MATH60604A
Statistical modelling
§2g - Interactions

Léo Belzile

HEC Montréal
Department of Decision Sciences

Interactions

- **Interactions:** combinations of covariates may affect the response differently than when taking in isolation.
- e.g., health premium are different if the person is a smoker (or not) versus and if he/she is obese (or not), but obese smokers also pay an extra premium.
- We say that the covariates X_1 and X_2 interact on Y when the effect of X_1 on Y depends on the value of X_2, and vice-versa.
- We consider the idealized fictious data `interaction` for the sake of illustration.

Interaction between a continuous and a binary variable

- We will only use two variables, `sex` and `fixation`, to model `intention`.
- The base model, without interaction, is

$$\texttt{intention} = \beta_0 + \beta_1 \texttt{sex} + \beta_2 \texttt{fixation} + \varepsilon,$$

where `sex` is a binary variable taking value unity for female and zero for male.

SAS code to fit a linear model

```
proc glm data=statmod.interaction;
class sex(ref="0");
model intention=sex fixation / ss3 solution;
run;
```

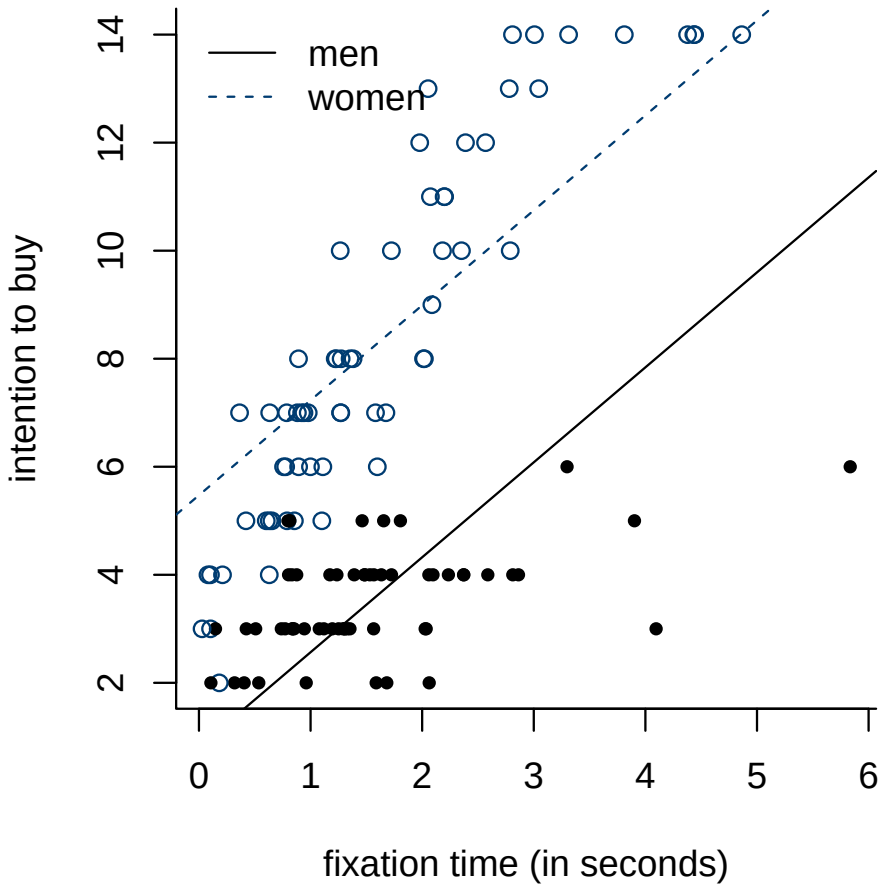
Interaction between a continuous and a binary variable

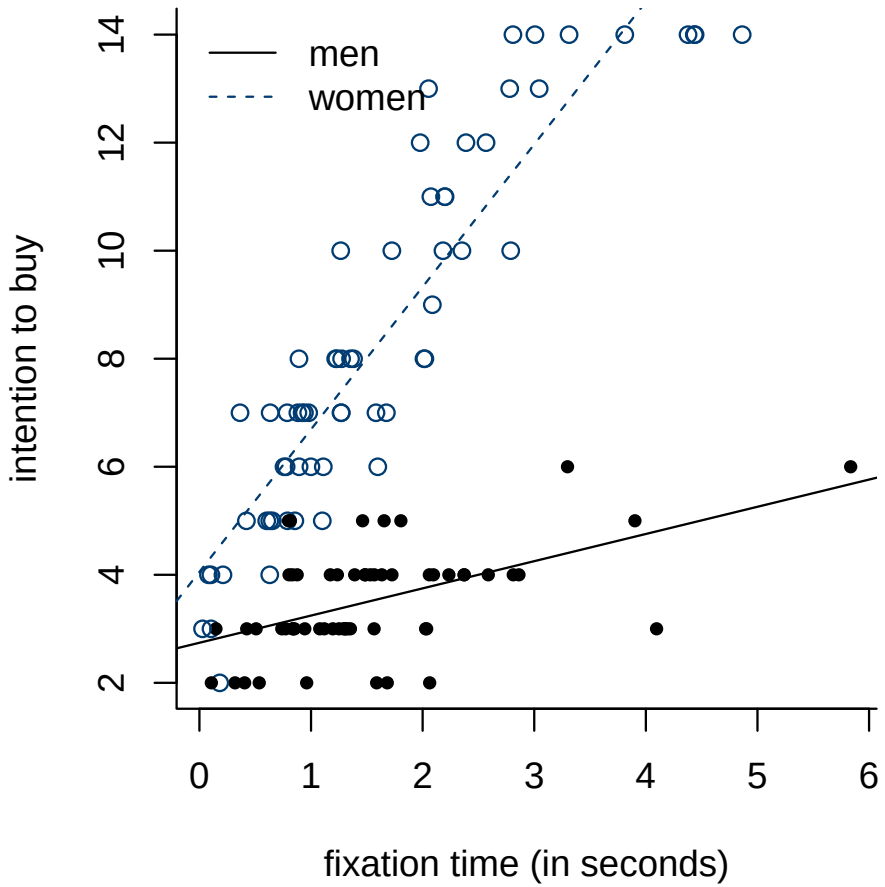
- This model includes no interaction between `fixation` and `sex`.
- The model assumes that the effect of the continuous variable `fixation` is the same for the two values of the binary variable.
- Likewise, the effect of the binary variable is assumed to be the same for all possible values of the continuous variable. We can see this on the plot, as the difference between the lines represents the effect of `sex`, is the same for all values of `fixation`; the lines are parallel.

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		0.807845274	B	0.30299531	2.67	0.0088
sex	**1**	4.664372686	B	0.30054877	15.52	<.0001
sex	**0**	0.000000000	B	.	.	.
fixation		1.758068738		0.13793124	12.75	<.0001

All parameters are statistically significant at level $\alpha = 0.05$.

Illustration of interaction between continuous and binary variable





Modelling interactions

- The previous figure shows that a better model would include a **different slope** for men and women.
- In order to add a different slope for men and women, we can create a new variable equal to the product $\texttt{fixation} \cdot \texttt{sex}$ and add it to the model,

$$\texttt{intention} = \beta_0 + \beta_1 \texttt{sex} + \beta_2 \texttt{fixation} + \beta_3 \texttt{fixation} \cdot \texttt{sex} + \varepsilon.$$

- Depending on the value of the binary variable $\texttt{sex}$, we get

$$\texttt{intention} = \begin{cases} (\beta_0 + \beta_1) + (\beta_2 + \beta_3)\texttt{fixation} + \varepsilon, & \text{if } \texttt{sex} = 1, \\ \beta_0 + \beta_2 \texttt{fixation} + \varepsilon, & \text{if } \texttt{sex} = 0. \end{cases}$$

- We recover the so-called main effect model when β_3 is zero.

Interaction between a continuous and a binary variable

SAS code to fit a linear model with an interaction

```
proc glm data=statmod.interaction;
class sex(ref="0");
model intention=sex fixation fixation*sex
    / ss3 solution;
run;
```

Parameter estimates of the interaction model

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	16.3846075	16.3846075	11.92	0.0008
fixation	1	340.0022198	340.0022198	247.44	<.0001
fixation*sex	1	156.9755099	156.9755099	114.24	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		2.741227810	B	0.28173140	9.73	<.0001
sex	**1**	1.311707333	B	0.37986044	3.45	0.0008
sex	**0**	0.000000000	B	.	.	.
fixation		0.503538480	B	0.15311557	3.29	0.0013
fixation*sex	**1**	2.134908186	B	0.19974154	10.69	<.0001
fixation*sex	**0**	0.000000000	B	.	.	.

Interaction between a continuous and a binary variable

- Testing whether the interaction is significant boils down to using the test $\mathrm{H}_0 : \beta_3 = 0$.
- If we reject H_0, then there is a significant interaction between the two variables (in this case, the p-value is less than 0.0001).
- The fitted model is
  - when $\texttt{sex} = 0$, we have $\mathrm{E}\left(\texttt{intention}\right) = 2.74 + 0.50\texttt{fixation}$;
  - when $\texttt{sex} = 1$, we have $\mathrm{E}\left(\texttt{intention}\right) = 4.05 + 2.64\texttt{fixation}$.
- The concept of interactions readily extends to categorical variables with k levels/categories.
  - In this case, we need to use the global F-test to check if the interaction is statistically significant.

Technical note

- The tests for `fixation` in the two tables are not the same because `fixation` is included in an interaction with a `class` variable.
- In the table of coefficients, the p-value corresponds to the t-test for the two-sided hypothesis $\mathrm{H}_0 : \beta_2 = 0$, i.e., the effect of `fixation` when `sex`=0.
- In the table above, the test is rather a test for the mean effect of `fixation`,

$$\mathrm{H}_0 : \{\beta_2 + (\beta_2 + \beta_3)\}/2 = 0$$

- These tests are **not** of interest. We cannot remove the main effect fixation unless we remove the interaction first.

Main effects and arbitrary choice of baseline for factors

- In the model with buying `intention` as a fonction of `sex` and `fixation` time, we would **not** remove the main effect of `fixation` while keeping the interaction term `fixation*sexe`, even if we fail to reject $\mathrm{H}_0 : \beta_2 = 0$.
- the parameter β_2 is the slope of `fixation` for men. Without it, the model would become

$$\texttt{intention} = \begin{cases} (\beta_0 + \beta_1) + \beta_3 \texttt{fixation} + \varepsilon, & \text{if } \texttt{sex} = 1, \\ \beta_0 + \varepsilon, & \text{if } \texttt{sex} = 0; \end{cases}$$

  this implies that intention to buy is constant for me, regardless of the fixation time.
- The choice of baseline is arbitrary, but changing the dummy `sex` (`0` for women, `1` for men), would yield a different model and so potentially different inferences.
- The baseline would not anymore be arbitrary!

Interactions between categorical variables

- For two categorical variables with respectively k_1 and k_2 levels, the interaction model has

$$k_1 k_2 = 1 + (k_1 - 1) + (k_2 - 1) + (k_1 - 1)(k_2 - 1)$$

  parameters — one for each combination.
- The number of restrictions to go from interaction model to main effect model is thus $(k_1 - 1)(k_2 - 1)$.
- The interpretation of the main effects are as before, i.e., they represent contrasts relative to a baseline, but the latter is level-dependent.
- We consider interaction terms **only if** the corresponding main effects are included.
- If the variance of the subgroup are equal, we can test the restriction using the F-statistic for global effects.

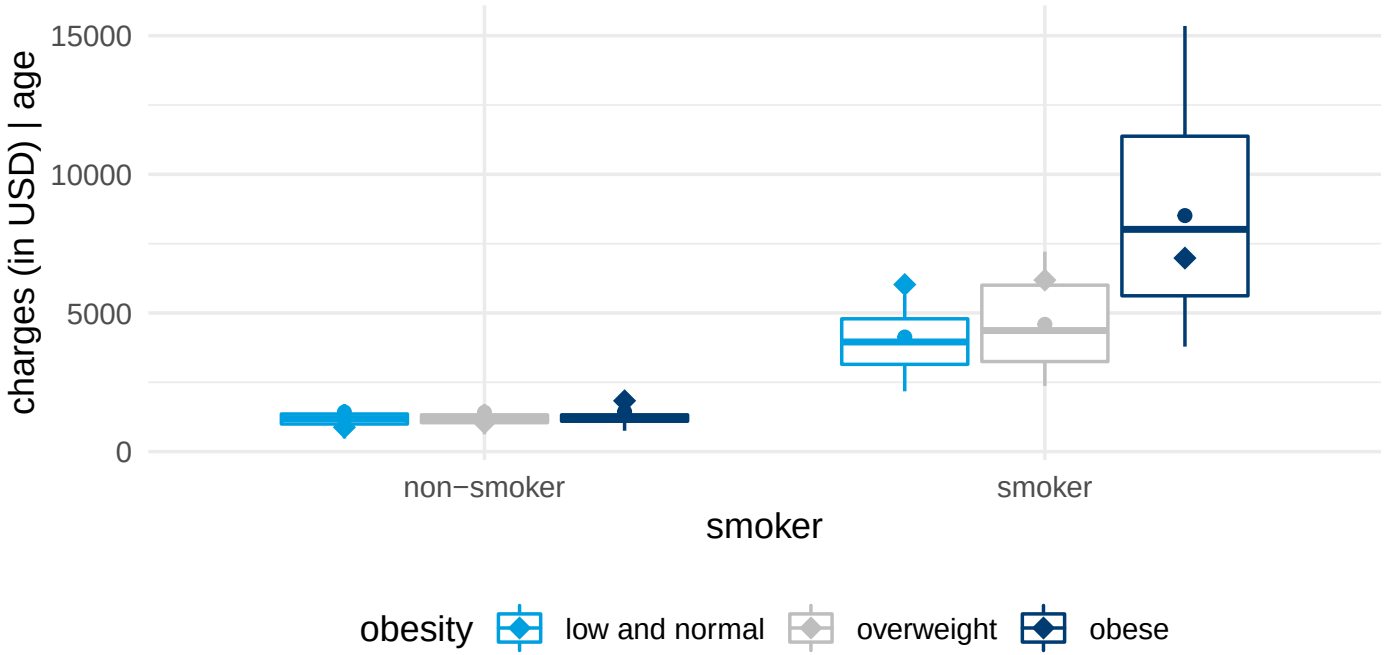
Interactions between categorical variables

- Consider a model for residual health insurance charges as a function of smoking and obesity indicators, after accounting for the effect of age.
- The fitted average for each group is based on the model

$$\texttt{rcharges} = \beta_0 + \beta_1 \texttt{smoker} + \beta_2 \texttt{obese}_1 + \beta_3 \texttt{obese}_2 + \varepsilon.$$

where $\texttt{obese}_1 = 1$ if $25 \leq \texttt{bmi} < 30$ (overweight) and $\texttt{obese}_2 = 1$ if $\texttt{bmi} \geq 30$ (obese).

Graphical representation of interactions between categorical variables



The diamonds indicate the fitted value for each group for the main effect, whereas the dots show the fitted values for the interaction model (mean of each group). The health insurance charges are clearly higher for obese smokers, something the main effect model fails to capture: it underpredicts the charges of obese smokers and overpredicts that of non-obese smokers.

Interaction model with two categorical variables

The linear model with interaction is

$$\texttt{rcharges} = \beta_0 + \beta_1 \texttt{smoker} + \beta_2 \texttt{obese}_1 + \beta_3 \texttt{obese}_2$$
$$\beta_4 \texttt{smoker} \cdot \texttt{obese}_1 + \beta_5 \texttt{smoker} \cdot \texttt{obese}_2 + \varepsilon.$$

The average charge for

- non-smokers with body mass index less than 25 is β_0;
- overweight non-smokers is $\beta_0 + \beta_2$;
- obese non-smokers is $\beta_0 + \beta_3$;
- obese smokers is $\beta_0 + \beta_1 + \beta_3 + \beta_5 \ldots$

Testing for the interaction amounts to $\mathrm{H}_0 : \beta_4 = \beta_5 = 0$.

Interpretation of parameters in the interaction model

The linear model with interaction is

$$\texttt{rcharges} = \beta_0 + \beta_1 \texttt{smoker} + \beta_2 \texttt{obese}_1 + \beta_3 \texttt{obese}_2 \\ \beta_4 \texttt{smoker} \cdot \texttt{obese}_1 + \beta_5 \texttt{smoker} \cdot \texttt{obese}_2 + \varepsilon.$$

- The interpretation is as before (but less straightforward…)
  - β_0, the baseline, is the mean charges of people whose body mass index (BMI) is less than 25 who do not smoke.
  - β_1 is the difference between the mean charges of smokers and non-smokers for individuals whose BMI is less than 25.
  - β_2 is the difference between the mean charges for overweight non-smokers and non-smokers whose BMI is less than 25.
  - β_3 is the difference between the mean charges of obese non-smokers and that of non-smokers whose BMI is less than 25.
  - $\beta_2 + \beta_4$ is the difference between the mean charges for overweight smokers and smokers whose BMI is less than 25.
  - $\beta_3 + \beta_5$ is the difference between the mean charges of obese smokers and that of smokers whose BMI less than 25.

Higher-order interaction

- In theory, we could consider an interaction between any number of variables. However, in practice we rarely go higher than third-order because it quickly becomes difficult to interpret the effects. Additionally, estimating an interaction between several variables requires a very large sample size.
- The basic principle is still the same. To create an interaction of a given order between several variables, we also need to include all the lower-order terms between the variables included in the higher-order interaction term.
- We interpret the variable effects while fixing the values of all the other variables in the interaction term.

Final remarks on interactions

- Don't remove a lower-order term, even if it's not significant. The lower-order terms are needed for proper inference!
- While it is tempting to include interactions between many categorical variables, beware of sparsely populated sub-categories.
- Algorithms performing model selection often base their variable choice on predictive performance.
  - removing lower-order term may not matter for the development of a black-box predictive model.
- However, removing main effects implies that the baseline is **not arbitrary** and inference is **invalid**.