Introduction: Applied Regression [1]

Emilija Perković

Dept. of Statistics
University of Washington

[1] (based on lectures of Elena Erosheva)

Basics

- Syllabus: Text, schedule, grades.
- Lectures and lab sessions, laptops, homework.
- Project.

Course Projects

- STAT/CSSS 504 is a project-based course.
- Students identify a research question and a corresponding data set, and carry out a regression analysis to answer the research question.
- Everyone proposes a project idea. Instructor selects projects. Best projects start with a question or idea, then find data.
- If not enough viable project ideas are proposed, there will be an in-class final exam and the class schedule will be revised accordingly.
- Project groups deliver short oral presentations in class.
- Project groups present final results in poster format during the finals week.

Objectives

- To gain statistical background necessary to understand regression analysis.
- To gain practical skills necessary to formulate a research question, carry out analyses, interpret results, and present findings addressing the research question from a regression study.
- To become a critical consumer of research that employs regression techniques.

Topics

The course will cover:

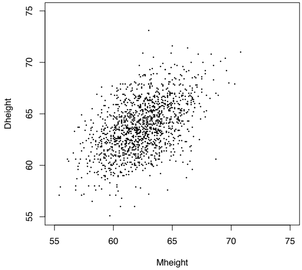
- Basic and multiple linear regression.
- Estimation methods (maximum likelihood, least squares, weighted least squares).
- Interpretation.
- Categorical independent variables, interactions.
- Violations of assumptions. Remedies.
- Model selection.
- Robust regression.
- Logistic regression.

Basic ideas of regression

- Regression is by far the most frequently used statistical model.
- Regression is used for explaining or modeling the relationship between a quantitative variable Y, called the **dependent** variable, and one or more **independent** variables, $X_1, \ldots, X_p$.
- When $p = 1$, the analysis is called **simple** regression; when $p > 1$, it is called **multiple** regression.

Inheritance of Heights: Mothers and Daughters

From R package `alr4` dataset `Heights`.



Basic ideas of regression

Other names for the X variables are:

- predictor
- input
- explanatory variable

Other names for Y are:

- response
- output
- outcome

Linear regression

A very general form for linear regression of Y on $\mathbf{X} = \{X_1, \ldots X_p\}$ is

$$y = f(\mathbf{x}) + \epsilon,$$

where

- y is the observed response of variable Y,
- $\mathbf{x} = \{x_1, \ldots, x_p\}$ are observed values of predictors $\mathbf{X} = \{X_1, \ldots, X_p\}$,
- $f(\mathbf{x})$ is a function of $x_1, \ldots, x_p$, linear in parameters (coefficients).
- ϵ is the error term with mean zero and some variance σ^2.

Note: the function f is not necessarily linear in the predictors!

Linear regression

Which equation is not a linear regression function?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$
$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_3 \log(x_1) x_2 + \beta_4 x_2^2 + \epsilon$$
$$y = \beta_0 + \beta_1 x_1 + x_2^{\beta_2} + \epsilon$$

Linear regression

The error term ϵ represents deviations from an exact linear relationship between Y and $\mathbf{X}$. This may be due to:

- measurement error on both $\mathbf{X}$ and Y,
- unobserved variables that also affect Y,
- deviations of the true relationship from linearity,
- rounding errors on $\mathbf{X}$ and Y,
- inherent randomness (unpredictable aspects of Y).

Regression objectives

1. **Prediction of future observations.**
    - Can we predict time to the next eruption by the duration of the current eruption of a geyser?
    - Can we predict the actual number of geese in a flock by using a visual estimate of a Wildlife Service member?

2. **Description and Inference:** Assessment of the relationship between explanatory variables and the response.
    - What is the relationship between mothers' and daughters' heights?
    - What is the relationship between education and voting Democrat?

Regression objectives

Note that the **prediction and decision-making** objectives are characteristic of problems where understanding the mechanism is important only to the extent that it aids better prediction.

On the other hand, **description and inference** objectives are characteristic of problems where understanding the mechanism is the key issue and predictions are by-products.

Before you fit a regression model

Examine appropriate:

- numerical summaries (min, max, cor, st deviation, etc),
- univariate graphical summaries (boxplots, histograms, density plots),
- scatterplots.

Example: Forbes Data

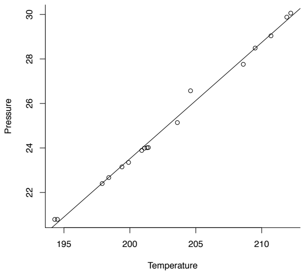
Forbes Data: Data on the relationship between atmospheric pressure and the boiling point of water were collected in the Alps and in Scotland.

The pressure was measured with a barometer (in inches of mercury) and the boiling point was measured using a thermometer (in F), at each location ($n = 17$).

Assuming we have already examined numerical summaries and univariate plots, let us look at the scatterplots.

Data: `Forbes` in R package `alr4`.

Example: Forbes data

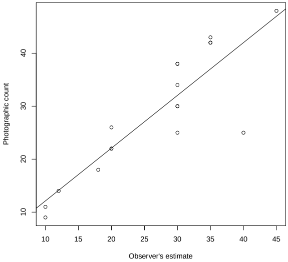


Example: Forbes data

Forbes Data: Observations from the scatterplot.

- Points appear to lie close to a line, however some curvature can be seen (by theory, log(pressure) is linearly related to temperature).
- One point does not “fit”.

Example: Snow Geese



Data: `snowgeese` in R package `alr3`.

Example: Snow Geese

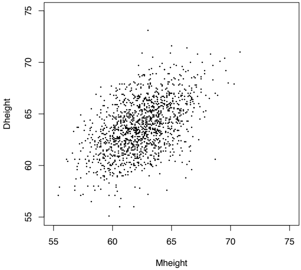
Snow geese: Observations from the scatterplot.

- Small sample size; some x values have multiple y values recorded; some data points may be duplicated (we are not able to see this on the plot).
- Although a non-constant variance (heteroscedasticity) is not easily spotted on the plot due to a relatively small sample size, we expect it to be present because estimation errors by wildlife service members are likely to increase with the size of a flock.

Example: Inheritance of height

- The sample size is $n = 1375$ (pairs of mothers and daughters).
- The original heights are rounded to the nearest inch.
- For the graph, data were jittered (uniform, $U(-0.5, 0.5)$, random noise added to mothers' and daughters' heights).

Example: Inheritance of height



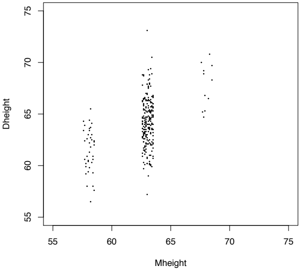
Data: `Heights` from R package `alr4`.

Example: Inheritance of height

Inheritance of height: Observations from the scatterplot.

- Ranges of heights for mothers and daughters appear the same.
- Mothers' and daughters' heights are clearly not independent, although the variability is high compared to the first two examples.
- The scatter appears elliptically shaped (rather typical if (X, Y) is a bivariate normal random vector).
- What about variance in the daughter's height for short, about average, and tall mothers?

Example: Inheritance of height



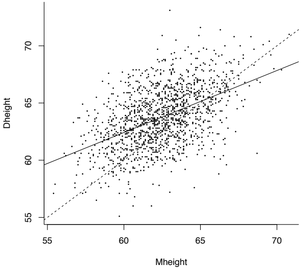
Data: `Heights` from R package `alr4`.

Example: Inheritance of height

Inheritance of height: Examining daughters' heights for mothers who are about 58, 64 and 68 inches tall, we find that the mean is increasing.

- The variance might be about the same (notice many more data points in the middle).
- The next figure illustrates two possible regression lines.

Example: Inheritance of height



Data: `Heights` from R package `alr4`.

Example: Fuel consumption

Goal: Describe how fuel consumption varies in the United States. What is the effect of the state gasoline tax on fuel consumption?

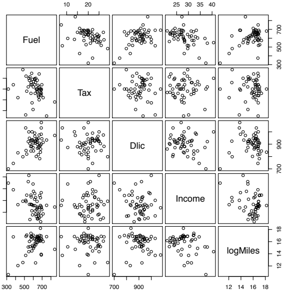
Variables:

- Dlic - 1000×[number of licensed drivers in the state]/[population of the state older than 16 in 2001].
- Income - yearly personal income in the year 2000.
- Fuel - 1000×[gasoline sold in thousands of gallons]/[population of the state older than 16 in 2001].
- logMiles - log(Miles), where Miles denotes the miles of Federal-aid highway in the state.
- Tax - Gasoline state tax rate in cents per gallon.

For multiple regression, scatterplot matrices can be useful.

Data: `fuel2001` from R package `alr4`.

Example: Fuel consumption



Example: Fuel consumption

Scatterplot matrix observations:

- The first row/column shows scatterplots of marginal relationship between fuel consumption and each of the predictors.
- Because marginal relationships among the pairs of the predictors is weak, marginal plots for fuel versus the predictors are informative for the multiple regression problem.

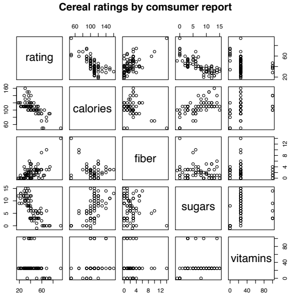
Example: Healthy breakfast data

Goal: describe how Consumer Reports' ratings of breakfast cereal are related to nutritional information.

Variables:

- rating - a rating of the cereals
- calories - calories per serving
- fiber - grams of dietary fiber
- sugars - grams of sugars
- vitamins - vitamins and minerals: $0, 25$, or 100, indicating the typical percentage of FDA recommended daily intake.

Example: Healthy breakfast data



Example: Healthy breakfast data

Observations:

- The rating seems to be related to calories and sugars, however calories and sugar content also seem to be related to each other.
- Note: If three or more predictors were linearly related, such as

$$X_1 + X_2 - X_3 \approx 0,$$

we would not be able to see this sort of relationship on a matrix of scatterplots.