Statistical modelling

#2.b Linear transformations

Dr. Léo Belzile
HEC Montréal

Linear transformations

Consider the log number of Bixi rentals per day as a function of the temperature in degrees Celcius (or in Farenheit).

Suppose that the true effect of temperature on log of bike rentals is

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

- The interpretation of α_1: *the average increase in the number of log rental per day when temperature increases by* $1^{\circ}\mathrm{C}$.

The model for log-rentals with temperature expressed in Farenheits is

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

SAS output

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

The two units are **linearly** related,

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

so we find that $\alpha_0 = \gamma_0 + 32\gamma_1$ and $\alpha_1 = 1.8\gamma_1$.

Uniqueness of the solution

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 give the same fitted values.

For $k \in \mathbb{R}$, $\beta_0 = k\alpha_0 + (1-k)\gamma_0$, $\beta_1 = k\alpha_1$ and $\beta_2 = (1-k)\gamma_1$ are equivalent.

The rank of $\mathbf{X}$ is 2, but the design matrix has 3 columns

- $\mathbf{X}^\top\mathbf{X}$ is not invertible.
- the solution to the normal equation is **not unique**.

Collinearity

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	.	.	.

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.`