p_2}{1-p_2}\right)}$$

- OR= 1 when $p_1 = p_2$.
- Interpretation of odds ratios is difficult!

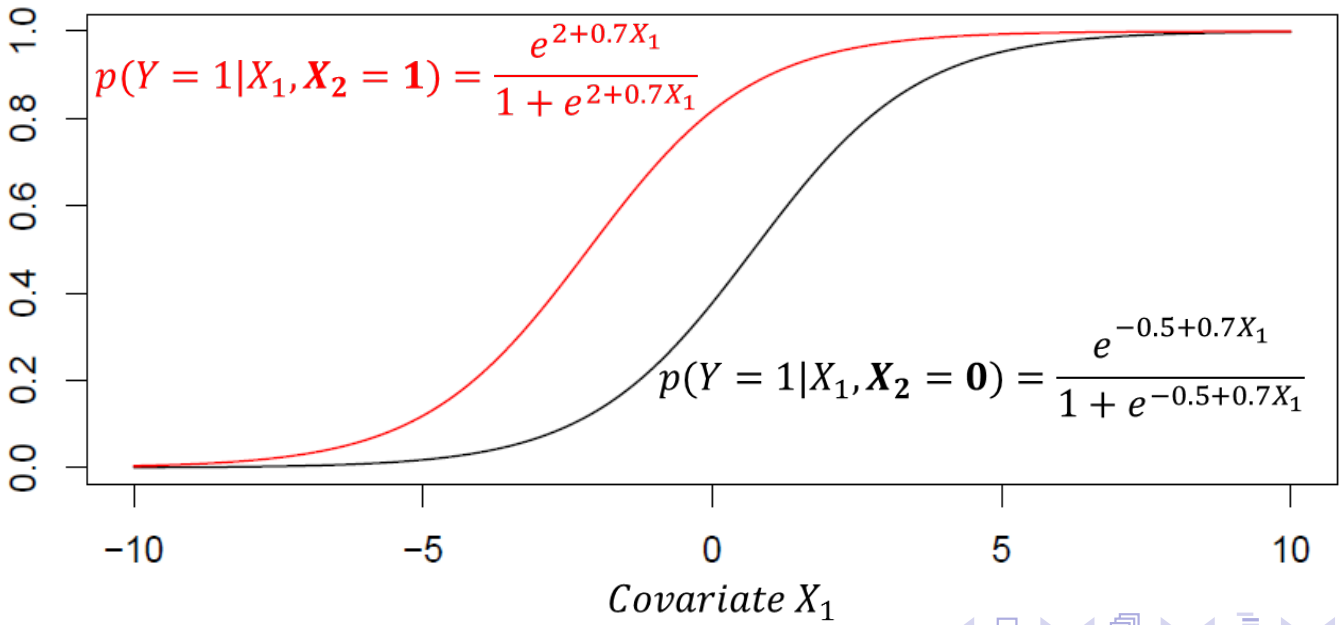


Multiple logistic regression

Consider a multiple logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Let X_1 be a continuous variable, X_2 an indicator variable (e.g. treatment or group).
- Set $\beta_0 = -0.5$, $\beta_1 = 0.7$, $\beta_2 = 2.5$.



Data example: CHD events

Data from Western Collaborative Group Study (WCGS).
For this example, we are interested in the outcome

$$Y = \begin{cases} 1 & \text{if develops CHD} \\ 0 & \text{if no CHD} \end{cases}$$

1. How likely is a person to develop coronary heart disease (CHD)?
2. Is hypertension associated with CHD events?
3. Is age associated with CHD events?
4. Does weight confound the association between hypertension and CHD events?
5. Is there a differential effect of CHD events for those with and without hypertension depending on weight?

How likely is a person to develop CHD?

- The WCGS was a prospective cohort study of 3524 men aged $39 - 59$ and employed in the San Francisco Bay or Los Angeles areas enrolled in 1960 and 1961.
- Follow-up for CHD incidence was terminated in 1969.
- 3154 men were CHD free at baseline.
- 275 men developed CHD during the study.
- The estimated probability a person in WCGS develops CHD is $257/3154 = 8.1\%$.
- This is an unadjusted estimate that does not account for other risk factors.
- How do we use logistic regression to determine factors that increase risk for CHD?

Getting ready to use R

Make sure you have the package **epitools** installed.

```
# install.packages("epitools")
library(epitools)
data(wcgs)
```

```
## Can get information on the dataset:
str(wcgs)
```

```
## Define hypertension as systolic BP > 140 or diastolic BP > 80:
wcgs$HT <- as.numeric(wcgs$sbp0>140 | wcgs$dbp0>90)
```

Is hypertension associated with CHD events?

The OR can be obtained from the 2x2 table:

```
table_2by2 <- data.frame(
  Hypertensive=c("No","Yes"),
  "No CHD event"=c(sum(wcgs$chd69==0 & wcgs$HT==0),
                   sum(wcgs$chd69==0 & wcgs$HT==1)),
  "CHD event"=c(sum(wcgs$chd69==1 & wcgs$HT==0),
                sum(wcgs$chd69==1 & wcgs$HT==1)),
  check.names=FALSE)
```

Hypertensive	No CHD event	CHD event
No	2312	173
Yes	585	84

$$\text{OR} = (2312 \times 84)/(585 \times 173) = 1.92.$$

The OR can also be obtained from the logistic regression model:

$$logit\left[P(\text{CHD})\right] = \log\left[\frac{P(\text{CHD})}{1-P(\text{CHD})}\right] = \beta_0 + \beta_1 \times hypertension.$$

```
logit_HT <- glm(chd69 ~ HT, data = wcgs, family = "binomial")
coefficients(summary(logit_HT))
```

```
##                Estimate Std. Error    z value      Pr(>|z|)
## (Intercept) -2.5925766 0.07882162 -32.891693 2.889272e-237
## HT           0.6517816 0.14080842   4.628854  3.676954e-06
```

OR from logistic regression is the same as the 2x2 table!

$$\exp(\beta_1) = \exp\left(0.6517816\right) = 1.92$$

- The effect of HT is significant ($p = 3.68 \times 10^{-6}$)
- The odds of developing CHD is 1.92 times higher in hypertensives than non-hypertensives; 95% C.I. (1.46, 2.53)

Is age associated with CHD events?

$$logit\left[P(\text{CHD})\right] = \log\left[\frac{P(\text{CHD})}{1-P(\text{CHD})}\right] = \beta_0 + \beta_1 \times age.$$

```
logit_age <- glm(chd69 ~ age0, data = wcgs, family = "binomial")
coefficients(summary(logit_age))

##                Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -5.93951594 0.54931839 -10.812520 3.003058e-27
## age0         0.07442256 0.01130234   6.584705 4.557900e-11
```

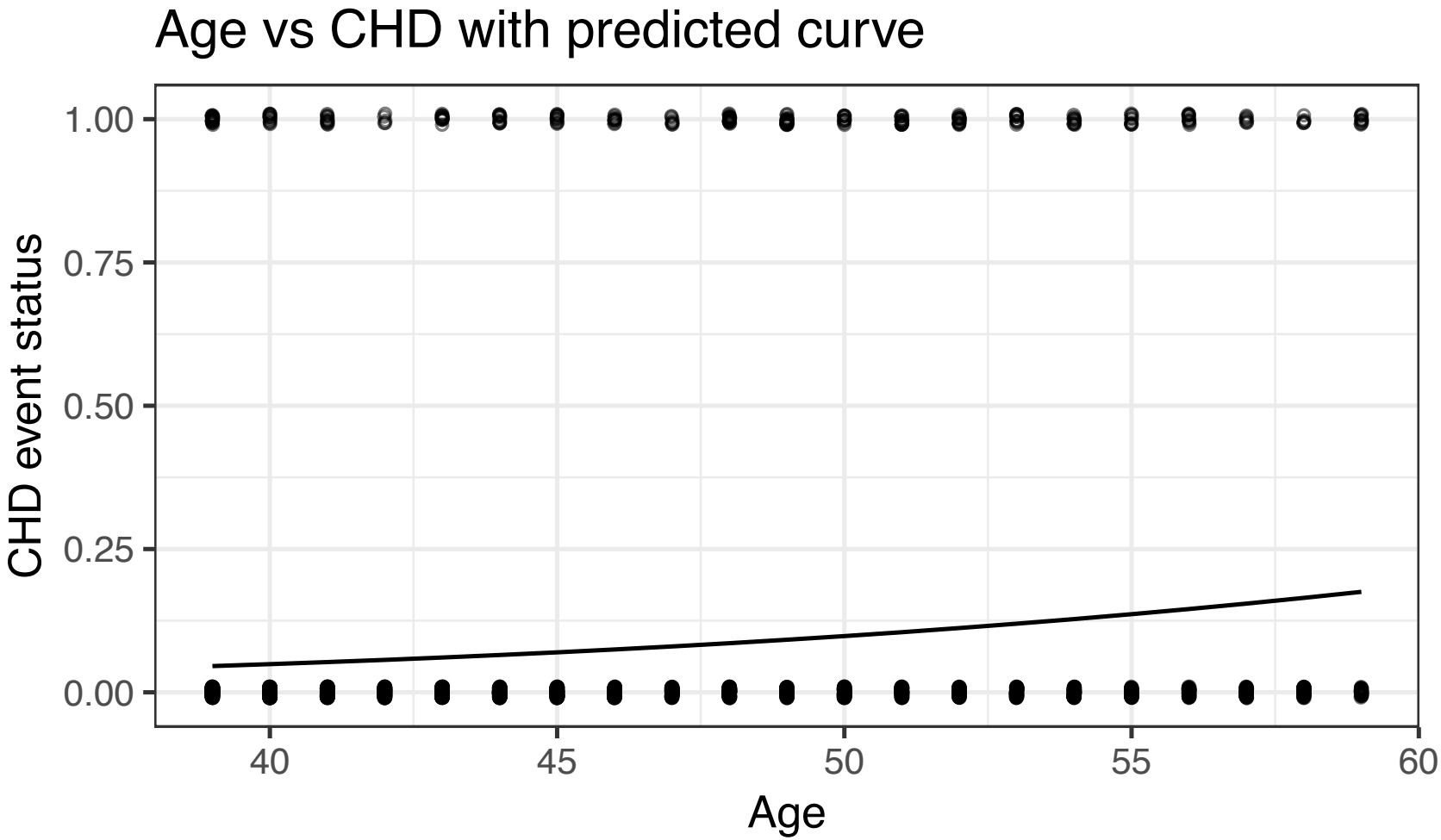
- Yes, CHD risk is significantly associated with increased age ($p = 4.56 \times 10^{-11}$)
- The OR $= \exp(0.0744) = 1.08$; 95% C.I. (1.05, 1.1).
- For a 1-year increase in age, the log odds of a CHD event increases by 7.4%, or the odds of a CHD event increase by 1.08.

What does the logistic model for age look like?

$$logit(\text{CHD}) = -5.94 + 0.07 \times age$$

$$P(\text{CHD}) = \frac{\exp\left[-5.94 + 0.07 \times age\right]}{1 + \exp\left[-5.94 + 0.07 \times age\right]}$$

```
library(ggplot2)
wcgs$pred_age <-
  predict(logit_age, data.frame(age0=wcgs$age0), type="resp")
ggplot(wcgs, aes(age0, chd69)) +
  geom_point(position=position_jitter(h=0.01, w=0.01),
             shape = 21, alpha = 0.5, size = 1) +
  geom_line(aes(y = pred_age)) +
  ggtitle("Age vs CHD with predicted curve") + xlab("Age") +
  ylab("CHD event status") + theme_bw()
```


Age vs CHD with predicted curve
CHD event status
1.00
0.75
0.50
0.25
0.00
40
45
50
55
60
Age

Does weight confound the association between hypertension and CHD events?

Recall that the OR for HT was 1.92 (the β value was 0.6518). Fit the model $logit(\text{CHD}) = \beta_0 + \beta_1 HT + \beta_2 weight$.

```
logit_weight <- glm(chd69 ~ HT + weight0, data = wcgs,
                    family = "binomial")
coefficients(summary(logit_weight))

##                  Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -3.928507302 0.51403008 -7.642563 2.129397e-14
## HT           0.568375813 0.14480630  3.925077 8.670213e-05
## weight0      0.007898806 0.00297963  2.650935 8.026933e-03
```

Look at the change in coefficient for HT between the unadjusted and adjusted models:

- $(0.6518 - 0.5684)/0.6518 = 12.8\%$.
- Since the change in effect size is $> 10\%$, we would consider weight a confounder.

Is there a differential effect of weight on CHD for those with and without HT?

In other words, is there an interaction between weight and hypertension?
Fit the model

$$logit[P(\text{CHD})] = \beta_0 + \beta_1 HT + \beta_2 weight + \beta_3 (HT \times weight).$$

```
logit_HTweight <- glm(chd69 ~ HT + weight0 + HT:weight0, data = wcgs,
                      family = "binomial")
coefficients(summary(logit_HTweight))
```

```
##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -4.82255032 0.671632476 -7.180341 6.953768e-13
## HT           2.82407466 1.096531902  2.575461 1.001067e-02
## weight0      0.01311598 0.003871862  3.387512 7.052961e-04
## HT:weight0  -0.01279195 0.006184812 -2.068285 3.861323e-02
```

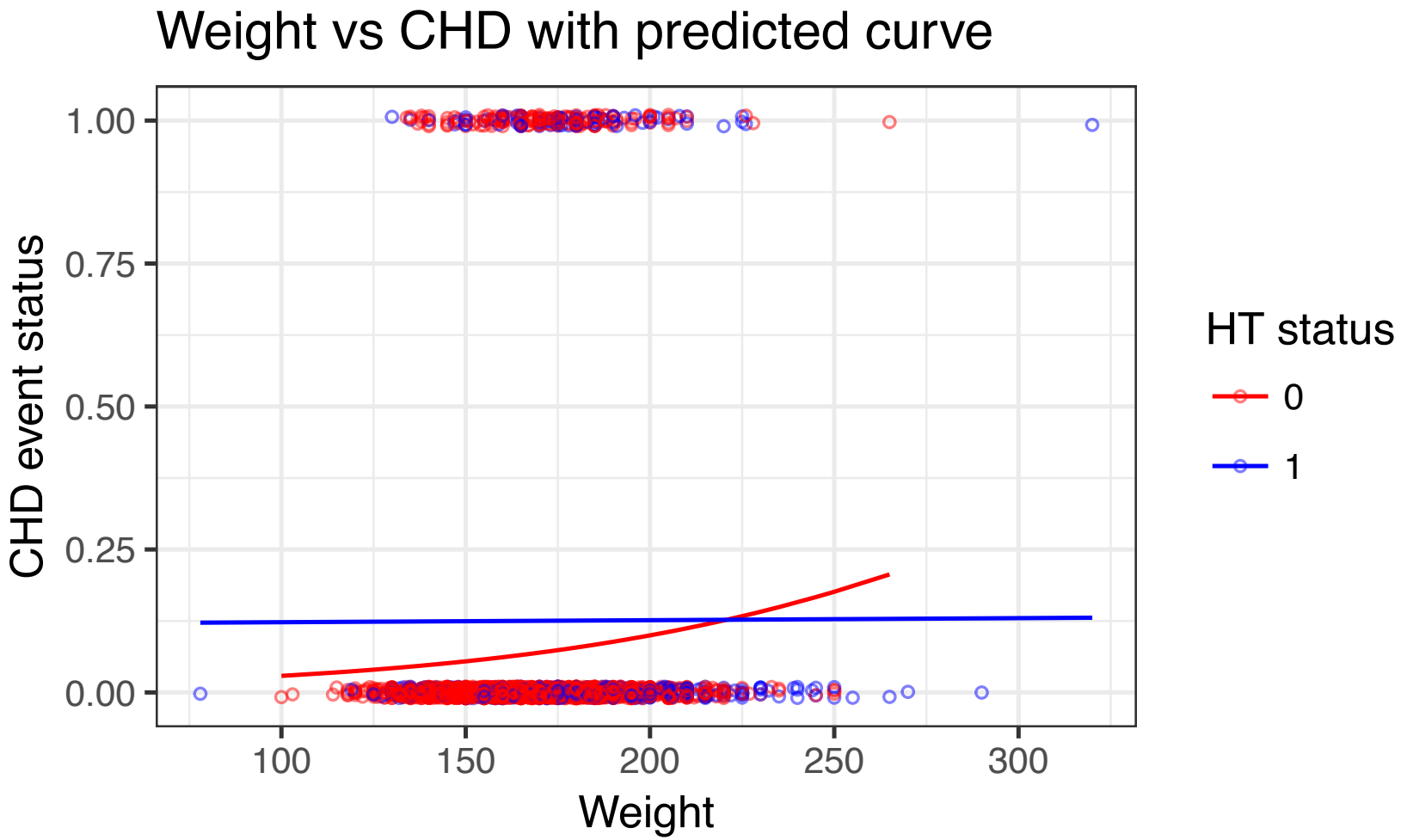
Interaction model interpretation

- The interaction effect is significant ($p = 0.0386$).
- Odds ratio for 1lb. increase in weight for those without hypertension: $\exp(0.013116) = 1.01$.
- Odds ratio for 1lb. increase in weight for those with hypertension: $\exp(0.013116 - 0.012792) \approx 1$.

Plot of interaction model:

```
wcgs$pred_interaction <-
  predict(logit_HTweight, data.frame(weight0=wcgs$weight0, HT=wcgs$HT),
          type="resp")
ggplot(wcgs, aes(weight0, chd69, color=as.factor(HT))) +
  geom_point(position=position_jitter(h=0.01, w=0.01),
             shape = 21, alpha = 0.5, size = 1) +
  geom_line(aes(y = pred_interaction, group=HT)) +
  scale_colour_manual(name="HT status", values=c("red","blue")) +
  ggtitle("Weight vs CHD with predicted curve") + xlab("Weight") +
  ylab("CHD event status") + theme_bw()
```

Plot of interaction model



Plot of interaction model – interpretation

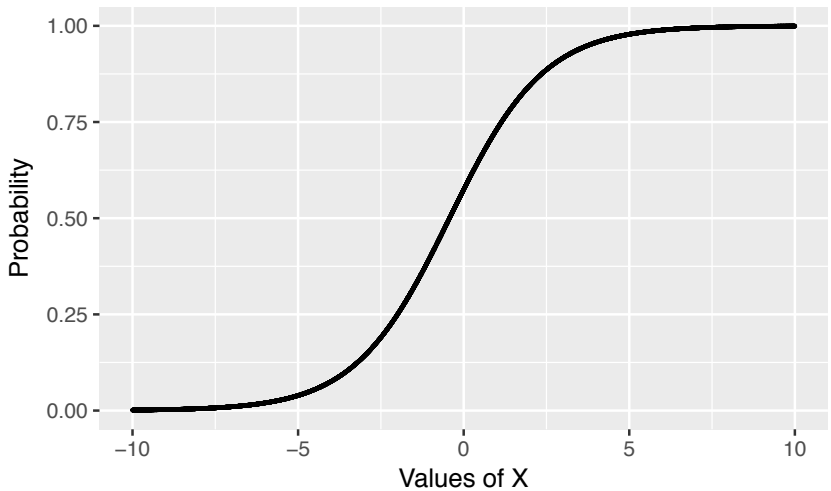
```
##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -4.82255032 0.671632476 -7.180341 6.953768e-13
## HT           2.82407466 1.096531902  2.575461 1.001067e-02
## weight0      0.01311598 0.003871862  3.387512 7.052961e-04
## HT:weight0  -0.01279195 0.006184812 -2.068285 3.861323e-02
```

- The effect of increasing weight on CHD risk is different between those with and without hypertension.
- For those without hypertension, increase in weight leads to an increase in CHD risk.
- For those with hypertension, the risk of CHD is nearly constant with respect to weight.

Predicted probabilities

- Fit model and obtain the estimated coefficients.
- Calculate predicted probability $\widehat{p}$ for each person depending on their characteristics X:

$$\widehat{p} = \frac{\exp\left(\widehat{\beta_0} + \widehat{\beta_1} X\right)}{1 + \exp\left(\widehat{\beta_0} + \widehat{\beta_1} X\right)}$$



Predicted probability of CHD by weight

The model is $logit[P(\text{CHD})] = \beta_0 + \beta_1 \times weight$.

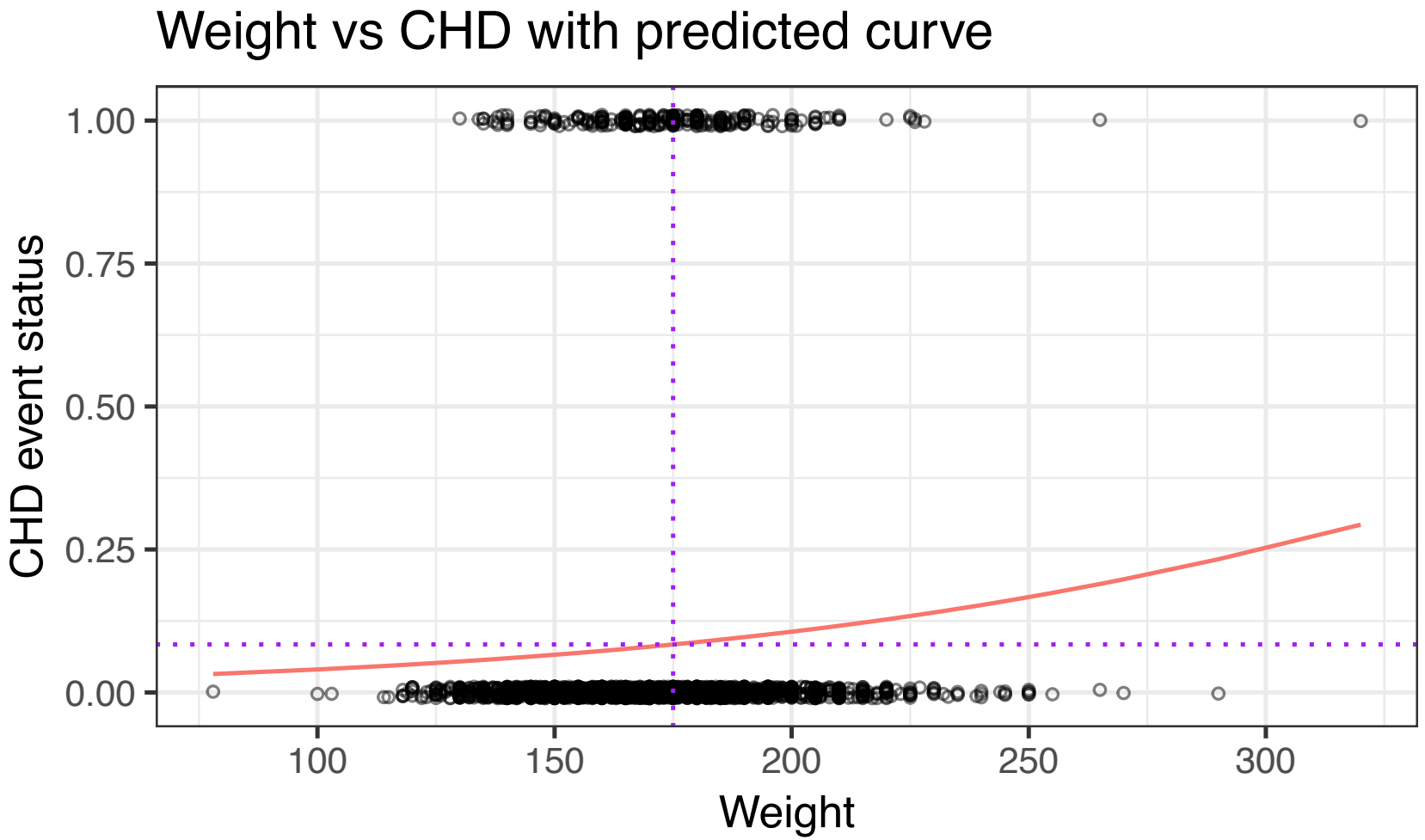
```
logit_weight_noHT <- glm(chd69 ~ weight0, data = wcgs,
                         family = "binomial")
coefficients(summary(logit_weight_noHT))

##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -4.21470593 0.51206319 -8.230832 1.859181e-16
## weight0      0.01042419 0.00291957  3.570455 3.563615e-04
```

Based on the model, the predicted probability for a person weighing 175 lbs is

$$P(\text{CHD}|175\text{lbs}) = \frac{\exp(-4.2147059 + 0.0104242 \times 175)}{1 + \exp(-4.2147059 + 0.0104242 \times 175)}$$
$$= 0.0839 \text{ or } 8.4\%.$$

Plot of predicted probability of CHD by weight



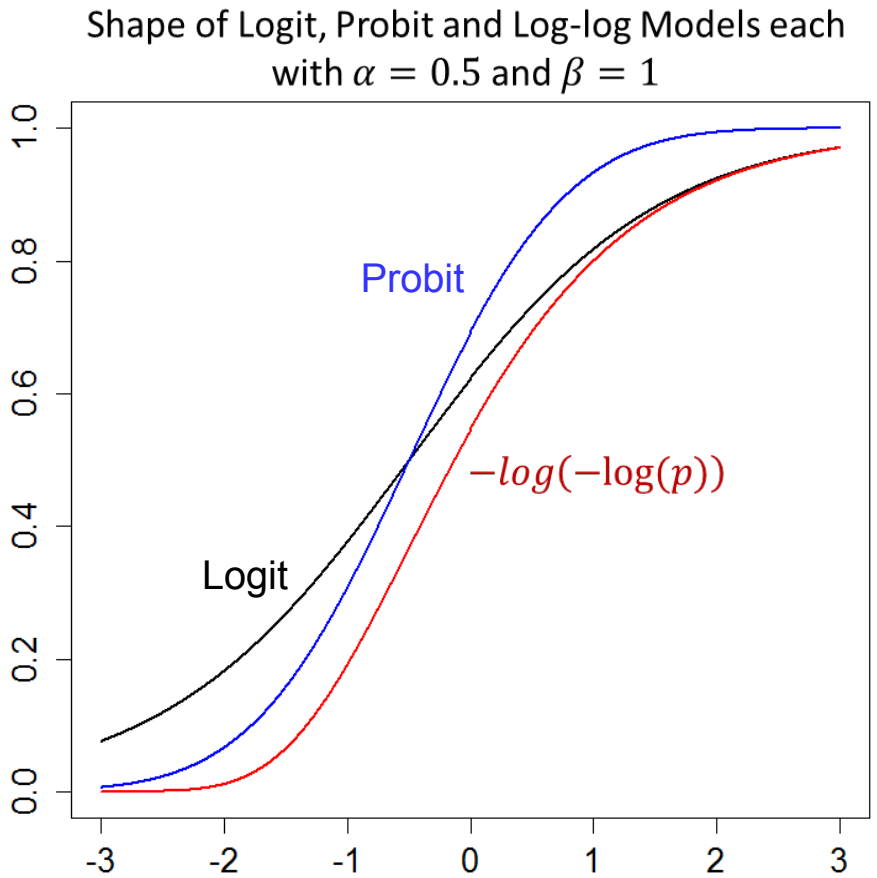
Alternative models for binary outcomes

The logit function induces a specific shape for the relationship between the covariate X and the probability of success $p = P(Y = 1|X)$.

Logit: $log[p/(1-p)] = \alpha + \beta X$.

Probit: $\Phi^{-1}(p) = \alpha + \beta X$ where Φ is the Normal CDF.

Log-log: $-\log[\log(p)] = \alpha + \beta X$.



Summary

- Logistic regression models the log of the odds of an outcome.
  - Used when the outcome is binary.
- We interpret odds ratios (exponentiated coefficients) from logistic regression.
- We can control for confounding factors and assess interactions in logistic regression.
- Many of the concepts that apply to multiple linear regression continue to apply in logistic regression.