MATH60604A
Statistical modelling
§2h - Collinearity

HEC Montréal
Department of Decision Sciences

Multicollinearity

- We say that two variables X_1 and X_2 are collinear if
  - X_1 and X_2 are both correlated with Y
  - X_1 and X_2 are strongly correlated with each other — so much so that they contain essentially the same information.
- There could be multicollinearity between more than two variables...in the same way that there could be more than one confounding variable.
- In such a case, multicollinearity (or simply collinearity) describes when an explanatory variable (or several) is strongly correlated with a linear combination of other explanatory variables.
- One potential harm of multicollinearity is a decrease in precision in parameter estimation, as it increases the standard errors of the parameters.

A stupid illustration of multicollinearity

- Consider the log number of Bixi rentals per day as a function of the temperature in degrees Celcius and in Farenheit, rounded to the nearest unit. The postulated linear model is

$$\texttt{lognuser} = \beta_0 + \beta_{\texttt{c}}\texttt{celcius} + \beta_{\texttt{f}}\texttt{farenheit} + \varepsilon.$$

- The interpretation of $\beta_{\texttt{c}}$ is "the average increase in number of rental per day when temperature increases by 1°C, keeping the temperature in Farenheit constant"...
- The two temperatures units are linearly related,

$$1.8\texttt{celcius} + 32 = \texttt{farenheit}.$$

- Suppose that the true effect (fictional) effect of temperature on bike rental is

$$\texttt{lognuser} = \alpha_0 + \alpha_1 \texttt{celcius} + \varepsilon.$$

- The coefficients for the model that only includes Farenheit are thus

$$\texttt{lognuser} = \gamma_0 + \gamma_1 \texttt{farenheit} + \varepsilon.$$

where $\alpha_0 = \gamma_0 + 32\gamma_1$ and $1.8\gamma_1 = \alpha_1$.
- The parameters of the postulated linear model with both predictors,

$$\texttt{lognuser} = \beta_0 + \beta_{\texttt{c}} \texttt{celcius} + \beta_{\texttt{f}} \texttt{farenheit} + \varepsilon,$$

are not **identifiable**, since any linear combination of the two solutions gives the same answer.

Bixi and multicollinearity

We consider a simple illustration with temperature at 16:00 in Celcius and Farenheit (rounded to the nearest unit for `rfarenheit`) to explain log of daily counts of Bixi users for 2014–2019.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	8.844327052	0.02819099	313.73	<.0001
celcius	0.048566261	0.00135205	35.92	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	7.980926861	0.05132678	155.49	<.0001
farenheit	0.026981256	0.00075114	35.92	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	8.844327052	B	0.02819099	313.73	<.0001
celcius	0.048566261	B	0.00135205	35.92	<.0001
farenheit	0.000000000	B	.	.	.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	9.555086770	1.14747585	8.33	<.0001
celcius	0.088592866	0.06461502	1.37	0.1706
rfarenheit	-0.022227045	0.03587330	-0.62	0.5356

SAS prints a warning if the data are exactly collinear.

> *Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.*

Effects of collinearity

Generally, collinearity has the following effects:

- The regression coefficients change drastically when new observations are included, or when we include/remove new covariates.
- The standard errors of the coefficients in the multiple regression model are very high, since the β cannot be precisely estimated.
- Consequently, the confidence intervals for these coefficients will be very wide.
- The individual parameters are not statistically significant, but the global F-test indicates some covariates are nevertheless relevant.

How do we detect collinearity or confounders?

- If the variables are exactly collinear, SAS or R will drop redundant ones.
  - The variables that are not **perfectly** collinear (e.g., due to rounding) will not be captured by software and will cause issues.
- Look at the **linear correlation** between explanatory variables and look at changes in estimated coefficients between regression models with and without a potential collinear variable.
- The problem is that, when more than two variables are collinear, detection is hard.
- One explanatory variable could be strongly correlated with a linear combination of other variables even though the individual correlations between the variables are not high.

Variance inflation factor

- Another tool we can use is the variance inflation factor (VIF); in SAS, use the option `vif` inside `proc reg`.
- For a given explanatory variable X_j, its VIF is

$$\mathrm{VIF}(j) = \frac{1}{1 - R^2(j)}$$

  where $R^2(j)$ is the R^2 of the model obtained by regressing X_j on all the other explanatory variables.
- The tolerance factor, $\mathrm{TOL} = 1 - R^2(j)$, is the reciprocal of `VIF`.

When is collinearity an issue?

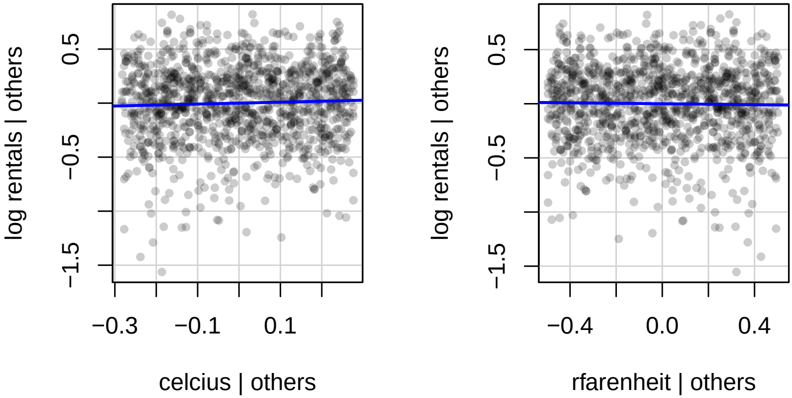
- $R^2(j)$ represents the proportion of the variance of X_j that is explained by all the other predictor variables.
- When is collinearity problematic? There is no general agreement, but practitioners typically choose an arbitrary cutoff (rule of thumb)

  - $\text{VIF}(j) > 4$ or $\text{TOL} < 0.25$ implies that $R^2(j) > 0.75$
  - $\text{VIF}(j) > 5$ or $\text{TOL} < 0.2$ implies that $R^2(j) > 0.8$
  - $\text{VIF}(j) > 10$ or $\text{TOL} < 0.1$ implies that $R^2(j) > 0.9$

Observations for Bixi multicollinearity example

- The value of the F statistic for the global significance for the simple linear model with Celcius (not reported) is 1292 with associated p-value less than 0.0001, suggesting that temperature is statistically significant (5% increase in number of users for each increase of 1°C).
- Yet, when we include both Celcius and Farenheit (rounded), the individual coefficients are not significant anymore at the 5% level.
- Moreover, the sign of `rfarenheit` change relative to that of `farenheit`!
- Note that the standard errors for Celcius are 48 times bigger when including the two covariates.
- The variance inflation factors of both `rfarenheit` and `celcius` are enormous (2454.68), suggesting identifiability issues.

Added-variable plots for Bixi multicollinearity example



Fictional example of multicollinearity

- We consider a fictional example with 100 observations on the outcome variable Y as well as five predictor variables X_1 through X_5.
- The Y values were actually randomly generated under the following model

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 + \varepsilon$$

- The parameter associated with each variable is 1.
- The data can be found in `simcollinear.sas7bdat`.

Fictional example of multicollinearity

- The correlation matrix between all the variables

SAS code for correlation

```
proc corr data=statmod.simcollinear noprob;
var y x1-x5;
run;
```

Pearson Correlation Coefficients, N = 100						
	Y	**X1**	**X2**	**X3**	**X4**	**X5**
Y **Y**	1.00000	0.45184	0.45549	0.64572	0.41047	0.34706
X1 **X1**	0.45184	1.00000	0.05607	0.68896	0.14553	0.01874
X2 **X2**	0.45549	0.05607	1.00000	0.64534	0.07247	-0.02981
X3 **X3**	0.64572	0.68896	0.64534	1.00000	0.15883	0.00667
X4 **X4**	0.41047	0.14553	0.07247	0.15883	1.00000	0.11266
X5 **X5**	0.34706	0.01874	-0.02981	0.00667	0.11266	1.00000

Simple regression for the fictional example of multicollinearity

- The correlation between Y and each predictor variable is significant and positive.
- Consequently, if we fit a separate model for each predictor variable, the parameter for each variable would be significant and positive for each one. This is consistent with the true model from which we simulated the data.
- This shows that there are enough observations to estimate the parameters, and to make proper conclusions about their effects, at least when considering one predictor at a time.

Fictional example of multicollinearity

- However, X_1, X_2 et X_3 are highly correlated with each other which could cause multicollinearity problems.
- We fit the model containing all the predictor variables with `proc reg`, while requesting multicollinearity diagnostics.

SAS code to compute variance inflation factor

```
proc reg data=statmod.simcollinear;
model y=x1-x5 / vif;
run;

proc glm data=statmod.simcollinear;
model y=x1-x5 / ss3 solution tolerance;
run;
```

SAS procedures: `reg` versus `glm`

The `glm` procedure doesn't include an option to compute the `vif`; one can either use `tol` (reciprocal of VIF) or else resort to the `reg` procedure for fitting linear models.

- By default, diagnostic plots are produced by `reg` (`plots=diagnostics` option with `glm`).
- The `reg` procedure includes more model selection diagnostics (not used in inference).
- The table of coefficients is automatically printed by `reg` (`solution` option in `glm`).
- The main drawback of `reg` is that it doesn't handle categorical variables: these must be manually coded using binary indicators (`0-1`) (**frequent programming mistake**).

Parameter estimates and VIF

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-0.76110	2.43241	-0.31	0.7551	0
X1	X1	1	0.43149	0.45829	0.94	0.3488	3.75609
X2	X2	1	0.68894	0.45638	1.51	0.1345	3.38306
X3	X3	1	1.94048	0.77306	2.51	0.0138	6.42789
X4	X4	1	1.06329	0.24587	4.32	<.0001	1.04162
X5	X5	1	1.14430	0.23231	4.93	<.0001	1.01507

Dependent Variable: Y

Tolerances

Variable	Type I Tolerance	Type II Tolerance
Intercept	100	6.3051461638
X1	1	0.2662342154
X2	0.9968564885	0.2955903718
X3	0.1560669286	0.1555721533
X4	0.9722559013	0.9600474856
X5	0.9851577945	0.9851577945

Goodness of fit and model summary

- Overall, the model seems adequate. The R^2 is 62%.
- However, the variables X_1 and X_2 are no longer significant once other explanatories are accounted for.
- The VIF of X_3 is quite large (6.43) and the variance inflation factors for X_1 and X_2 are between 3 and 4.
- **This indicates a possible problem of collinearity**. The estimation precision for these parameters is not as good as it would be if there were no multicollinearity.
- Note that the VIF is an individual measure. It does not tell us which particular variables are correlated with each other.