

STATS 3001 / STATS 4104 / STATS 7054

Statistical Modelling III

Practical 4 - Logistic regression

Week 7

GOAL

The purpose of this practical is to explore the application of the `glm` function to logistic regression.

OVERVIEW

- To enter and prepare data for logistic regression analysis in R.
- To perform a logistic regression in R and interpret the output.
- To extract estimated probabilities from a logistic regression.

DATA

The incidence of non-melanoma skin cancer among women in Minneapolis-St Paul, Minnesota, and Dallas-Fort Worth, Texas was recorded in a study.

The data are available in the file `skin.xlsx`.

STEPS

1. Load data into R.
2. Add column that is the proportion of cases for each row.
3. Plot proportion against age with colour of plots for each town. Describe the relationships, do the relationships make sense?
4. Fit the following logistic regression models

$$M_0 : \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \text{constant}$$

$$M_1 : \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \text{Town}_i$$

$$M_2 : \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \text{Age}_i$$

$$M_3 : \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \text{Age}_i + \text{Town}_i$$

5. Choose the best model.
6. For you final model, give an interpretation of the coefficient `TownSt Paul`.
7. Check the assumptions of the model.
8. Use the model to predict the probability of skin cancer for a 51 year old living in Texas.