Lecture 12
Logistic regression

BIOST 515

February 17, 2004

Outline

- Review of simple logistic model
- Further motivation for logistic regression (why is it so popular?)
- Extending the logistic model (multiple predictors)
- Estimation
- Testing
- Model checking

Review of logistic regression

In logistic regression, we model the log-odds,

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi},$$

where

- $\pi_i = E[y_i]$ and
- y_i is a binary outcome.

So far, we've only looked at the simple case,

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i.$$

We showed that the odds ratio for a unit increase in x is

$$OR = \exp(\beta_1),$$

and the predicted probability that $y_i = 1$ is

$$\hat{\pi}_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

Example

Of 2332 patients who underwent cardiac catheterization at Duke University Medical Center, 1129 were found to have significant diameter narrowing of at least one major coronary artery. In this subset of patients, investigators were interested in knowing whether the time from the onset of symptoms of coronary artery disease was related to the probability that the patient has severe disease.

We can assess this using logistic regression fitting the following model,

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 cad.dur_i,$$

where $\pi_i = Pr(i^{th}$ patient has severe disease$|cad.dur_i)$ and $cad.dur_i$ is the time from the onset of symptoms.

Fitting this model in R, we get the following results

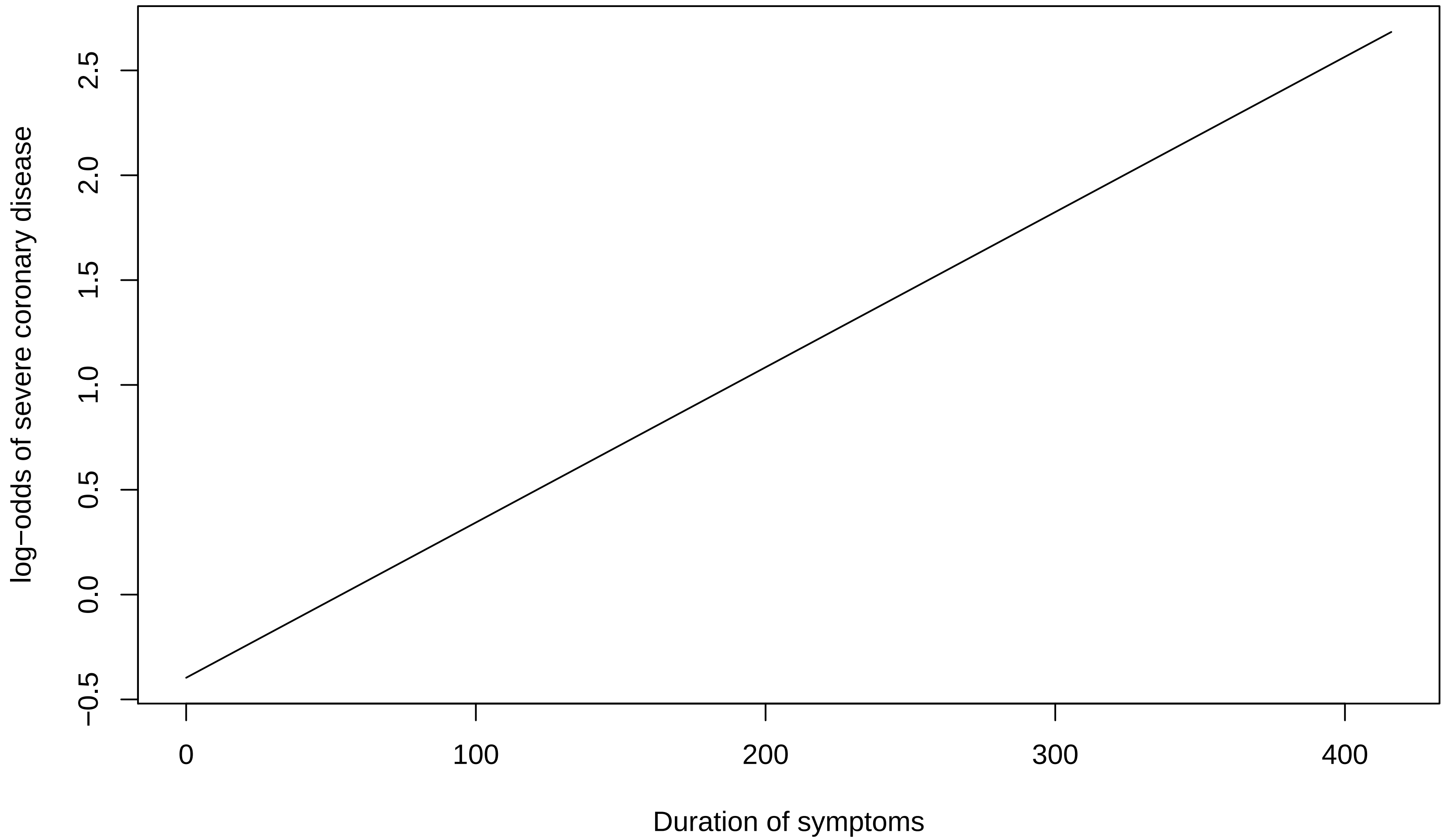
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.3966	0.0542	−7.32	0.0000
cad.dur	0.0074	0.0008	9.31	0.0000

The fitted model is

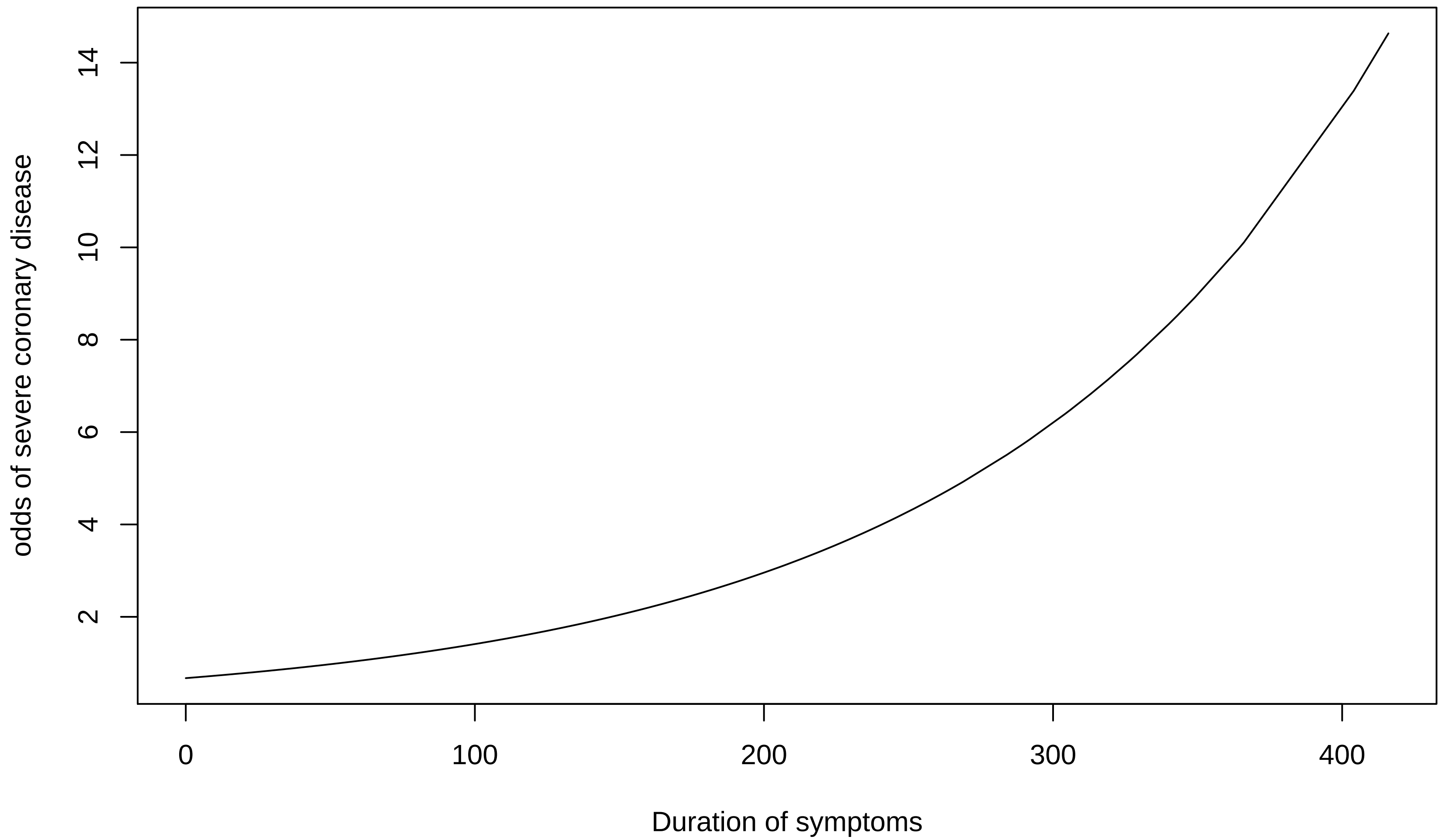
$$\text{logit}(\hat{\pi}_i) = -0.3966 + 0.0074 cad.dur_i.$$

How do we interpret this?

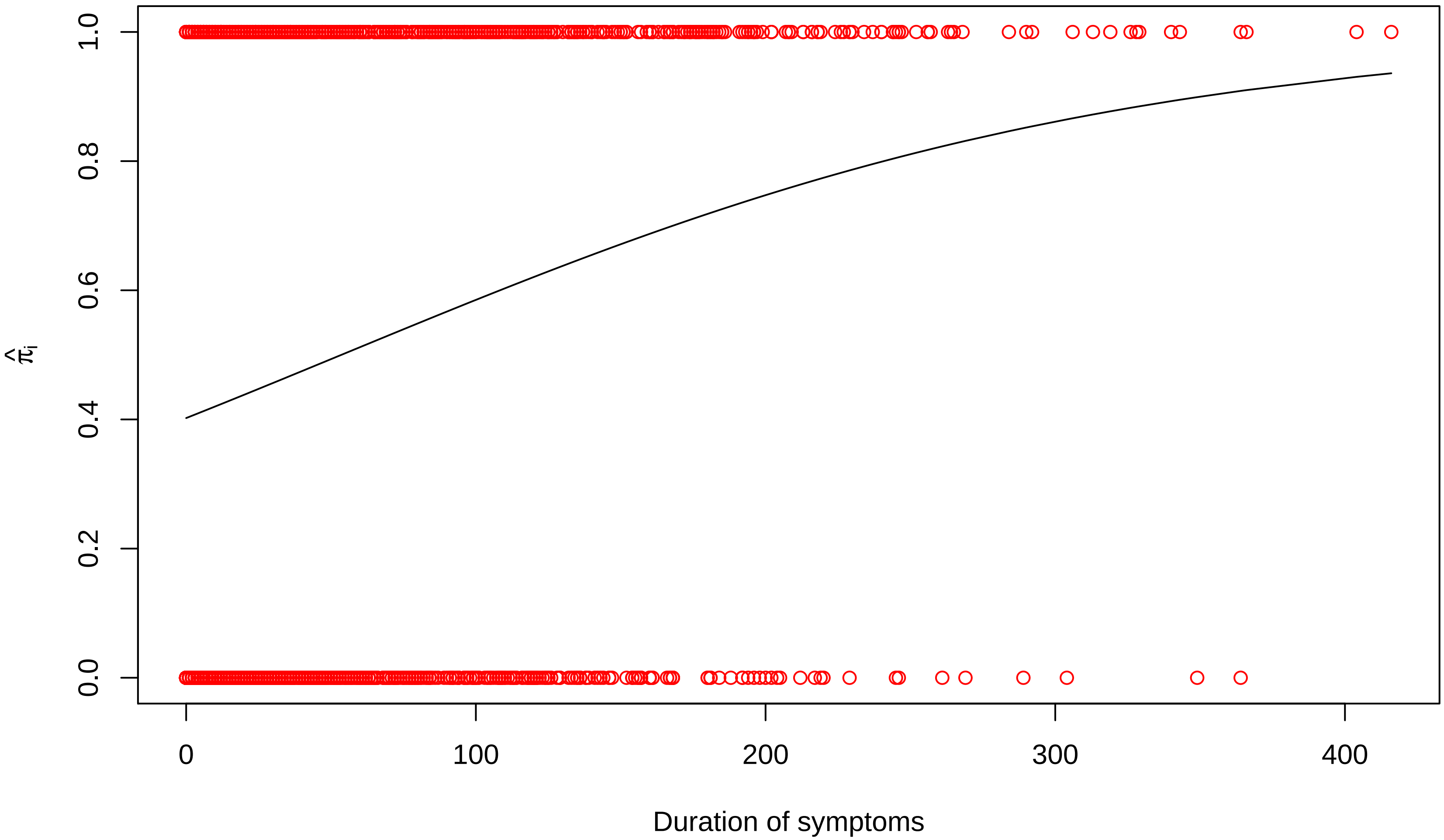
Fitted model on log-odds scale



Fitted model on odds scale



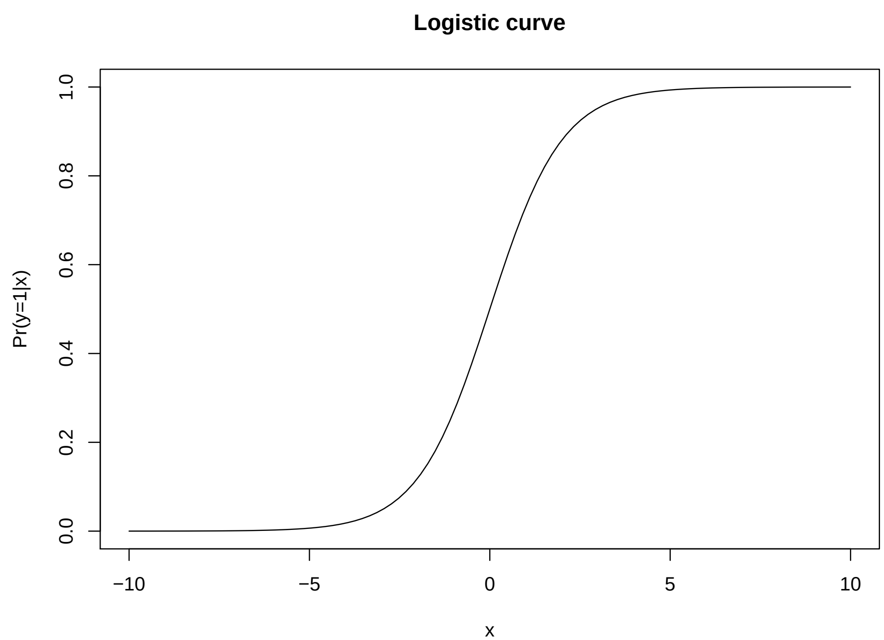
Fitted model on probability scale



Why is logistic regression so popular?

- Custom
- The shape of the logistic curve
- Estimates force to lie between 0 and 1
- Case-control studies

Shape of the logistic curve



The shape suggests that for some values of the predictor(s), the probability remains low. Then, there is some threshhold value of the predictor(s) at which the estimated probability of event begins to increase.

Study Design

We will touch on two major study designs.

- Case-control study: sampling is based on the outcome of interest
  - $Pr(E|D)$ is estimable, but $Pr(D|E)$ is not
  - **Only odds ratio** and not risks **can be estimated validly**.

- Cohort study: sampling is based on the predictor of interest
  - $Pr(D|E)$ is estimable, but not $Pr(E|D)$
  - Odds ratios and risks can be estimated.

Assumptions of the logistic regression model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$$

Limitations on scientific interpretation of the slope

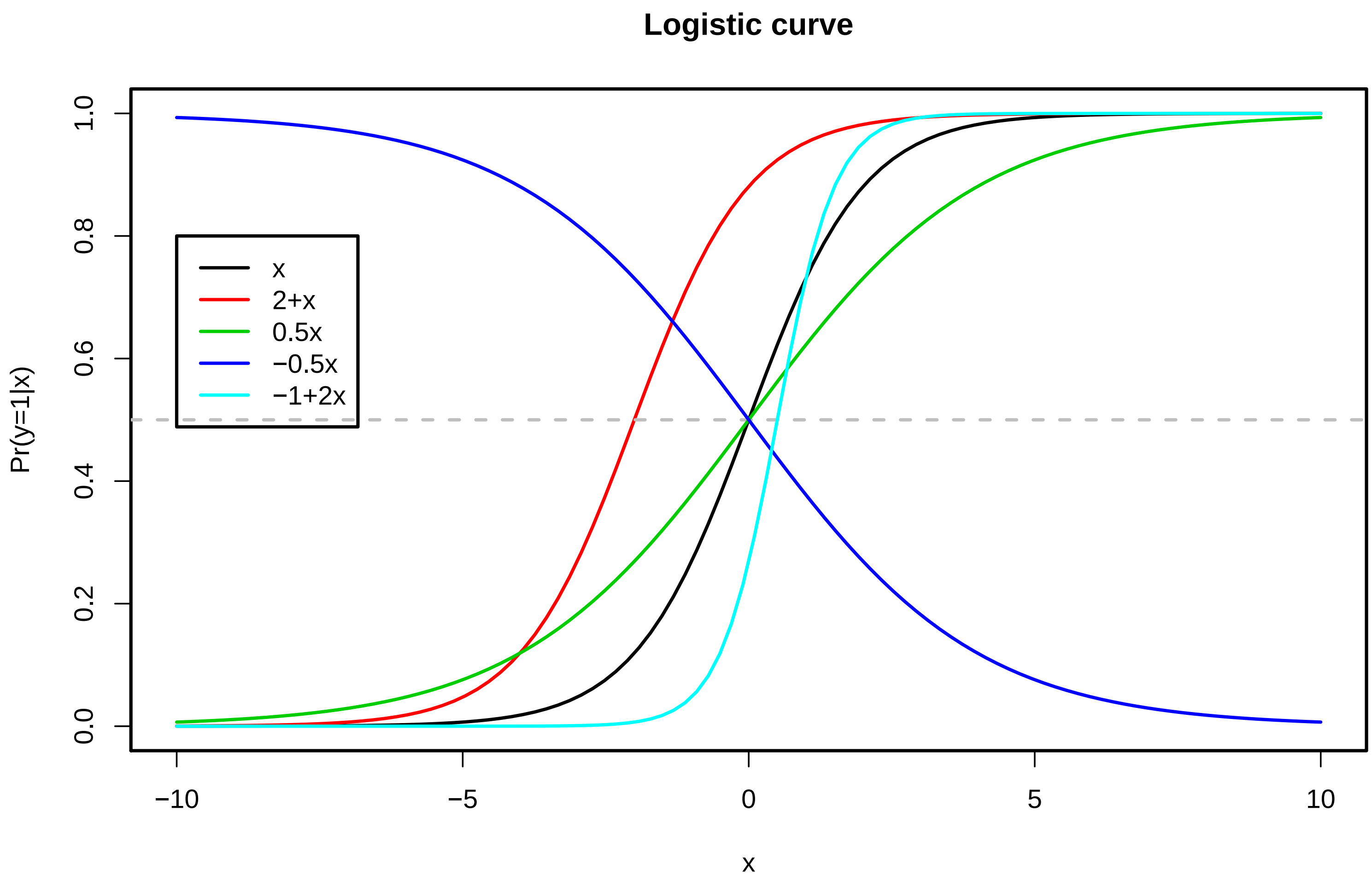
- If the log odds truly lie on a straight line, $\exp(\beta_1)$ is the odds ratio for **any** two groups that differ by 1 unit in the value of the predictor
  - $\exp(k\beta_1)$ for **any** k unit difference
- If the true relationship is nonlinear, then the odds ratio describes a "general trend" in the ratio over the distribution of the predictor values
  - "On average, the odds is $\exp(\beta_1)$ times larger for every unit increase in predictor values."

As we move towards using logistic regression to test for associations, we will be looking for first order (linear) trends in the log odds of response across groups defined by the predictor.

- If the response and predictor of interest were totally indepe-dent, the odds of response in each group would be the same (a flat line would describe the log odds of response across groups).

- A nonzero slope for the best fitting line on log odds suggests the presence of an association between the odds of response and a predictor.

How coefficients effect the shape of the logistic curve.



Example 2

Descriptive statistics for two groups of men. Variables are AGE and whether or not a subject had seen a physician (PHY) within the last six months (1=yes, 0=no).

	Group 1		Group 2	
	Mean	SD	Mean	SD
PHY	0.30		0.80	
AGE	40.18	5.34	48.45	5.02

Interest is whether there is an association between $GROUP$ and PHY.

The odds ratio estimated from this table is

$$OR = \frac{0.8/0.2}{0.3/0.7} = 9.3!$$

What issue do you see in this simple example? What do you think about AGE?

In summary, we have

- a binary predictor of interest ($GROUP$)
- a binary outcome of interest (PHY)
- a continuous control variable (AGE)

We can fit a logistic model where PHY is the response, GP is the predictor of interest and AGE is a control variable,

$$\text{logit}(Pr(PHY_i = 1|GP_i, AGE_i)) = \beta_0 + \beta_1 GP_i + \beta_2 AGE_i.$$

	Estimate	Std. Error
Intercept	-4.739	1.998
GP	1.599	0.577
AGE	0.096	0.048

The "age-adjusted odds ratio" in this example is $\exp(1.599) = 4.75 \ll 9.33$. Therefore, much of the intitially observed difference between the groups was really due to AGE.

What assumptions are we making when we model predictors additively on the odds and odds ratio scale?

Logistic regression with multiple predictors

Where there are no interacations, the predictors are assumed to act additively on the log-odds,

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

The odds ratio for a one unit increase in $x_j,\ j = 1, \ldots, p$ is

$$OR = \exp(\beta_j).$$

Although the predictors act additively on the log-odds scale, they are not additive on the odds or risk (probability) scales,

odds of disease given $x_{1i}, \ldots, x_{pi} = \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})$

and

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}.$$

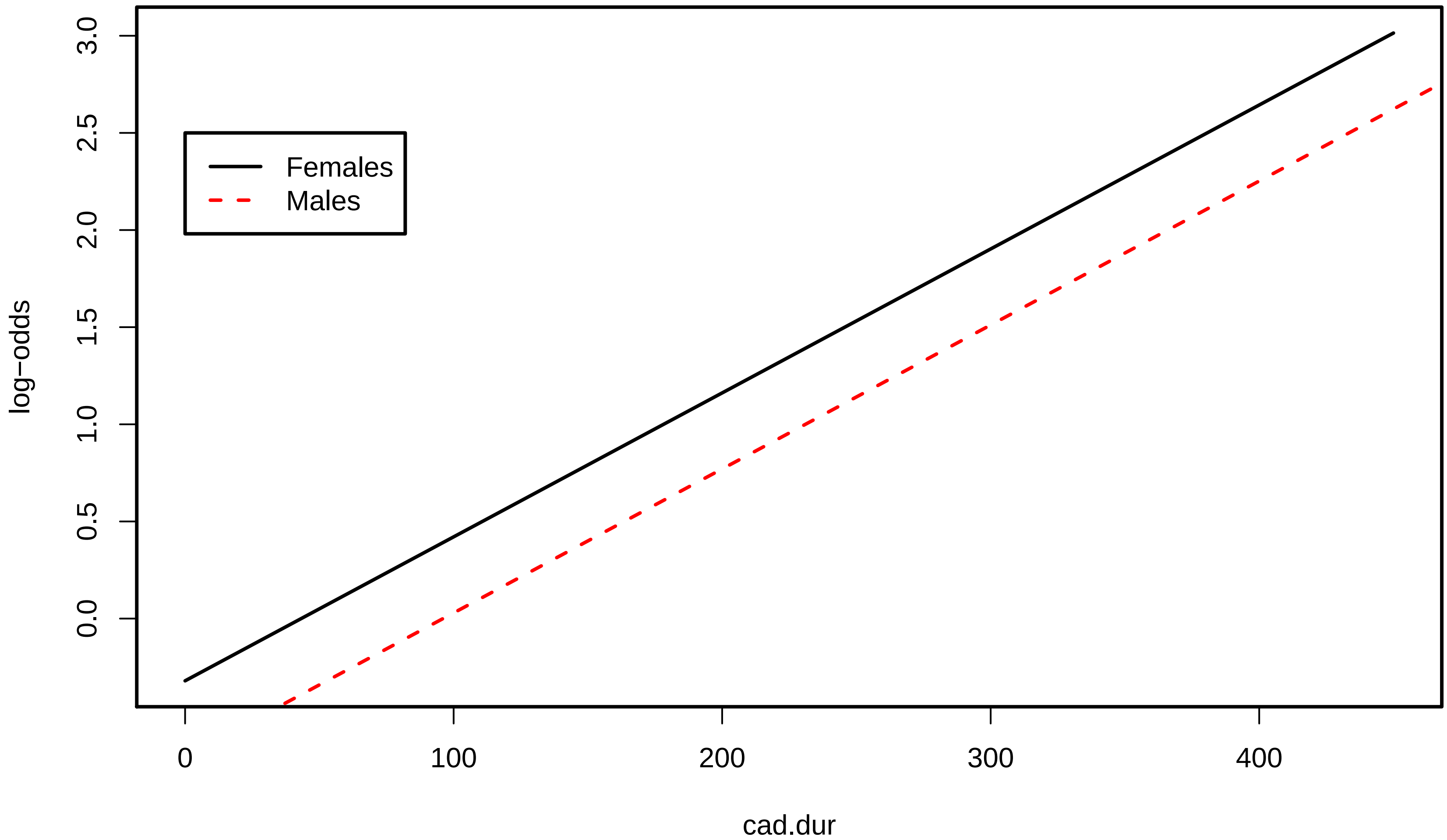
Example

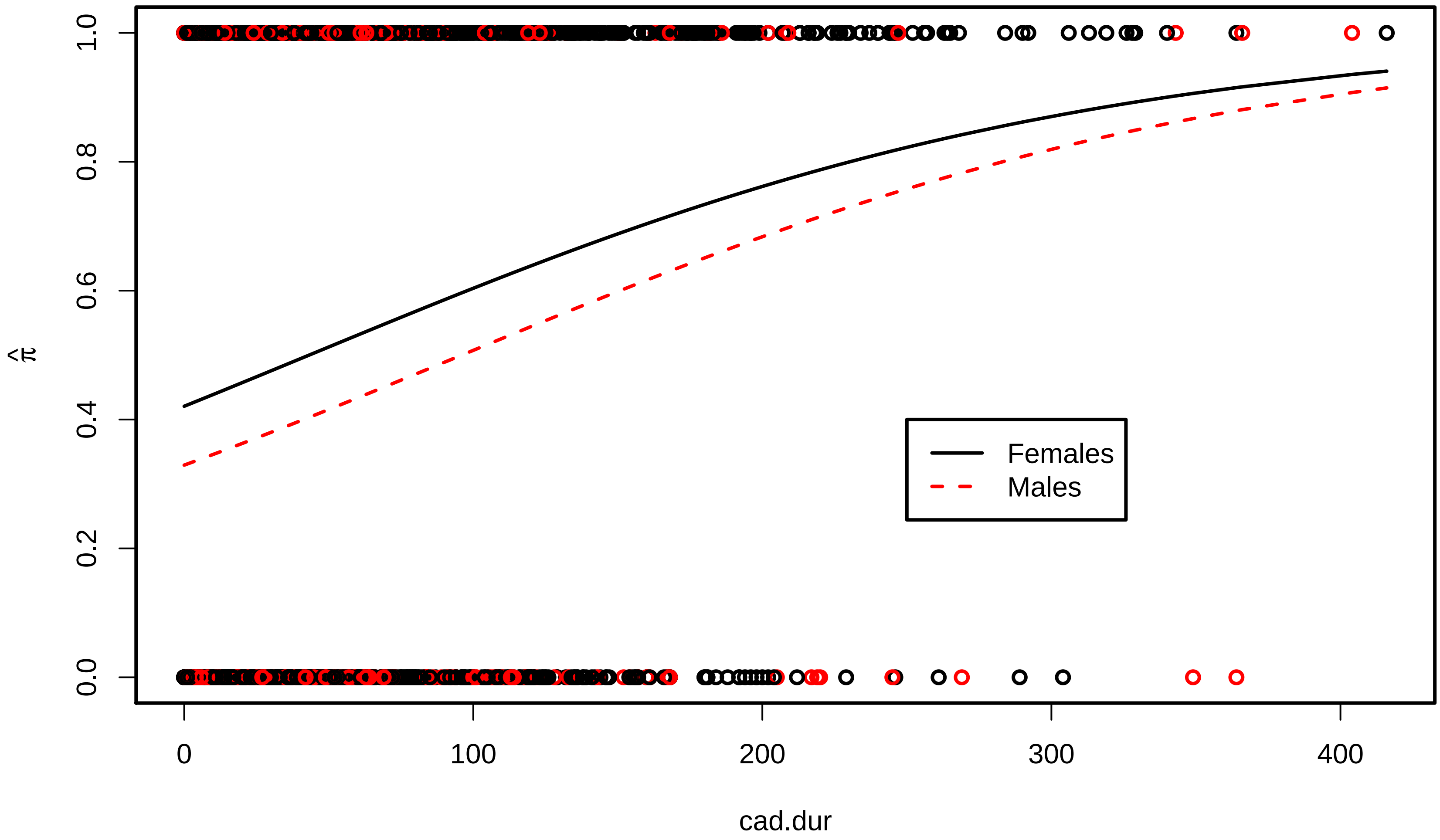
Following the cardiac catheterization example from the beginning of lecture, we will model the association between severe disease and time from onset of symptoms adjusted for gender. The model is

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 cad.dur_i + \beta_2 gender_i.$$

How do we interpret π_i here?

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.3203	0.0579	−5.53	0.0000
cad.dur	0.0074	0.0008	9.30	0.0000
sex	−0.3913	0.1078	−3.63	0.0003


Females
Males
log−odds
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
100
200
300
400
cad.dur



1.0
0.8
0.6
0.4
0.2
0.0
$\hat{\pi}$
0
100
200
300
400
cad.dur
Females
Males

Multiplicative interactions

Assume you have two binary predictors of disease, A and B. The risk of disease given the values of A and B are given in the following table,

		B = 1	B = 0
A	1	π_{11}	π_{10}
	0	π_{01}	π_{00}

where $\pi_{ij} = Pr(D = 1|A = i, B = j),\ j = 0, 1.$

With multiple predictors and interactions, we're often interested in the odds ratios over differences in two or more exposures. In this case we set one of the groups of predictors to be the reference group. In this case, our reference group is $(A = 0,\ B = 0)$ and

$$OR_{ij} = \frac{\text{odds of disease given } A = i,\ B = j}{\text{odds of disease given } A = 0,\ B = 0}.$$

The possible odds ratios of interest are

$$OR_{11} = \frac{\pi_{11}(1-\pi_{00})}{\pi_{00}(1-\pi_{11})},$$

$$OR_{10} = \frac{\pi_{10}(1-\pi_{00})}{\pi_{00}(1-\pi_{10})}$$

and

$$OR_{01} = \frac{\pi_{01}(1-\pi_{00})}{\pi_{00}(1-\pi_{01})}.$$

If there is no interaction,

$$OR_{11} = OR_{10} \times OR_{01}.$$

What does this mean?

Interaction in logistic regression

How can we relate this back to the regression model?

no interaction: $\text{logit}(\pi_i) = \beta_0 + \beta_1 A + \beta_2 B$

- odds of disease given $A = 1$, $B = 1$: $\exp(\beta_0 + \beta_1 + \beta_2)$
- odds of disease given $A = 0$, $B = 0$: $\exp(\beta_0)$
- $OR_{11} = \exp(\beta_1 + \beta_2) = \exp(\beta_1) \times \exp(\beta_2) = OR_{10} \times OR_{01}$

interaction: $\text{logit}\pi_i = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 A \times B$

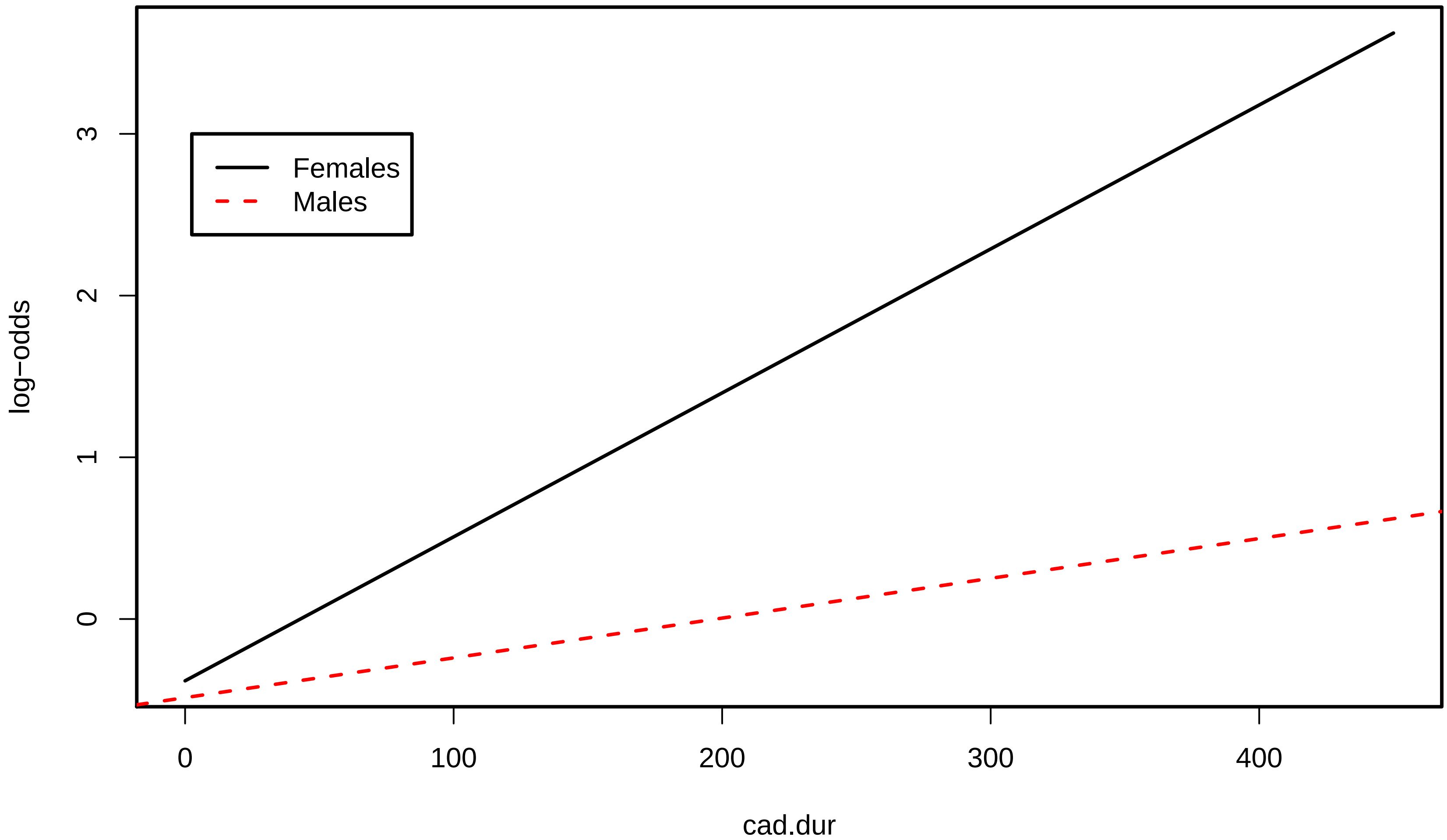
- odds of disease given $A = 1$, $B = 1$: $\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)$
- odds of disease given $A = 1$, $B = 0$: $\exp(\beta_0 + \beta_1)$
- odds of disease given $A = 0$, $B = 1$: $\exp(\beta_0 + \beta_2)$
- odds of disease given $A = 0$, $B = 0$: $\exp(\beta_0)$
- $OR_{11} = \exp(\beta_1 + \beta_2 + \beta_3) \neq OR_{10} \times OR_{01}$

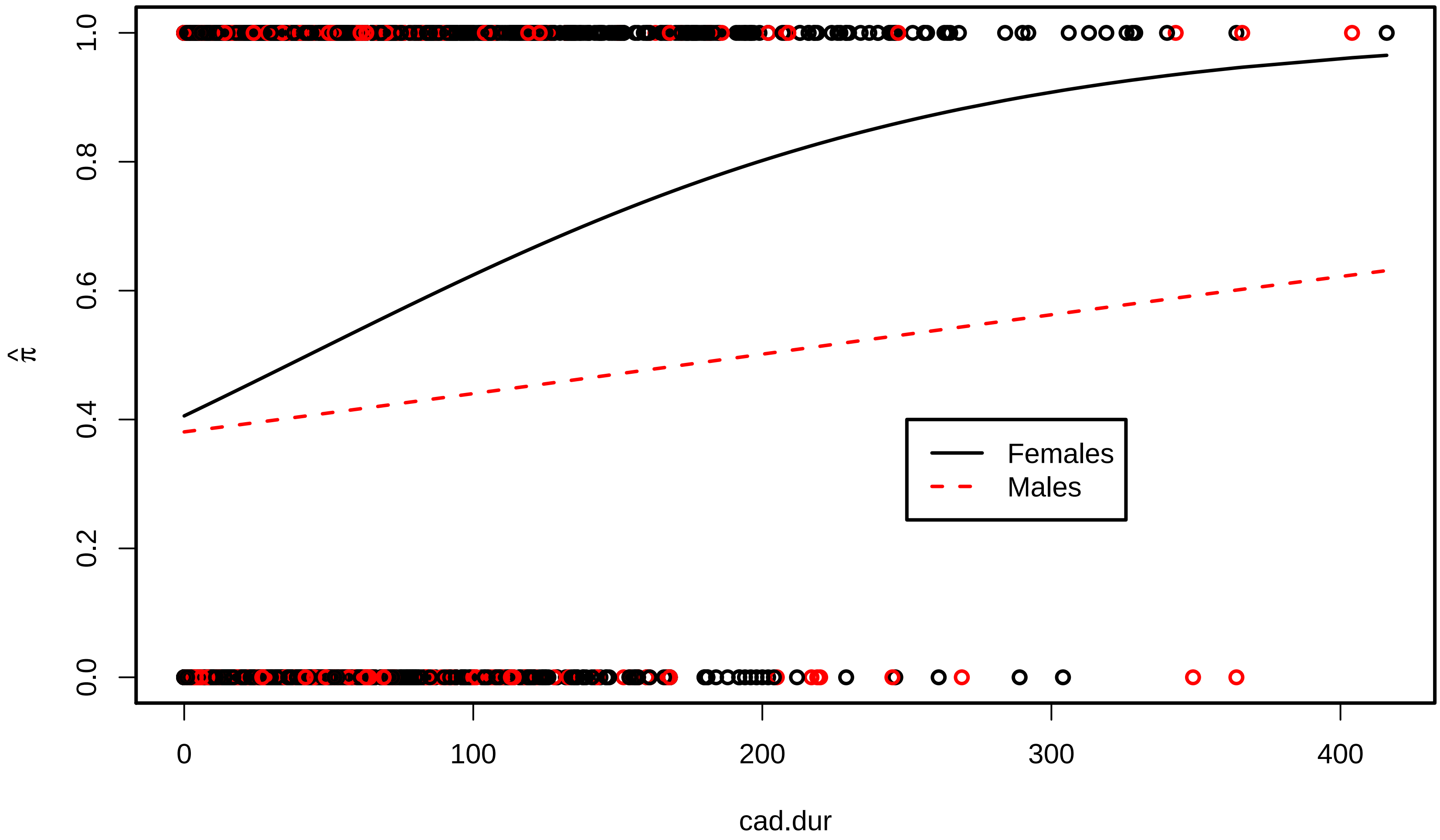
How could we assess interaction?

Interaction in catheterization example

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 cad.dur_i + \beta_2 gender_i + \beta_3 cad.dur_i \times gender_i$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.3822	0.0609	−6.28	0.0000
cad.dur	0.0089	0.0009	9.56	0.0000
sex	−0.1040	0.1342	−0.78	0.4382
cad.dur:sex	−0.0064	0.0018	−3.53	0.0004


Females
Males
log−odds
cad.dur
0
1
2
3
0
100
200
300
400


1.0
0.8
0.6
0.4
0.2
0.0
$\hat{\pi}$
0
100
200
300
400
cad.dur
Females
Males