

Logistic Regression

Global MECOR Course
Kenya, 2011

Linear Regression Review

$$FEV_1 = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Height}$$

$$E(FEV_1) = \mu = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Height}$$

Where data are assumed to be normally distributed with mean equal to μ

Models such as these are appropriate for **continuous** outcome measures such as FEV_1 , weight, blood pressure

What if our outcome is Binary?

Common Binary Outcome Measures

- **Healthy vs unhealthy**
 - E.g., heart disease (y/n), Cancer (y/n), COPD (y/n)
- **Progressive disease vs stable disease**
 - Based on, e.g., cancer stage
- **Alive vs dead**

Convenient Coding of Binary Outcomes

- COPD = 0 if $FEV_1/FVC \geq 0.70$
= 1 if $FEV_1/FVC < 0.70$
- Worse = 0 if tumor staging is stable
= 1 if tumor staging is increasing
- Dead = 0 if alive
= 1 if deceased

Note use of 0/1 coding and descriptive names that define “1”

The Logistic Regression Model

- Consider the case of a binary indicator of vital status
 - $\text{Dead} = 0$ if alive
 - $\text{Dead} = 1$ if deceased
- If Dead is coded 0/1, then its expected value is equal to the probability that $\text{Dead}=1$. i.e.,
$$E(\text{Dead}) = P = \text{Probability of death}$$

The Logistic Regression Model

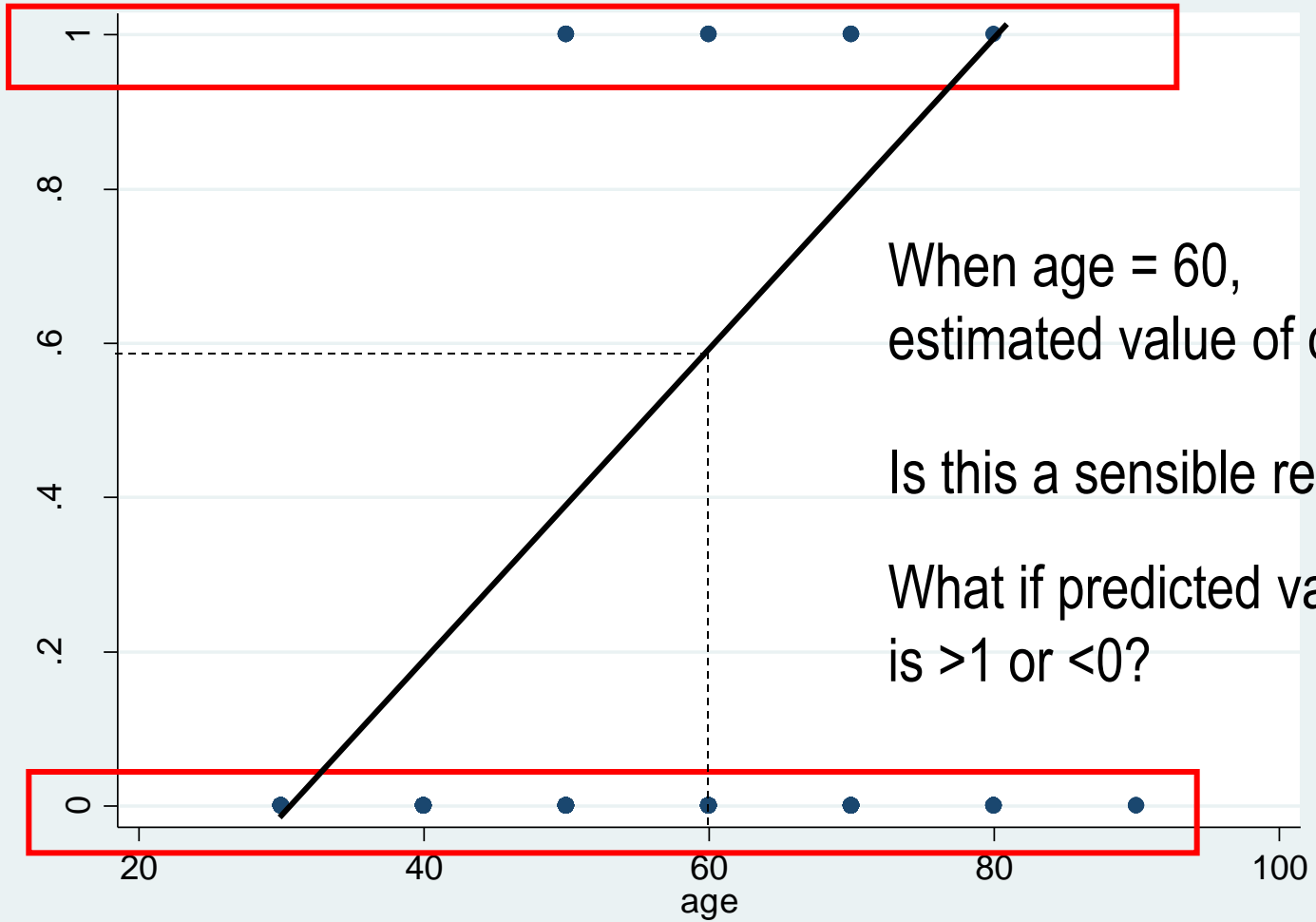
Suppose we want to model the association between vital status and age...

- If we fit the data using standard linear regression, our model would be of the form

$$\text{Exp(Dead)} = P = \beta_0 + \beta_1 \text{Age}$$

- That is, we assume the probability of death varies in a linear manner with age.

The Logistic Regression Model



The Logistic Regression Model

- Logistic regression analysis is a tool for modeling binary data that overcomes some of the limitations of linear regression.
- Rather than assuming the data are normally distributed, which we know isn't true, we first assume the data follow a **binomial distribution**, which implicitly assumes we have a series of 0/1 observations each with probability P of being dead (i.e., $Dead = 1$).

The Logistic Regression Model

Rather than assuming P is a linear combination of variables of interest, e.g.,

$$P = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}$$

we instead assume

$$P = \frac{e^{\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}}}{1 + e^{\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}}}$$

or equivalently,

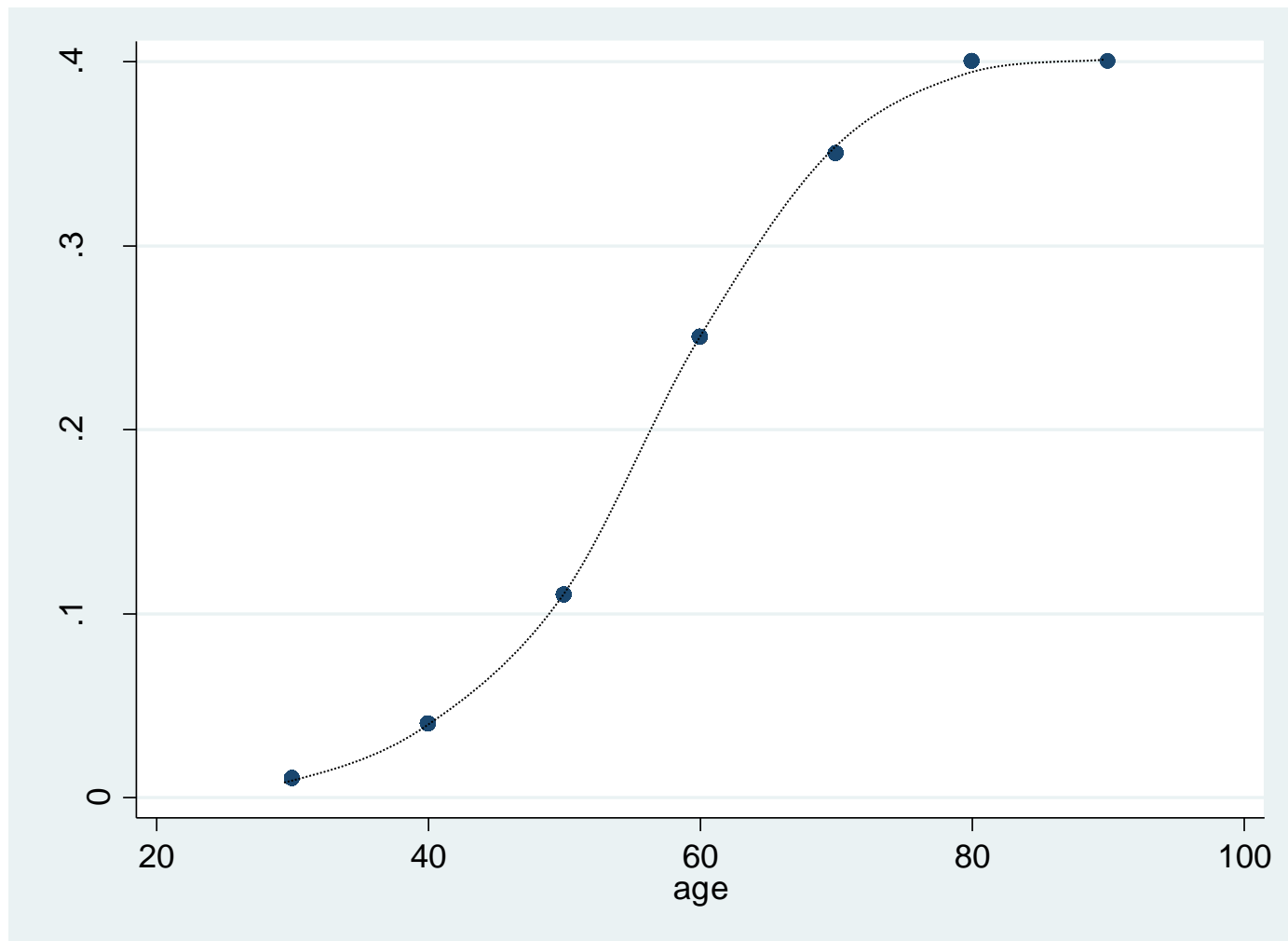
$$\ln[P/(1-P)] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}$$

The Logistic Regression Model

$$\ln[P/(1-P)] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Height}$$

- The function $\ln[P/(1-P)]$ is referred to as the “logit” of P , hence the term “logistic” regression!
- Unlike the linear regression model, the logit function has the desirable property that it is always between 0 and 1.
- It also turns out to have some statistical properties that makes it a particularly desirable function of P to estimate.

Sample logistic function



The Logistic Regression Model

Interpretation of coefficients

$$\ln[P/(1-P)] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}$$

- Recall that $P/(1-P)$ is the odds of our outcome of interest, in this case death.
- Hence the logit of P is the same as the $\ln(\text{odds})$ of death, and so the odds of death can be written

$$\text{odds} = e^{\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}}$$

The Logistic Regression Model

Interpretation of coefficients

$$\text{odds} = e^{\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}}$$

$$\begin{aligned} \text{OR (male vs female)} &= \frac{e^{\beta_0 + \beta_1 \text{Age} + \beta_2}}{e^{\beta_0 + \beta_1 \text{Age}}} \\ &= e^{\beta_2} \end{aligned}$$

$$\Rightarrow \beta_2 = \ln(\text{OR}_{\text{males vs females}})$$

The Logistic Regression Model

Interpretation of coefficients

$$\text{odds} = e^{\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}}$$

Similarly we can calculate the OR associated with an increase in age of 10 years as

$$\begin{aligned} \text{OR (10 yr incr in age)} &= \frac{e^{\beta_0 + \beta_1 (\text{Age} + 10) + \beta_2 \text{Male}}}{e^{\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}}} \\ &= e^{10\beta_1} = (e^{\beta_1})^{10} \end{aligned}$$

$$\begin{aligned} \Rightarrow \ln(\text{OR}_{10 \text{ year increase in age}}) &= 10\beta_1 \\ &= 10 * \ln(\text{OR}_{1 \text{ yr incr}}) \end{aligned}$$

Programming Logistic Regression in Stata

Suppose we want to use logistic regression to relate the probability of death to history of ever smoking. We have a 0/1 variable “dead” and a 0/1 variable “evsmk”. In Stata, we would write

```
logit (dead) (evsmk)
```

“logit” tells Stata to run a logistic regression model

The first variable, dead, is the outcome. Stata assumes by default that we want to model the odds of the higher value (1=dead) vs. the lower value (0=alive). Not all programs work this way!

All remaining variables define the predictor variables.

Programming Logistic Regression in SPSS

If you are using SPSS for Windows, you do the same thing (specify analytic model, outcome variable, and predictor variables), but you do so by selecting options and variables from various drop down menus.

In the background, SPSS is creating code analogous to Stata's "logit dead evsmk".

Programming Logistic Regression in Stata

So when we write

```
logit dead evsmk
```

we are asking Stata to fit the model

$$\text{logit}(P) = \ln(\text{odds of death}) = \beta_0 + \beta_1 * \text{evsmk}$$

How do we use the output of the model to test whether the odds of death depends on ever smoking status?

Test whether the coefficient of evsmk, β_1 , equals zero.

The Logistic Regression Model

Hypothesis testing and confidence intervals

- Testing $H_0: \ln(\text{OR}) = \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$ is equivalent to testing
$$H_0: \text{OR} = e^{\beta_1} = 1 \text{ vs. } H_a: e^{\beta_1} \neq 1$$
- Use large sample normality of β_1 to compute p-values and to construct confidence limits
 - $\beta_1 \div \text{SE}(\beta_1)$ should look like a z-score under H_0
... use this to compute p-value and test null hypothesis
 - $\beta_1 \pm 1.96 * \text{SE}(\beta_1)$ is an approximate 95% confidence interval

The Logistic Regression Model

An example

Vital status	Ever smoker (evsmk=1)	Never smoker (evsmk=0)
Dead (dead=1)	62 ($p_1 = 14\%$)	18 ($p_0 = 9\%$)
Alive (dead= 0)	384 (86%)	186 (91%)

100%

100%

$$\text{Odds ratio for smokers to never-smokers} = \frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{p_0}{1-p_0}\right)} = 1.67$$

STATA logistic regression output for: logit dead evsmk **OR = $e^{.512} = 1.67$**

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
evsmk	.5118667	.28225	1.81	0.070	-.0413331	1.065067
_cons	-2.335375	.2468438	-9.46	0.000	-2.81918	-1.85157

ln(OR)

The Logistic Regression Model

An example

- Test null hypothesis that probability of death is not associated with ever smoking status
 - under H_0 , $\beta_1 / SE(\beta_1)$ should look like a z-score (normal with mean 0 and variance 1)... use to compute p-value
- Now construct confidence limits
 - $\beta_1 \pm 1.96 * SE(\beta_1)$ is an approximate 95% confidence interval

STATA logistic regression output for: logit dead evsmk

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
evsmk	.5118667	.28225	1.81	0.070	[-.0413331 1.065067]
_cons	-2.335375	.2468438	-9.46	0.000	[-2.81918 -1.85157]

The Logistic Regression Model

Computing confidence intervals for the odds ratio

Because β_1 is more normally distributed than e^{β_1} , we construct CIs for the $\ln(\text{OR})$ and then exponentiate these to get corresponding CIs for the OR.

$$95\% \text{ CI for } \ln(\text{OR}) = (-0.04, 1.07)$$

$$\Rightarrow 95\% \text{ CI for OR} = e^{(-0.04, 1.07)} = (0.96, 2.90)$$

Variations in Software Output

Letting computer transform data for us

STATA logistic regression output: `logit dead evsmk`

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
evsmk	.5118667	.28225	1.81	0.070	-.0413331	1.065067
_cons	-2.335375	.2468438	-9.46	0.000	-2.81918	-1.85157

Default output in the log scale, Stata labels output as “coefficient”

STATA logistic regression output: `logit dead evsmk, or`

	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
evsmk	1.668403	.4709067	1.81	0.070	.9595094	2.901032

Output requested in the transformed scale

The Logistic Regression Model

Adjusting for potential confounder variables

Might the association between ever smoking and death be confounded by age?

Does it meet the criteria for a potential confounder?

- Is age likely associated with death?
- Is age likely associated with smoking history?

So how might we go about adjusting for the potentially confounding effects of age?

The Logistic Regression Model

Adjusting for potential confounder variables

So we might fit the following model

$$\ln[P/(1-P)] = \beta_0 + \beta_1 \text{evsmk} + \beta_2 \text{age}$$

Under this model, we say the effect of ever smoking is now adjusted for the potentially confounding effect of age. If age was categorical, the resulting odds ratio would be analogous to the pooled OR that you would get from a stratified 2x2 table analysis that crosses dead by evsmk for each age category.

We could adjust for additional potential confounders by adding them to the model as other main effects.

Adjusting for Confounder Variables

An example

STATA output from: **logit dead evsmk, or**

	Odds Ratio	Std. Err.	Z	P> z	[95% Conf. Interval]
evsmk	1.668403	.4709067	1.81	0.070	.9595094 2.901032

STATA output from: **logit dead evsmk age, or**

	Odds Ratio	Std. Err	Z	P> z	[95% Conf. Interval]
evsmk	4.219142	1.397128	4.35	0.000	2.204738 8.074048
age	1.108467	.0147424	7.74	0.000	1.079946 1.137742

The order in which I list the variables doesn't matter.

- The second model gives the OR for death associated with ever smoking *after adjusting for age*
- Note the change in the size of the smoking OR between the two models. This is a classic example of confounding.

The Logistic Regression Model

Adjusting for potential effect modification

Now suppose we want to know whether the effect of ever smoking on death differs for men and women. In classical epidemiologic terms, this means we want to know if the OR associated with ever smoking varies by gender.

How would we test for the presence of effect modification in our logistic model?

As we learned previously, we use interaction terms!

The Logistic Regression Model

Adjusting for potential effect modification

$$\ln[(P/(1-P))] = \beta_0 + \beta_1 \text{evsmk} + \beta_2 \text{male} + \beta_3 \text{evsmk} * \text{male}$$

male	evsmk	model
0	0	β_0
0	1	$\beta_0 + \beta_1$
1	0	$\beta_0 + \beta_2$
1	1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

$\beta_1 = \ln(\text{OR})$ for ever vs. never smoking in women

$\beta_2 = \ln(\text{OR})$ for ever male vs. female gender in never smokers

$\beta_3 =$ difference in $\ln(\text{OR})$ for ever smoking between men & women

$=$ difference in $\ln(\text{OR})$ for male sex between ever & never smokers

Testing $H_0: \beta_3 = 0$ is a test of whether there is effect modification.

Adjusting for Effect Modification

An example (a different dataset)

STATA output from: logit evsmk, or

	Odds Ratio	Std. Err	Z	P> z	[95% Conf. Interval]
evsmk	2.001	0.498	2.79	0.005	1.229 3.259

STATA output from: logit evsmk if male==0, or (women)

	Odds Ratio	Std. Err	Z	P> z	[95% Conf. Interval]
evsmk	0.914	0.366	-0.22	0.823	0.417 2.005

STATA output from: logit evsmk if male==1, or (men)

	Odds Ratio	Std. Err	Z	P> z	[95% Conf. Interval]
evsmk	2.550	0.916	2.61	0.009	1.261 5.155

The overall OR appears to be a weighted average of the separate ORs. This is the classic pattern in effect modification.

Adjusting for Effect Modification

An Example

So now we formally test for effect modification.

STATA output from: `logit dead evsmk male evsmk_male`, or
(`evsmk_male = evsmk*male`)

	Odds Ratio	Std. Err	Z	P> z	[95% Conf. Interval]	
evsmk	0.915	0.366	-0.22	0.823	0.417	2.005
male	1.029	0.459	0.06	0.949	0.429	2.465
evsmk_male	2.788	1.500	1.91	0.057	0.972	8.001

The p-value for the interaction is 0.057, so we would fail to reject the null hypothesis of no effect modification at the 0.05 level, but the data are clearly very suggestive of effect modification. Note that the evsmk output is identical to that for women on the previous slide, just as it should be!

The Logistic Regression Model

Adjusting for potential effect modification

Technically we can fit interactions with either categorical or continuous variables, though conceptually it is easiest to think about what the model means when one or both of the terms is categorical.

- $\ln[P/(1-P)] = \beta_0 + \beta_1 \text{evsmk} + \beta_2 \text{male} + \beta_3 \text{evsmk} * \text{male}$
- $\ln[P/(1-P)] = \beta_0 + \beta_1 \text{evsmk} + \beta_2 \text{age} + \beta_3 \text{evsmk} * \text{age}$

In the latter case we can think of evsmk as a binary variable of primary interest with age as the continuous effect modifier, or of age as a continuous variable of interest with ever smoking status as the binary effect modifier.