Logistic Regression: Interaction Terms

Interactions in Logistic Regression

- For linear regression, with predictors X_1 and X_2 we saw that an interaction model is a model where the interpretation of the effect of X_1 depends on the value of X_2 and *vice versa*.
- Exactly the same is true for logistic regression.
- The simplest interaction models includes a predictor variable formed by multiplying two ordinary predictors:

$$\text{logit}(\mathbb{P}(Y=1)) = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \boxed{\beta_3 \times X_1 \times X_2}$$

- Interaction term

Interactions in Logistic Regression

We will look at the interpretation of interactions in 3 cases:

1. Interaction between two dummy variables.
2. Interaction between a dummy and a continuous variable.
3. Interaction between two continuous variables.

Interaction Between 2 Dummy Variables

- Consider a logistic model for the risk of suffering a heart attack over a year in terms gender and smoking status:

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \mathsf{sex} + \beta_2 \mathsf{smoke} + \beta_3 (\mathsf{sex} \times \mathsf{smoke})$$

- sex indicates gender (male=1, female=0)
- smoke indicates smoking status (smokes=1, does not=0).

Interpreting the Intercept

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \mathsf{sex} + \beta_2 \mathsf{smoke} + \beta_3(\mathsf{sex} \times \mathsf{smoke})$$

- In order to interpret β_0 we need to find a situation in which the final three terms in the equation vanish.
- This happens when an observation corresponds to a female non-smoker, for then sex=0 and smoke=0.

$$\begin{aligned}\operatorname{logit} \mathbb{P}(Y = 1) &= \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 + \beta_3(0 \times 0) \\ &= \beta_0\end{aligned}$$

Female Non-smoker

- Consequently, β_0 *is the log odds in favour of a female non-smoker suffering from a heart attack.*

Interpretations of Other Quantities Involving β_0

We can also give interpretations on the odds scale and on the probability scale:

- $\exp(\beta_0)$ is the odds in favour of a female non-smoker suffering from a heart attack.
- $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ is the probability of a female non-smoker suffering from a heart attack.

Interpreting β_1 and β_2

$$\text{logit}\,\mathbb{P}(Y=1) = \beta_0 + {\color{red}\beta_1}\mathsf{sex} + \beta_2\mathsf{smoke} + \beta_3(\mathsf{sex} \times \mathsf{smoke})$$

- We would know how to interpret β_1 if the interaction term was not there.
- Since in that case would just have an ordinary multivariate logistic model.
- This happens when an observation corresponds to a non-smoker, for then smoke=0.

$$\begin{aligned}\text{logit}\,\mathbb{P}(Y=1) &= \beta_0 + \beta_1 \times \mathsf{sex} + \beta_2 \times 0 + \underbrace{\beta_3(\overset{1}{\mathsf{sex}} \times 0)}_{0} \\ &= \beta_0 + \beta_1 \times \mathsf{sex}\end{aligned}$$

1
Males

Non smoker
to analyze
only males

Interpreting β_1 and β_2

$\text{Sex} \cdot \text{Smoker} = 1 \cdot 0$
$= 0$

- Amongst non-smokers

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \times \mathsf{sex}$$

- We know how to interpret β_1 in this case as its a univariate logistic model.

- β_1 is the log-odds ratio comparing males and females **amongst non-smokers**.

- $\exp(\beta_1)$ is the odds ratio comparing males and females amongst non-smokers.

Interpreting β_1 and β_2

$$\operatorname{logit} \mathbb{P}(Y=1) = \beta_0 + \beta_1 \mathsf{sex} + \beta_2 \mathsf{smoke} + \beta_3(\mathsf{sex} \times \mathsf{smoke})$$

- To interpret β_2 we need to get rid of the interaction term without getting rid of the $\beta_2\mathsf{smoke}$ term.
- Same argument as before but now set sex=0 (female): Female

Female

$$\begin{aligned}\operatorname{logit} \mathbb{P}(Y=1) &= \beta_0 + \beta_1 \times 0 + \beta_2 \times \underbrace{\mathsf{smoke}}_{1} + \beta_3(0 \times \underbrace{\mathsf{smoke}}_{1}) \\ &= \beta_0 + \beta_2 \times \mathsf{smoke}\end{aligned}$$

$0 \cdot 1 = 0.$

- β_2 is the log-odds ratio comparing smokers with non-smokers **amongst females**.

Interpreting β_3

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \textsf{sex} + \beta_2 \textsf{smoke} + \beta_3 (\textsf{sex} \times \textsf{smoke})$$

- To interpret β_3 rewrite the regression equation:

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + [\beta_1 + \beta_3 \textsf{smoke}]\textsf{sex} + \beta_2 \textsf{smoke}$$

- This looks like a multivariate regression model with sex and smoke as predictors where:
  - $\beta_1 + \beta_3 \textsf{smoke}$ is the log-odds ratio for males *vs.* females;
  - β_2 is the log odds ratio for smokers *vs.* non-smokers.

- β_3 is the difference between the log-odds ratio comparing males *vs* females in smokers and the log-odds ratio comparing males *vs.* females in non-smokers.

Interpreting β_3

$$\text{logit}\,\mathbb{P}(Y=1) = \beta_0 + \underbrace{\overset{1}{\beta_1\text{sex}}}_{\text{Male}} + \underbrace{\overset{1}{\beta_2\text{smoke}}}_{\text{Smoker}} + \beta_3(\overset{1}{\text{sex}} \times \overset{1}{\text{smoke}})$$

- We could just as well have rewritten the equation this way:

$$\text{logit}\,\mathbb{P}(Y=1) = \beta_0 + \beta_1\text{sex} + [\beta_2 + \beta_3\text{sex}]\text{smoke}$$

- β_3 is the difference between the log-odds ratio comparing smokers *vs* non-smokers in males and the log-odds ratio comparing smokers *vs.* non-smokers in females.

- So we have two ways of thinking about β_3:
  1. either as modification of the effect of smoke by sex
  2. or the modification of the effect of sex by smoke.

Quick Lookup Table

We can draw up a table for the 4 types of observation:

	sex	smoke	$\text{logit}(\mathbb{P}(Y = 1))$
1	Male	Yes	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
2	Male	No	$\beta_0 + \beta_1$
3	Female	Yes	$\beta_0 + \beta_2$
4	Female	No	β_0

- This allows us to find the function of the parameters corresponding to a log-odds ratio and vice versa.
- e.g. 3 - 4 shows us that the log-odds ratio for smokers *vs.* non-smokers amongst females is β_2
- e.g. 1 - 2 shows us that the log-odds ratio for smokers *vs.* non-smokers amongst males is $\beta_2 + \beta_3$

Interaction Between a Dummy Variable and a Continuous Variable

- Consider a logistic model where the main predictors are sex (a dummy coded as before) and age (in years)

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \mathsf{sex} + \beta_2 \mathsf{age} + \beta_3 (\mathsf{sex} \times \mathsf{age})$$

- β_0 is the log-odds in favour of a female age 0 suffering from a heart attack.

Interaction Between a Dummy Variable and a Continuous Variable

- Consider a logistic model where the main predictors are sex (a dummy coded as before) and age (in years)

$$\operatorname{logit}\mathbb{P}(Y=1)=\beta_0+\underset{\text{Male}}{\beta_1}\overset{1}{\mathsf{sex}}+\beta_2\overset{0}{\mathsf{age}}+\beta_3(\overset{1}{\mathsf{sex}}\times\overset{0}{\mathsf{age}})$$

- β_1 is the log-odds ratio for males *vs.* females amongst people of age 0.

Interaction Between a Dummy Variable and a Continuous Variable

- Consider a logistic model where the main predictors are sex (a dummy coded as before) and age (in years)

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \overset{0}{\mathsf{sex}} + \beta_2 \overset{z}{\mathsf{age}} + \beta_3 (\overset{0}{\mathsf{sex}} \times \overset{z}{\mathsf{age}})$$

Female

- β_2 is the log-odds ratio corresponding to an increase in age by 1 year amongst females.

Interaction Between a Dummy Variable and a Continuous Variable

- Consider a logistic model where the main predictors are sex (a dummy coded as before) and age (in years)

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \overset{1}{\mathsf{sex}} + \beta_2 \overset{Z}{\mathsf{age}} + \beta_3 (\overset{1}{\mathsf{sex}} \times \overset{Z}{\mathsf{age}})$$

- β_3 is the difference between the log-odds ratio corresponding to a change in age by 1 year amongst males and the the log-odds ratio corresponding to an increase in age by 1 year amongst females.

- β_3 is also difference between the log-odds ratios for males *vs.* females in two age homogenous groups which differ by 1 year.

Quick Lookup Table

Again we can draw up a table, this time considering groups of individuals aged z and $z+1$

	sex	age	$\text{logit}(\mathbb{P}(Y=1))$
1	Male	$z+1$	$\beta_0+\beta_1+\beta_2(z+1)+\beta_3(z+1)$
2	Male	z	$\beta_0+\beta_1+\beta_2 z+\beta_3 z$
3	Female	$z+1$	$\beta_0+\beta_2(z+1)$
4	Female	z	$\beta_0+\beta_2 z$

- e.g. 3 - 4 shows us that the log-odds ratio corresponding to an increase in age by 1 year amongst females is β_2
- e.g. 2 - 4 shows us that the log-odds ratio for males *vs.* females amongst people aged z is $\beta_1+\beta_3 z$

Interaction Between 2 Continuous Variables

- Consider a logistic model where the main predictors are BP (blood pressure in mmHg) and age (in years)

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \mathsf{BP} + \beta_2 \mathsf{age} + \beta_3 (\mathsf{BP} \times \mathsf{age})$$

- β_0 is the log-odds in favour of a person with a BP of 0mmHg and age 0 suffering from a heart attack.

- Ridiculous interpretation (model can't apply when age or BP are close to 0, but we hope it is good for the ranges we are interested in.)

Interaction Between 2 Continuous Variables

- Consider a logistic model where the main predictors are BP (blood pressure in mmHg) and age (in years)

$$\operatorname{logit}\mathbb{P}(Y=1) = \beta_0 + \beta_1\mathsf{BP} + \beta_2\mathsf{age} + \beta_3(\mathsf{BP}\times\mathsf{age})$$

- β_1 is the log-odds ratio corresponding to an increase in BP by 1mmHg amongst people aged 0.

Interaction Between 2 Continuous Variables

- Consider a logistic model where the main predictors are BP (blood pressure in mmHg) and age (in years)

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \mathsf{BP} + \beta_2 \mathsf{age} + \beta_3 (\mathsf{BP} \times \mathsf{age})$$

- β_2 is the log-odds ratio corresponding to an increase in age by 1 year amongst people with a BP of 0mmHg.

Interaction Between 2 Continuous Variables

- Consider a logistic model where the main predictors are BP (blood pressure in mmHg) and age (in years)

$$\operatorname{logit} \mathbb{P}(Y = 1) = \beta_0 + \beta_1 \mathsf{BP} + \beta_2 \mathsf{age} + \beta_3 (\mathsf{BP} \times \mathsf{age})$$

- β_3 is the difference between the log-odds ratios corresponding to an increase in age of 1 year for two BP homogenous groups which differ by 1 mmHg.

- β_3 is also difference between the difference between the log-odds ratios corresponding to an increase in BP of 1 mmHg for two age homogenous groups which differ by 1 year.

Quick Lookup Table

Again we can draw up a table, this time considering individuals with *BP* w and $w+1$ and aged z and $z+1$

	BP	age	$\text{logit}(\mathbb{P}(Y=1))$
1	$w+1$	$z+1$	$\beta_0 + \beta_1(w+1) + \beta_2(z+1) + \beta_3(w+1)(z+1)$
2	$w+1$	z	$\beta_0 + \beta_1(w+1) + \beta_2 z + \beta_3(w+1)z$
3	w	$z+1$	$\beta_0 + \beta_1 w + \beta_2(z+1) + \beta_3 w(z+1)$
4	w	z	$\beta_0 + \beta_1 w + \beta_2 z + \beta_3 wz$

- e.g. 3 - 4 shows us that the log-odds ratio corresponding to an increase in age by 1 year amongst those of BP w is $\beta_2 + \beta_3 w$.
- e.g. 2 - 4 shows us that the log-odds ratio corresponding to an increase in BP by 1 mmHg amongst

Final Comment on Interpretation

- Remember whenever you give an interpretation of a quantity γ in terms of a log-odds ratio there is always an equivalent interpretation of $\exp(\gamma)$ as an odds-ratio.
- Whenever you give an interpretation of a quantity γ as the log-odds in favour of an event you can always give two equivalent interpretations
  1. of $\exp(\gamma)$ as the odds in favour of the event,
  2. of $\frac{\exp(\gamma)}{1+\exp(\gamma)}$ as the probability of the event.