Linear Regression (LR)

1 Simple to complex models

1.1 Simple Linear Regression

- Definition
  Dependent variable is a scalar. Independent variable is a scalar.

- Equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i, i = 1, \ldots, n \tag{1}$$

- Solution (Ordinary Least Square)

$$Q = \sum_i [y_i - (\beta_0 + \beta_1 x_{i1})]^2 \tag{2}$$

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_{i1}) = 0 \tag{3}$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_{i1}) x_{i1} = 0 \tag{4}$$

$$n\beta_0 + n\overline{\boldsymbol{x_1}}\beta_1 = n\overline{\boldsymbol{y}} \tag{5}$$

$$n\overline{\boldsymbol{x_1}}\beta_0 + (\sum_i x_{i1}^2)\beta_1 = \sum_i x_{i1} y_i \tag{6}$$

$$\hat{\beta}_0 = \overline{\boldsymbol{y}} - \hat{\beta}_1 \overline{\boldsymbol{x_1}} \tag{7}$$

$$\hat{\beta}_1 = \frac{\sum_i (x_{i1} - \overline{\boldsymbol{x_1}})(y_i - \overline{\boldsymbol{y}})}{\sum_i (x_{i1} - \overline{\boldsymbol{x_1}})^2} \tag{8}$$

1.2 Multiple Linear Regression

- Definition
  Dependent variable is a scale. Independent variable is a vector.

- Equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, i = 1, \ldots, n \tag{9}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{10}$$

- Solution (Ordinary Least Square)

$$Q = \sum_i [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2 \tag{11}$$

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}) = 0 \tag{12}$$

$$\frac{\partial Q}{\partial \beta_j} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}) x_{ij} = 0, j = 1, 2, \ldots, p \tag{13}$$

$$X^T X \boldsymbol{\beta} = X^T Y \tag{14}$$

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y \tag{15}$$

- Special case with two independent variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, i = 1, \ldots, n \tag{16}$$

$$\hat{\beta}_0 = \overline{\boldsymbol{y}} - \hat{\beta}_1 \overline{\boldsymbol{x_1}} - \hat{\beta}_2 \overline{\boldsymbol{x_2}} \tag{17}$$

$$\hat{\beta}_1 = \frac{(\overline{yx_1} - \bar{y}\bar{x}_1)(\overline{x_2^2} - \bar{x}_2^2) - (\overline{yx_2} - \bar{y}\bar{x}_2)(\overline{x_1 x_2} - \bar{x}_1 \bar{x}_2)}{(\overline{x_1^2} - \bar{x}_1^2)(\overline{x_2^2} - \bar{x}_2^2) - (\overline{x_1 x_2} - \bar{x}_1 \bar{x}_2)^2} \tag{18}$$

$$\hat{\beta}_2 = \frac{(\overline{x_1^2} - \bar{x}_1^2)(\overline{yx_2} - \bar{y}\bar{x}_2) - (\overline{yx_1} - \bar{y}\bar{x}_1)(\overline{x_1 x_2} - \bar{x}_1 \bar{x}_2)}{(\overline{x_1^2} - \bar{x}_1^2)(\overline{x_2^2} - \bar{x}_2^2) - (\overline{x_1 x_2} - \bar{x}_1 \bar{x}_2)^2} \tag{19}$$

If we substitute x_1 to be $x_1 + a$, $\hat{\beta}_1, \hat{\beta}_2$ remains the same but $\hat{\beta}_0$ will change.
If we substitute x_1 to be ax_1, $\hat{\beta}_0, \hat{\beta}_2$ remains the same but $\hat{\beta}_1$ will be $\hat{\beta}_1/a$.

- Matlab code snippet

```
clear

rng('default')
rng(1)

% N: number of subjects
N = 100;
% p: number of independent variables
p = 4;

% multiple linear regression and OLS solution
y = randn(N, 1);
X = [ones(N, 1), randn(N, p)];
b = (X'*X)\X'*y;
disp('original:')
disp(b)
```

```
% shift the 1st independent variable
X_shift = X;
X_shift(:, 2) = X_shift(:, 2) + 5;
b_shift = (X_shift'*X_shift)\X_shift'*y;
disp('shift the first independent variable by 5:')
disp(b_shift)

% scale the 1st independent variable
X_scale = X;
X_scale(:, 2) = X_scale(:, 2)*5;
b_scale = (X_scale'*X_scale)\X_scale'*y;
disp('scale the first independent variable by 5:')
disp(b_scale)
```

OUTPUT

original: -0.0582 -0.1065 0.0473 -0.0799 -0.1551

shift the first independent variable by 5: 0.4744 -0.1065 0.0473 -0.0799 -0.1551

scale the first independent variable by 5: -0.0582 -0.0213 0.0473 -0.0799 -0.1551

1.3 Multivariate Linear Regresssion OR General Linear Model

- Definition
  Dependent variable is a vector. Independent variable is a vector.

- Equation

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{i1} + \cdots + \beta_{pj}x_{ip} + \epsilon_{ij}, i = 1, \ldots, n, j = 1, \ldots, m \tag{20}$$

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{U} \tag{21}$$

- Solution
  Similar to multiple linear regression.

2 Questions

2.1 When should we do centering and standardization for X?

The following answer is from

https://stats.stackexchange.com/questions/19216/variables-are-often-adjusted-e-g-standardised-

Standardization is all about the weights of different variables for the model. If you do the standardisation "only" for the sake of numerical stability, there may be transformations that yield very similar numerical properties but different physical meaning that could be much more appropriate for the interpretation. The same is true for centering, which is usually part of the standardization.

Situations where you probably want to **standardize**:

- the variables are different physical quantities
- and the numeric values are on very different scales of magnitude
- and there is no "external" knowledge that the variables with high (numeric) variation should be considered more important.

Situations where you may **not** want to standardize:

- if the variables are the same physical quantity, and are (roughly) of the same magnitude, e.g. relative concentrations of different chemical species, absorbances at different wavelengths, emission intensity (otherwise same measurement conditions) at different wavelengths
- you definitively do not want to standardize variables that do not change between the samples (baseline channels) - you'd just blow up measurement noise (you may want to exclude them from the model instead)
- if you have such physically related variables, your measurement noise may be roughly the same for all variables, but the signal intensity varies much more. I.e. variables with low values have higher relative noise. Standardizing would blow up the noise. In other words, you may have to decide whether you want relative or absolute noise to be standardized.
- There may be physically meaningful values that you can use to relate your measured value to, e.g. instead of transmitted intensity use percent of transmitted intensity (transmittance T).

You may do something "in between", and transform the variables or choose the unit so that the new variables still have physical meaning but the variation in the numerical value is not that different, e.g.

- if you work with mice, use body weight g and length in cm (expected range of variation about 5 for both) instead of the base units kg and m (expected range of variation 0.005 kg and 0.05 m - one order of magnitude different).
- for the transmittance T above, you may consider using the absorbance A=-log10T

Similar for centering:

- There may be (physically/chemically/biologically/...) meaningful baseline values available (e.g. controls, blinds, etc.)
- Is the mean actually meaningful? (The average human has one ovary and one testicle)