MATH60604A
Statistical modelling
§2d - Tests for linear models

Léo Belzile

HEC Montréal
Department of Decision Sciences

Hypothesis testing for individual parameters

- Suppose we want to know whether the effect of an explanatory variable is significant, i.e.,

$$\mathscr{H}_0 : \beta_j = 0, \qquad \mathscr{H}_1 : \beta_j \neq 0$$

- To test this two-sided hypothesis, we use the Wald statistic,

$$t = \frac{\widehat{\beta}_j - 0}{\mathsf{se}(\widehat{\beta}_j)},$$

where $\mathsf{se}(\widehat{\beta}_j)$ is the estimated **standard error** of $\widehat{\beta}_j$. Its null distribution is Student-t, reported in the SAS output under the label **t Value**.

Are people willing to pay more with credit cards?

Reference

Prelec, D. et Simester, D. (2001). Always Leave Home Without It: A Further Investigation of the Credit-Card Effect on Willingness to Pay. *Marketing Letters*, **12**, 5-12.

- **Study question**: Does paying by credit card encourage consumers to pay more?
- **Context**: In potentially expensive purchases, the amount that consumers are willing to spend when paying by credit card could be higher than that of consumers who pay by cash.
- **Objectives**: Present new evidence in favour of the fact that individuals are willing to spend more when paying by credit card.

Example: `tickets` data

- The product being sold: two tickets for the last regular season game of the Boston Celtics (NBA basketball team). This game was important, since it would decide who would end up in first place in the division.
- 64 subjects were **randomly** placed into the two groups:
  - Group 1 (33 subjects) who must pay by cash
  - Group 2 (31 subjects) who must pay by credit card
- Every subject had to fill out a questionnaire asking how much they were willing to pay for the two tickets.

Example: `tickets` data

- The simulated data corresponding to this study are found in the database `tickets`
- `offer`: amount offered for the tickets (in dollars) by an individual
- `group`: group assignment
  - `group=0` for individuals who pay by cash
  - `group=1` for individuals who pay by credit card

General objective

Compare the **mean** (amount of money individuals are willing to spend for the tickets) between the two groups (payment by either credit card or cash).

Two-sample t-test in a regression setting

- We want to test if the amount of money consumers are willing to spend is the same between the "cash" and the "credit card" groups.
- Testing if the two means are equal, is equivalent to testing whether the effect of the variable group on the variable offer is zero.
- We can formulate the problem through the regression model

$$\texttt{offer} = \beta_0 + \beta_1 \texttt{group} + \varepsilon$$

- Testing $\mathscr{H}_0 : \mu_{\texttt{cash}} = \mu_{\texttt{card}}$ is the same as testing $\mathscr{H}_0 : \beta_1 = 0$.

Linear regression with the `tickets` data

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	56.60606061	3.00335700	18.85	<.0001	50.60243550	62.60968571
group	15.00684262	4.31535073	3.48	0.0009	6.38057942	23.63310582

- We find that $\widehat{\beta}_1 = \$15$, i.e., those who pay by credit card are willing to pay $15 more **on average** than those who pay with cash.
- This difference is statistically significant at the 5% level (*p*-value of 0.0009).
- The linear regression output only reports two-sided tests, but the *p*-value for the one-sided test is half of the one reported (because of the symmetry of the Student-t distribution).

Confidence intervals

- In SAS, confidence intervals can be obtained by adding the option `clparm` to the `model` call. The default confidence level is $\alpha = 5\%$.
- The following code shows this for the simple linear regression model with the `intention` data.

$$\texttt{intention}_i = \beta_0 + \beta_1 \texttt{fixation}_i + \varepsilon_i.$$

SAS code to fit the regression model with the `glm` procedure

```
proc glm data=statmod.intention;
model intention=fixation / clparm;
run;
```

Tests for individual parameters and confidence intervals

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits
Intercept	6.453188456	0.42849218	15.06	<.0001	5.604657280 7.301719632
fixation	1.144083751	0.22351452	5.12	<.0001	0.701464150 1.586703353

- The t-statistic for $\mathscr{H}_0 : \beta_1 = 0$ against $\mathscr{H}_1 : \beta_1 \neq 0$ is $t = 1.144/0.224 = 5.12$.
- The p-value for the two-sided test is less than 0.0001.
- Therefore, we reject $\mathscr{H}_0$, as there is a significant effect of `fixation` on `intention`.
- The 95% confidence interval for β_1, corresponding to the linear effect of `fixation`, is $[0.70; 1.59]$.
- Since the confidence interval does not contain the value 0, we can conclude that the parameter is significantly different from zero (at level $\alpha = 5\%$).

Interpretation of the test

If we reject $\mathscr{H}_0 : \beta_j = 0$ in favor of the alternative, we rule out the possibility that there is no **linear relationship** between X_j and Y once the effect of other variables have been factored in.
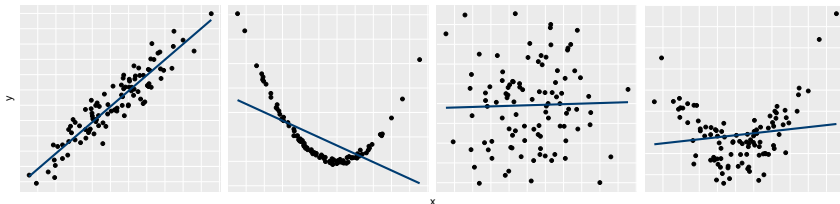
We reject $\mathscr{H}_0 : \beta_1 = 0$ (p-values less than 10^{-6}) in the two leftmost plots, but not in the two rightmost (p-values of 0.787 and 0.156). The coefficient of determination, from left to right, is 0.87, 0.3, 10^{-3} and 10^{-3}.

Hypothesis tests for significance of several variables

- There are also ways to test whether several β's are simultaneously equal to zero, e.g.,

$$\mathscr{H}_0 : \beta_{\texttt{age}} = \beta_{\texttt{emotion}} = 0.$$

- To see if we should include a categorical variable, we need to look at its global effect.
  - The variable educ has three levels and is modeled using two indicator variables educ1 and educ2; testing the global effect of educ amounts to testing

$$\mathscr{H}_0 : \beta_{\texttt{educ}_1} = \beta_{\texttt{educ}_2} = 0.$$

Comparison of nested models

- Consider the full model which contains p predictors,

$$\mathbb{M}_1 : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_g X_g + \beta_{k+1} X_{k+1} + ... + \beta_p X_p + \varepsilon.$$

- Suppose that we want to test

$$\mathscr{H}_0 : \beta_{k+1} = \beta_{k+2} = ... = \beta_p = 0.$$

- This hypothesis specifies that $(p - k)$ of the β parameters are zero.
- The restricted model contains only the covariates for which $\beta_j \neq 0$,

$$\mathbb{M}_0 : Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon.$$

Global *F*-test for testing effect of all covariates

- The first table output from the SAS `proc glm` procedure gives the results of the global test whether at least one of the variables in the model is useful for predicting the response variable Y.
- In the model that includes all possible variables, this corresponds to testing $\mathcal{H}_0 : \beta_{\mathtt{sex}} = \beta_{\mathtt{age}} = \cdots = \beta_{\mathtt{emotion}} = 0$.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	460.965432	51.218381	9.99	<.0001
Error	110	564.026235	5.127511		
Corrected Total	119	1024.991667			

- We reject $\mathcal{H}_0$, meaning that the model contains at least one variable useful for predicting intention to buy.

Sum of square decomposition

- It is important to note that model comparisons for continuous/categorical variables are done using **Type III sum of squares**.
  - this means that the reduced model includes all but one of the explanatory variables.
- In contrast, the Type I sum of squares table performs **sequential comparisons** with effect of covariates relative to those already included (in the order in which they entered)
  - e.g., if your first explanatory variable in the model statement is the variable of interest, the comparison is between a model with only this variable, compared to the intercept-only null model.
- To avoid confusion, add `ss3` in the SAS call.

Global F-test for categorical variables

- When the jth explanatory is continuous or binary, the F-test is the same as the t-test for $\beta_j = 0$. These variables have DF=1.
- When the variable is categorical (as defined through `class` in SAS), the output is different, e.g., the global test for a categorical variable `cat` with k levels corresponds to the hypothesis $\mathscr{H}_0 : \beta_{\mathtt{cat}_1} = \cdots = \beta_{\mathtt{cat}_{k-1}} = 0$
  - compare the model with and without the explanatory variable once other covariates are accounted for.

Suppose we fit a linear regression model with all the covariates for the intention data.

SAS code to fit the full linear model

```
proc glm data=statmod.intention noprint;
class sex educ revenue;
model intention= fixation emotion
    sex age revenue educ statut / ss3 solution;
run;
```

- The degrees of freedom for revenue and educ are equal to two, since each variable has three levels.
- e.g., the F-test compares the model with all covariates to the one with all explanatories, but education, i.e., $\beta_{\text{educ}_1} = \beta_{\text{educ}_2} = 0$.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	32.9762187	32.9762187	6.43	0.0126
age	1	37.1972654	37.1972654	7.25	0.0082
revenue	2	49.8556095	24.9278047	4.86	0.0095
educ	2	14.8380494	7.4190247	1.45	0.2397
marital	1	2.2349782	2.2349782	0.44	0.5105
fixation	1	184.7635934	184.7635934	36.03	<.0001
emotion	1	36.9122275	36.9122275	7.20	0.0084