10. Regression

Outline

- Definition
- Example applications
- Least squares
- Solving linear problems
- Diagnostics (R^2, adjusted R^2, AIC)

Definition

- Fit a function of (some) given type, say f, *approximately* through given points (x_i, y_i) so that for the given points, $f(x_i) \approx y_i, i = 1, 2, \ldots n$
- This function is usually intended to predict y given a predictor value x (this x can actually be a vector that contains several different relevant predictors)

Example applications

- Forecasting travel time for a route based on predictors such as time of day, day of week, closeness to holiday, amount of construction . . .
- Forecasting transit ridership based on similar factors, and perhaps ones that change more slowly such as population density, amount of commercial space, and economic activity
- Estimating house prices from predictors such as size, age, location, . . .

Example problem

Given the following data on u, v, z values, develop a function $f(u, v)$ that we can use to predict z where it's not known.

u	2	1	6	0	2	1
v	4	1	3	1	0	14
z	11	3	26	3	10	8

Least squares

- If the points are grouped as vectors $\boldsymbol{x}, \boldsymbol{y}$, the *residual* vector of the fitted function is $\boldsymbol{r} = \boldsymbol{y} - f(\boldsymbol{x})$ (i.e. $r_i = y_i - f(x_i)$)
- We'd like $\boldsymbol{r}$ to be close to all zeros (which would be the case for interpolation, where $y_i = f(x_i)$)
- Problem: Out of all the functions f in the given type, find the one that has the smallest the residual sum of squares, RSS $= \boldsymbol{r}^T \boldsymbol{r} = \sum_{i=1}^{n} r_i^2$

Linear least squares

- Suppose the function type is such that we can write $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{\beta}$ (linear regression)
  - $\boldsymbol{A}$ is a known $n \times m$ *design matrix* for the given $\boldsymbol{x}$
  - $\boldsymbol{\beta}$ is an unknown $m \times 1$ vector of 'parameters'
  - So $f(x_i) = A_{i,1}\beta_1 + A_{i,2}\beta_2 + \ldots A_{i,m}\beta_m$
- Then we can find using linear algebra methods the value of $\boldsymbol{\beta}$ that minimizes the RSS for the given $\boldsymbol{x}$ and $\boldsymbol{y}$

The normal equations

$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{A}^T\boldsymbol{y}$ is the system of *normal equations* for a linear regression problem. Solving it for $\boldsymbol{\beta}$ gives us the least squares (lowest RSS) set of parameter values.

Example problem

Suppose our function type is $f(u,v) = \beta_1 u + \beta_2 \sqrt{v}$. This is linear in the unknown β_i, so we can write $f(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{A}\boldsymbol{\beta}$, where the design matrix $\boldsymbol{A}$ is $\begin{pmatrix} u_1 & \sqrt{v_1} \\ u_2 & \sqrt{v_2} \\ u_3 & \sqrt{v_3} \\ u_4 & \sqrt{v_4} \\ u_5 & \sqrt{v_5} \\ u_6 & \sqrt{v_6} \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 1 & 1 \\ 6 & \sqrt{3} \\ 0 & 1 \\ 2 & 0 \\ 1 & \sqrt{14} \end{pmatrix}$

The normal equations are then

$$\begin{pmatrix} 46 & 19.134 \\ 19.134 & 23 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 209 \\ 102.97 \end{pmatrix}$$

Resulting in $\boldsymbol{\beta} = \begin{pmatrix} 4.100 \\ 1.066 \end{pmatrix}$

Diagnostics

- These are measures of how well a function fits given y values, meant to give an indication of how good it might be as a predictor
- $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$, where the total sum of squares TSS is $(\boldsymbol{y} - \bar{y})^T(\boldsymbol{y} - \bar{y})$, with $\bar{y}$ the average value of $\boldsymbol{y}$
- Adjusted R^2 : $R_a^2 = 1 - \frac{n-1}{n-m}\frac{\text{RSS}}{\text{TSS}}$, where n is the number of data points and m is the number of unknown parameters in $\boldsymbol{\beta}$
- Akaike information criterion (AIC): $n \log(\mathrm{RSS}/n) + 2mn/(n-m-1)$ (lower value denotes better fit)
- R_a^2 and AIC include m as well as RSS in their formulas to reflect that all else being equal, a more complicated function type (with more parameters that need to be determined) will not be as good at predicting unknown values

Example

- For the function type $f(u, v) = \beta_1 u + \beta_2 \sqrt{v}$, the least-squares $\boldsymbol{\beta}$ give RSS $= 12.3$, $R^2 = 0.966$, $R_a^2 = 0.957$, AIC $= 12.3$
- Changing the function type to one with another parameter with $g(u, v) = \beta_1 u + \beta_2 \sqrt{v} + \beta_3$, gives RSS $= 10.4$, $R^2 = 0.971$, $R_a^2 = 0.952$, AIC $= 21.3$
- Although R^2 improves with the additional parameter, R_a^2 and AIC both worsen, implying that this function type may not be better at predicting y given the points available for fitting

Nonlinear least squares

- The *regression model* or function type, $f(\boldsymbol{\beta}; x)$, may also be nonlinear in $\boldsymbol{\beta}$ – for example, it might look like x^{β} or $\cos(\beta x)$.
- In that case, we can still look for the least-squares values of β, but we typically need to use iterative numerical methods to approximate it, instead of just solving a linear system.
- Once the least-squares value of β is found, R^2 and other diagnostics can be calculated following the same formulas as before.