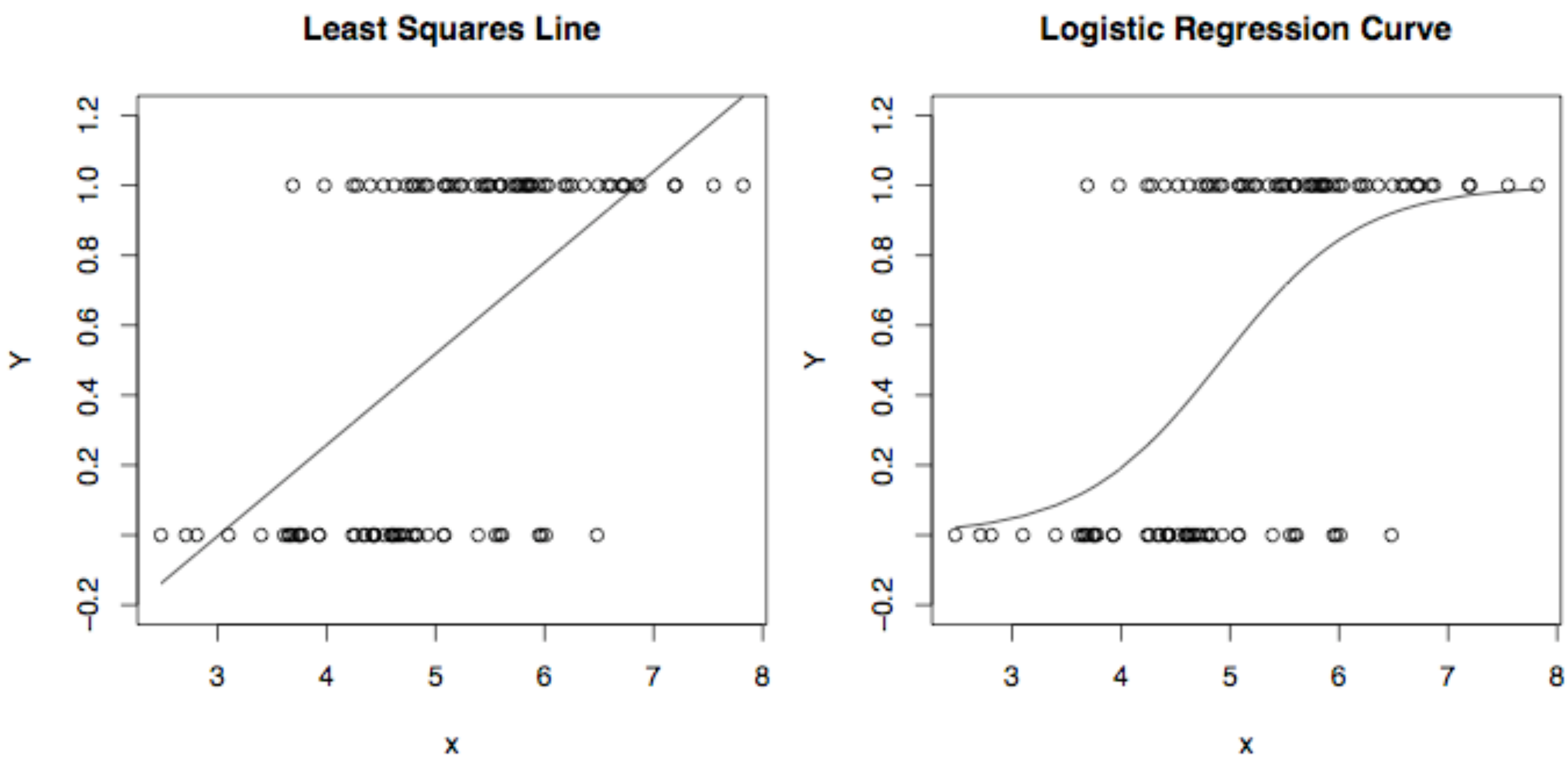
Logistic Regression

For a binary response variable:
1=Yes, 0=No

Least Squares vs. Logistic Regression



Linear regression model for the log odds of the event Y=1

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

Equivalent Statements

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

$$\begin{aligned}\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})} &= e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}} \\ &= e^{\beta_0} e^{\beta_1 x_1} \ldots e^{\beta_{p-1} x_{p-1}}\end{aligned}$$

$$P(Y=1|x_1, \ldots, x_{p-1}) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}}}$$

Non-linear Regression

- Regression is based on the conditional expected value of Y given **X**=**x**.
- For binary data, E(Y) = P{Y=1}
- Definitely a non-linear function of the β values.

$$P(Y=1|x_1,\ldots,x_{p-1}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}$$

$$P(Y=1|x_1,\ldots,x_{p-1}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}$$

$F(x) = \frac{e^x}{1+e^x}$ is called the *logistic distribution*.

- Could use any cumulative distribution function:

$$P(Y=1|x_1,\ldots,x_{p-1}) = F(\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1})$$

- CDF of the standard normal is sometimes used.
- Called probit analysis
- Can be closely approximated with a logistic regression.

In terms of log odds, logistic regression is like regular regression

$$\ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

In terms of plain odds,

- Logistic regression coefficients tell us about *odds ratios*
- For example, “Among 50 year old men, the odds of being dead before age 60 are three times as great for smokers.”

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = 3$$

Logistic regression

- X=1 means smoker, X=0 means non-smoker
- Y=1 means dead, Y=0 means alive

- Log odds of death = $\beta_0 + \beta_1 x$

- Odds of death = $e^{\beta_0} e^{\beta_1 x}$

$$\text{Odds of Death} = e^{\beta_0} e^{\beta_1 x}$$

Group	x	**Odds of Death**
Smokers	1	$e^{\beta_0} e^{\beta_1}$
Non-smokers	0	e^{β_0}

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = \frac{e^{\beta_0} e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Cancer Therapy Example

$$\text{Log Survival Odds} = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 x$$

Treatment	d_1	d_2	**Odds of Survival** $= e^{\beta_0} e^{\beta_1 d_1} e^{\beta_2 d_2} e^{\beta_3 x}$
Chemotherapy	1	0	$e^{\beta_0} e^{\beta_1} e^{\beta_3 x}$
Radiation	0	1	$e^{\beta_0} e^{\beta_2} e^{\beta_3 x}$
Both	0	0	$e^{\beta_0} e^{\beta_3 x}$

x = disease severity

For any given disease severity x,

$$\frac{\text{Survival odds with Chemo}}{\text{Survival odds with Both}} = \frac{e^{\beta_0} e^{\beta_1} e^{\beta_3 x}}{e^{\beta_0} e^{\beta_3 x}} = e^{\beta_1}$$

In general,

- When x_k is increased by one unit and all other independent variables are held constant, the odds of Y=1 are multiplied by e^{β_k}
- That is, e^{β_k} is an **odds ratio** --- the ratio of the odds of Y=1 when x_k is increased by one unit, to the odds of Y=1 when everything is left alone.
- As in ordinary regression, we speak of "controlling" for the other variables.

The conditional probability of Y=1

$$P(Y=1|x_1,\ldots,x_{p-1}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}$$

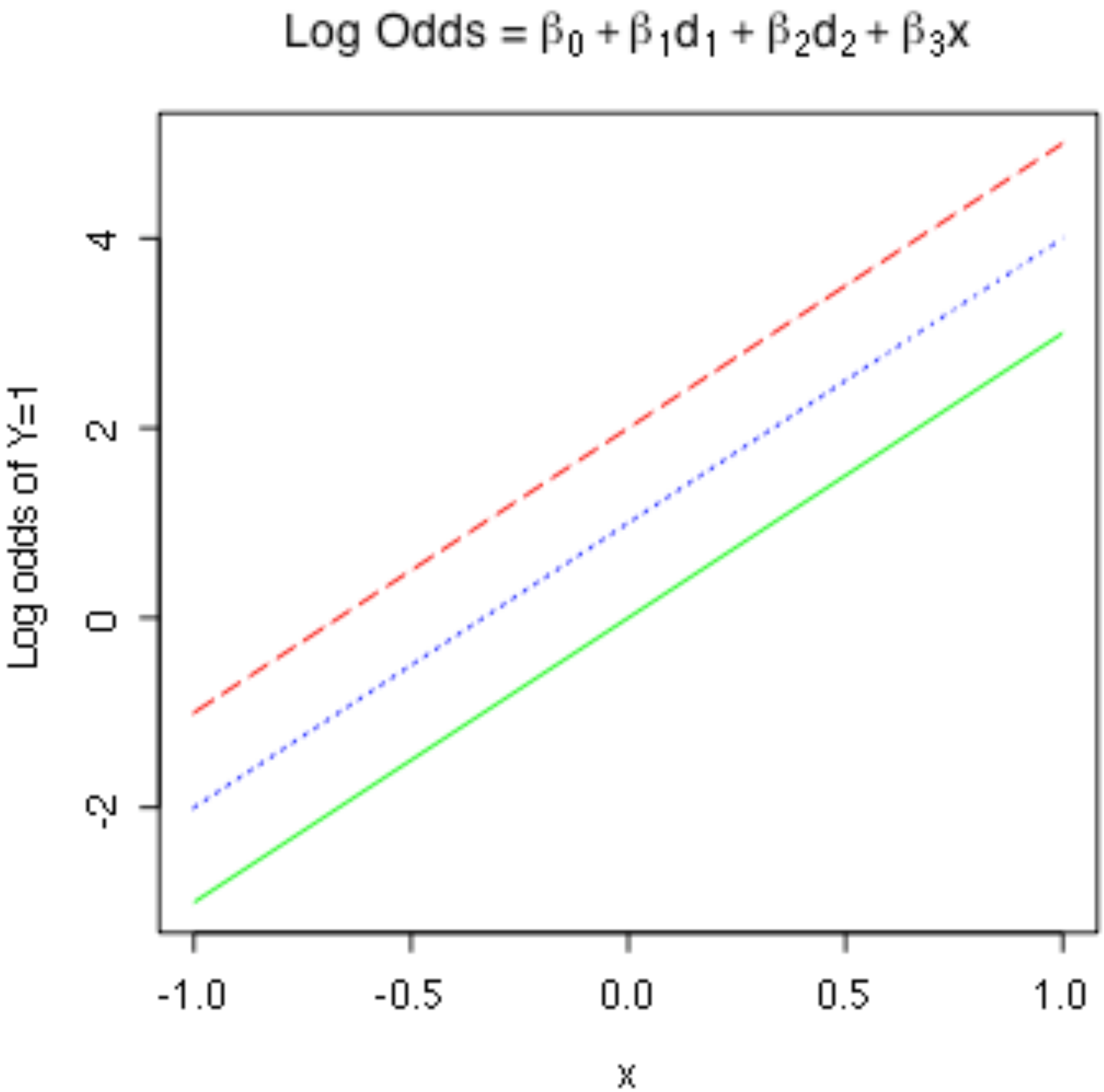
This formula can be used to calculate a predicted P(Y=1|**x**). Just replace betas by their estimates

It can also be used to calculate the probability of getting the sample data values we actually did observe, as a function of the betas.

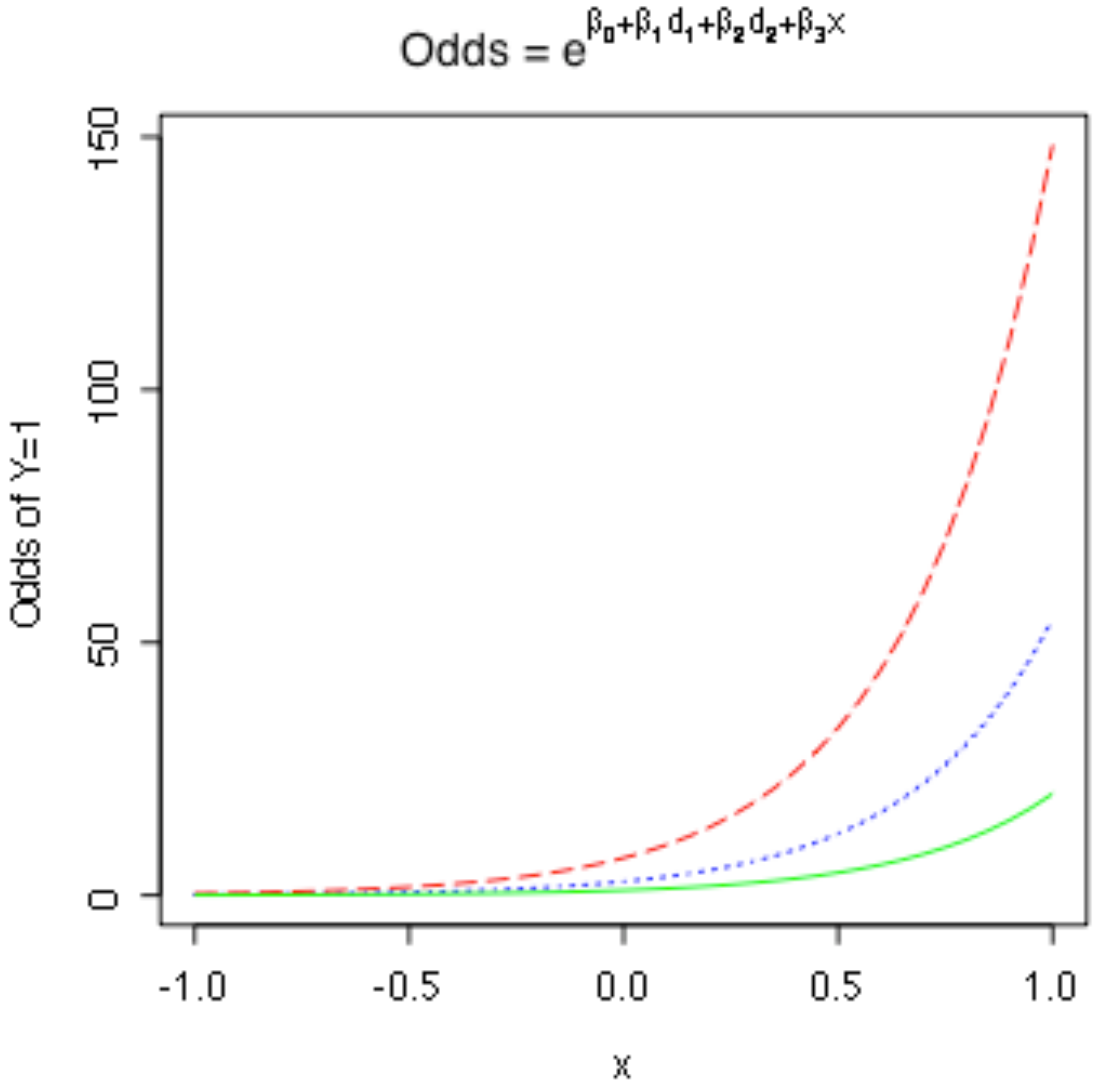
Maximum likelihood estimation

- Likelihood = Conditional probability of getting the data values we did observe,
- As a function of the betas
- Maximize the (log) likelihood with respect to betas.
- Maximize numerically (“Iteratively re-weighted least squares”)
- Likelihood ratio and Wald tests as usual

Equal slopes in the log odds scale

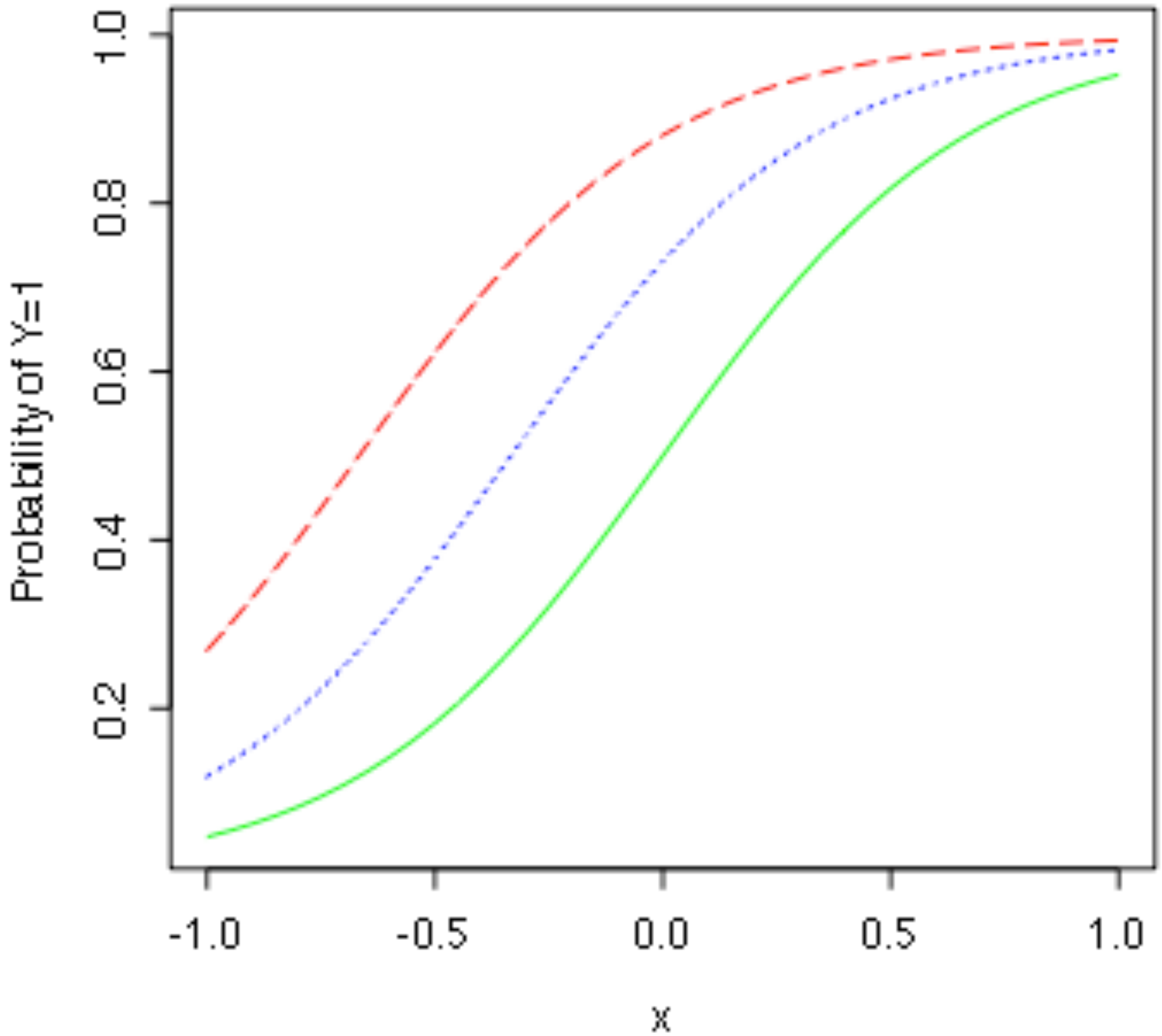


Equal slopes in the log odds scale means proportional odds



Proportional Odds in Terms of Probability

$$\text{Probability} = \frac{e^{\beta_0+\beta_1 d_1+\beta_2 d_2+\beta_3 x}}{1 + e^{\beta_0+\beta_1 d_1+\beta_2 d_2+\beta_3 x}}$$



Interactions

- With equal slopes in the log odds scale, *differences* in odds and *differences* in probabilities do depend on x.
- Regression coefficients for product terms still mean something.
- If zero, they mean that the *odds ratio* does not depend on the value(s) of the covariate(s).
- Odds ratio has odds of Y=1 for the reference category in the denominator.

Mathematical Symbols on R Plots

```
prob = odds/(1+odds)
plot(x,prob,pch=' ',ylab='Probability of Y=1')
lines(xpts,prob[d1==1],col='blue',lty=3)
lines(xpts,prob[d2==1],col='red',lty=5)
lines(xpts,prob[d1+d2==0],col='green',lty=1)
title(expression("Probability = " *
frac(e^{beta[0] + beta[1]*d[1] + beta[2]*d[2] + beta[3]*x},
     1+e^{beta[0] + beta[1]*d[1] + beta[2]*d[2] + beta[3]*x}) ))
```

See demo(plotmath) to see what else you can do!