Statistical modelling

#2.a Parameter interpretation in the linear model

Dr. Léo Belzile
HEC Montréal

Interpretation of coefficients of the mean model

We consider the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

where ε is a mean zero error term.

Interpretation of the parameters

- β_0 is the value when all of $X_1, \ldots, X_p$ are zero.
- β_j $(1 \leq j \leq p)$ is the mean change of Y when X_j increases by one unit, *ceteris paribus*.
  - provided no higher order terms or nonlinear functions of X_j, interactions, etc.

intention data

- In a study performed at Tech3Lab, subjects navigated a website that contained, among other things, an advertisement for candies.
- During the site navigation, an eye-tracker measured the location on the screen on which the subject's eyes were fixated.
- The tracker also recorded whether the subject looked at the ad and for how long it was in sight.
- A facial expression analysis software (FaceReader) was used to guess the subject's emotions when the ad was in sight.
- At the end of the study, a questionnaire measured the subject's intention to buy this type of candy and sociodemographic variables.

Study objectives

Evaluate whether

1. there is a link between the duration of fixation on the advertisement and the intention to buy and
2. whether perceived emotion is linked to the intention to buy.

Only the 120 subjects that had seen the ad in question are included in the data `intention`.

Data description

- `intention`: discrete variable ranging between `2` and `14`; larger values indicate higher interest in buying the product. Specifically, the score was constructed by summing the response of two questions, both measured using a Likert scale ranging from strongly disagree (`1`) to strongly agree (`7`).
- `fixation`: the total duration of fixation on the ad (in seconds).
- `emotion`: a measure of reaction during fixation; the ratio of the probability of showing a positive emotion to the probability of showing a negative emotion.

- `sex`: sex of subject, either man (`0`) or woman (`1`).
- `age`: age (in years).
- `marital`: civil status, either single (`0`) or in a relationship (`1`).
- `revenue`: categorical variable indicating the subject's annual income; one of (`1`) $[0, 20\,000]$; (`2`) $[20\,000, 60\,000]$; (`3`) $60\,000$ and above.
- `educ`: categorical variable indicating the highest educational achievement, either (`1`) high school or lower; (`2`) college; (`3`) university degree.

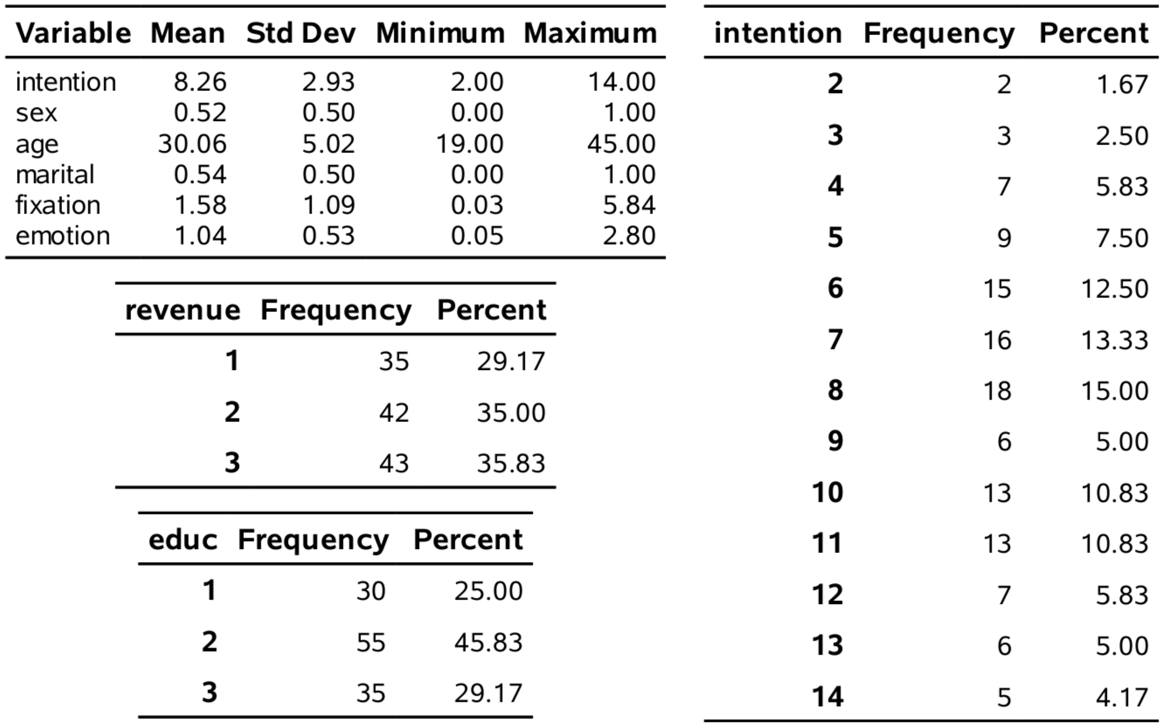
Exploratory data analysis

▪ SAS code ✚ SAS output (1) ✚ SAS output (2)

```
proc means data=statmod.intention mean std min max maxdec=2;
var intention sex age marital fixation emotion;
run;

proc freq data=statmod.intention;
tables intention revenue educ;
run;

 *Repeat this for other variables;
proc sgplot data=statmod.intention;
histogram intention emotion;
run;
```

Exploratory data analysis

SAS code SAS output (1) SAS output (2)



Variable	Mean	Std Dev	Minimum	Maximum
intention	8.26	2.93	2.00	14.00
sex	0.52	0.50	0.00	1.00
age	30.06	5.02	19.00	45.00
marital	0.54	0.50	0.00	1.00
fixation	1.58	1.09	0.03	5.84
emotion	1.04	0.53	0.05	2.80

revenue	Frequency	Percent
1	35	29.17
2	42	35.00
3	43	35.83

educ	Frequency	Percent
1	30	25.00
2	55	45.83
3	35	29.17

intention	Frequency	Percent
2	2	1.67
3	3	2.50
4	7	5.83
5	9	7.50
6	15	12.50
7	16	13.33
8	18	15.00
9	6	5.00
10	13	10.83
11	13	10.83
12	7	5.83
13	6	5.00
14	5	4.17

Exploratory data analysis

SAS code ✚ SAS output (1) ✚ SAS output (2)

Regression terminology

- **response variable** (Y): variable of interest
- **explanatory** variables, **covariates** or **predictors** ($\mathbf{X}$): the variables that are potentially associated with Y.

In our example,

- the response variable Y is `intention`;
- the explanatory variables are $\mathbf{X}$: `fixation`, `emotion`, `sex`, `age`, `revenue`, `educ`, `marital`.

We want to measure the effect of `fixation` and `emotion` on the `intention` to buy, while adjusting for sociodemographic variables.

Simple linear regression

Consider a linear model with `fixation` as only covariate.

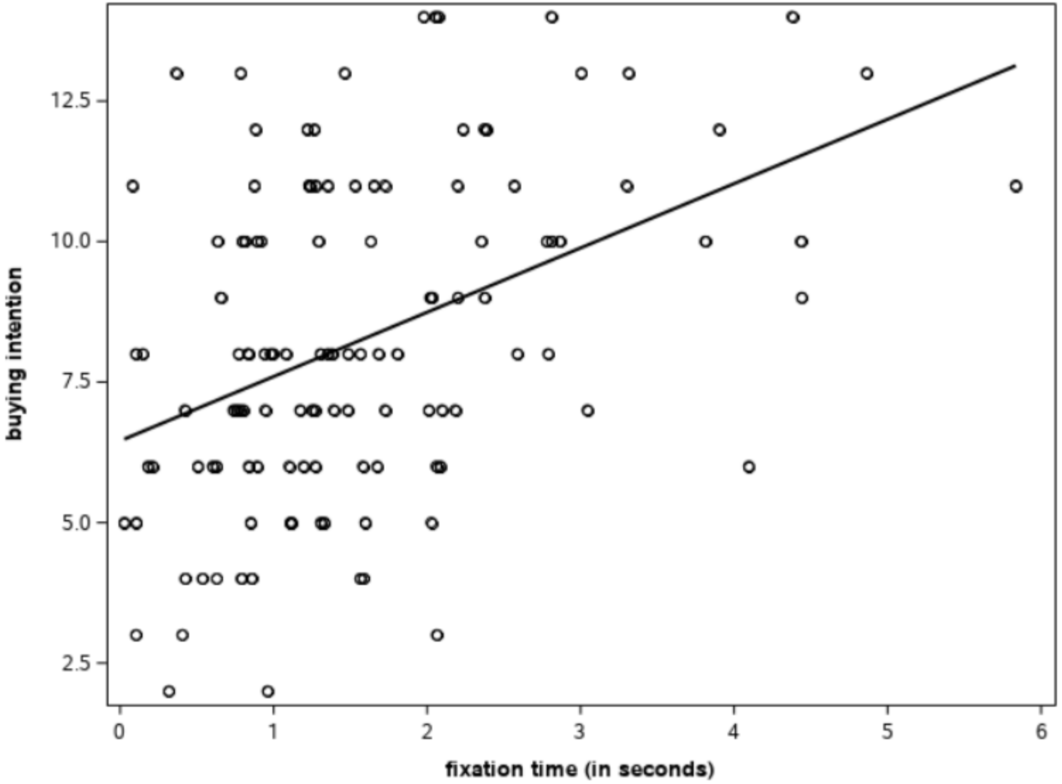
▪ SAS code ✚ Scatterplot ✚ Parameter estimates

```
proc sgplot data=statmod.intention noautolegend;
scatter y=intention x=fixation;
reg y=intention x=fixation;
yaxis label="buying intention";
xaxis label="fixation time (in seconds)";
run;

proc glm data=statmod.intention;
 *Only print coefficients table;
ods select ParameterEstimates;
model intention=fixation;
run;
```

Simple linear regression

Consider a linear model with `fixation` as only covariate.

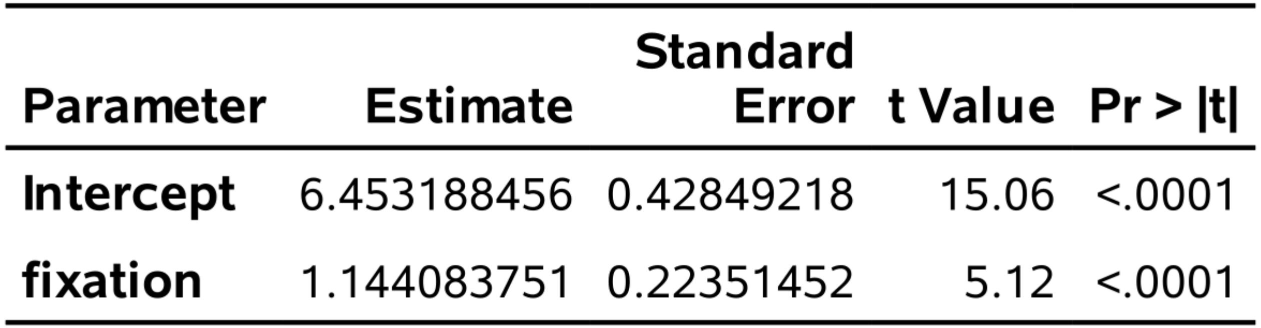
- SAS code ✚ Scatterplot ✚ Parameter estimates



Simple linear regression

Consider a linear model with `fixation` as only covariate.

SAS code + Scatterplot + **Parameter estimates**



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	6.453188456	0.42849218	15.06	<.0001
fixation	1.144083751	0.22351452	5.12	<.0001

The fitted regression line is

$$\widehat{\texttt{intention}} = 6.45 + 1.14\texttt{fixation}$$

Caveats?

Specification of categorical variables in SAS

- The **SAS** command `class` creates categorical variables that are interpreted as collection of indicators by the software.
- The baseline category is specified using `ref`.
- The **SAS** default is the first value encountered.
- In **R**, the analog is `factor`; the baseline is the first value in alphabetical or numerical order.

Binary explanatory variable

Consider a linear model with `sex` as only covariate.

▪ SAS code ✚ Parameter estimates ✚ Interpretation

```
proc glm data=statmod.intention;
ods select ParameterEstimates;
model intention=sex;
run;

/* If not coded 0/1, use class command */
proc glm data=statmod.intention;
class sex(ref="0");
model intention=sex / ss3 solution;
run;
```

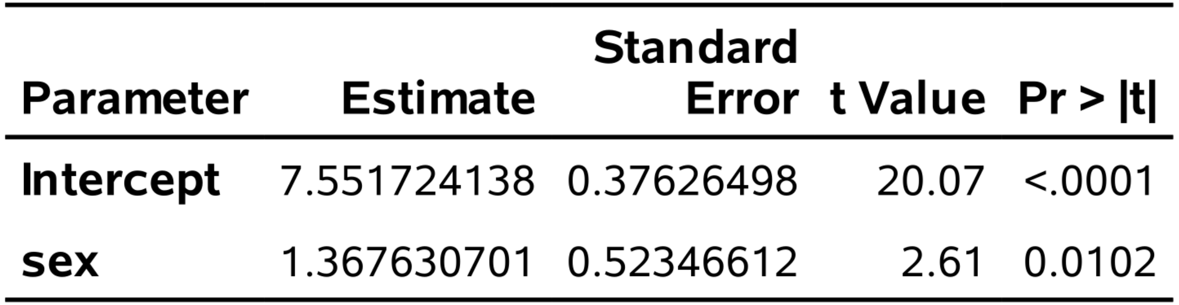
Binary explanatory variable

Consider a linear model with `sex` as only covariate.

SAS code | Parameter estimates | Interpretation

The postulated model is

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



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	7.551724138	0.37626498	20.07	<.0001
sex	1.367630701	0.52346612	2.61	0.0102

Binary explanatory variable

Consider a linear model with `sex` as only covariate.

▪ SAS code ✚ Parameter estimates ✚ Interpretation

- The mean intention to buy for men is 7.55 points
- The mean intention to buy for women is 8.92 points.
- The estimate of the slope is $\widehat{\beta}_1 = 1.37$, so the mean intention to buy score is 1.37 units higher for women than for men.

Categorical explanatory variables

- The variables `revenue` and `educ` are categorical, each with three levels.
- A categorical variable with k levels requires $k-1$ explanatory variables $\mathbf{x}$ in the model. For example, consider
  - `educ1 = 1` if `educ = 1` and zero otherwise.
  - `educ2 = 1` if `educ = 2` and zero otherwise.

Because the model includes an intercept, we don't need a third variable, since it would be redundant

educ	intercept	educ1	educ2
1	1	1	0
2	1	0	1
3	1	0	0

SAS code to fit the model with dummies

To fit the model, we include the two indicator variables in place of **`educ`**.

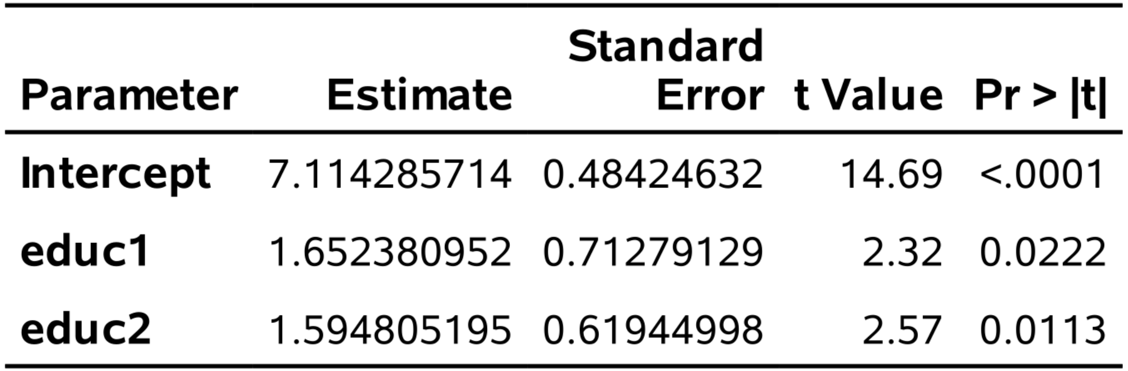
- SAS code (1) ✚ SAS output (1) ✚ SAS output (2)

```
data intention;
set statmod.intention;
educ1=(educ=1);
educ2=(educ=2);
run;

proc glm data=intention;
ods select ParameterEstimates;
model intention=educ1 educ2;
run;

 /* Alternative way with `class` */
proc glm data=statmod.intention;
ods select ParameterEstimates;
class educ(ref="3");
model intention=educ / ss3 solution;
run;
```

SAS code to fit the model with dummies

To fit the model, we include the two indicator variables in place of `educ`.

SAS code (1) + SAS output (1) + SAS output (2)

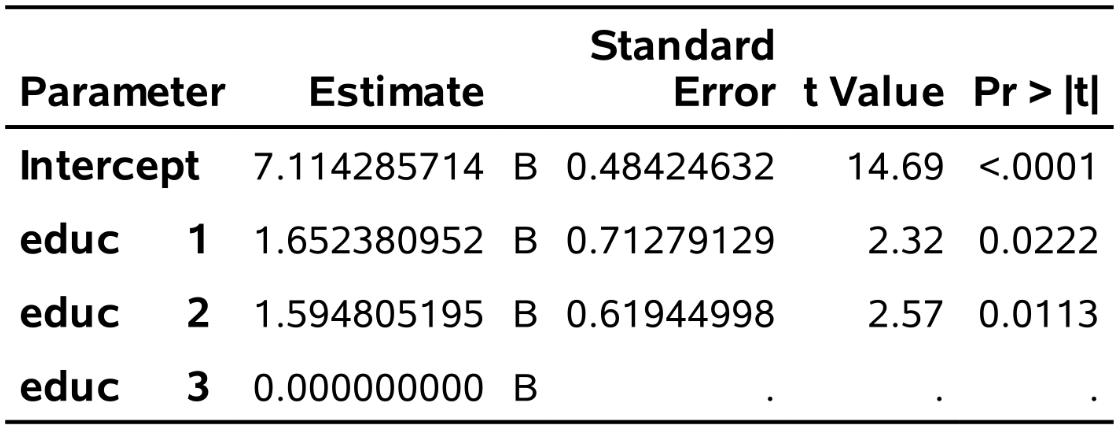


Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	7.114285714	0.48424632	14.69	<.0001
educ1	1.652380952	0.71279129	2.32	0.0222
educ2	1.594805195	0.61944998	2.57	0.0113

SAS code to fit the model with dummies

To fit the model, we include the two indicator variables in place of `educ`.

SAS code (1) + SAS output (1) + SAS output (2)



Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		7.114285714	B	0.48424632	14.69	<.0001
educ	1	1.652380952	B	0.71279129	2.32	0.0222
educ	2	1.594805195	B	0.61944998	2.57	0.0113
educ	3	0.000000000	B	.	.	.

The results are identical to those obtained by creating the indicator variables by hand.

Interpretation of the contrasts

- The estimated means of each of the three groups are $8.77, 8.71$, and 7.11 for education groups `1`, `2` and `3`, respectively.
- We can see that the mean of `intention` is 1.65 points higher for $\texttt{educ} = \texttt{1}$ than for $\texttt{educ} = \texttt{3}$, etc.
- To get the comparison between $\texttt{educ} = \texttt{1}$ and $\texttt{educ} = \texttt{2}$, we would need to refit the model after changing the reference category (exercise).

Comments about `class`

- In **SAS**, the levels of the categorical variable are case sensitive within `class`, e.g., `class rank(ref="AssistantProf")`
- **SAS** does not print the coefficient table if you use `class` unless the `/ solution` to the `model` call.