Inference 2: Fitting models to data

Last updated: October 10, 2017

Quantify the relationship between variables

Example: `numchar` vs `linebreak` in emails.
Example: `salary` vs `degree` in teacher.
Example: `survived` vs `age,sex,class` in titanic.

Descriptive: what do I see? (qualitative). Scatterplot, correlation.
Hypothesis testing: is what I see significant? (yes/no). t-test, z-test.
Model: **how much does blah affects blah?** (quantitative)

Quantify the relationship between variables

Example: `numchar` vs `linebreak` in emails.
Example: `salary` vs `degree` in teacher.
Example: `survived` vs `age,sex,class` in titanic.

Descriptive: what do I see? (qualitative). Scatterplot, correlation.
Hypothesis testing: is what I see significant? (yes/no). t-test, z-test.
Model: **how much does blah affects blah?** (quantitative)

eg: If a teacher gets a master degree, how much does his salary go up by?

Quantify the relationship between variables

Example: `numchar` vs `linebreak` in emails.
Example: `salary` vs `degree` in teacher.
Example: `survived` vs `age,sex,class` in titanic.

Descriptive: what do I see? (qualitative). Scatterplot, correlation.
Hypothesis testing: is what I see significant? (yes/no). t-test, z-test.
Model: **how much does blah affects blah?** (quantitative)

eg: If a teacher gets a master degree, how much does his salary go up by?
eg: survival rates for (adult male first class) vs (child female third class)?

Quantify the relationship between variables

Example: `numchar` vs `linebreak` in emails.
Example: `salary` vs `degree` in teacher.
Example: `survived` vs `age,sex,class` in titanic.

Descriptive: what do I see? (qualitative). Scatterplot, correlation.
Hypothesis testing: is what I see significant? (yes/no). t-test, z-test.
Model: **how much does blah affects blah?** (quantitative)

eg: If a teacher gets a master degree, how much does his salary go up by?
eg: survival rates for (adult male first class) vs (child female third class)?

Quantify the relationship between variables

Example: `numchar` vs `linebreak` in emails.
Example: `salary` vs `degree` in teacher.
Example: `survived` vs `age,sex,class` in titanic.

Descriptive: what do I see? (qualitative). Scatterplot, correlation.
Hypothesis testing: is what I see significant? (yes/no). t-test, z-test.
Model: **how much does blah affects blah?** (quantitative)

eg: If a teacher gets a master degree, how much does his salary go up by?
eg: survival rates for (adult male first class) vs (child female third class)?

Models we learn	cat. vs (cat., num)	num. vs (cat., num.)
linear regression		✓
logistic regression	✓	

Simple linear regression

Linear regression= find best line that goes through the scatterplot.

- Numerical variables X, Y.
- Data: pairs $(x_i, y_i), i = 1, \ldots, n$.

Model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

$\epsilon \sim N(0, \sigma^2)$. That is,

- For a fixed $X = x$ value, Y equals $\beta_0 + \beta_1 x$, plus some noise
- The noise has Normal distribution, mean 0, variance σ^2
- β_0, β_1 are unknown constants.
- The noise are independent for different data points.

Key assumptions

- $Y =$ linear in X plus noise.
- Noise have normal distribution, mean zero, constant variance.
- Noise are independent

Linear regression $=$ find β_0, β_1.

Terminologies

- X: regressors, exogenous, explanatory, covariate, **input**, predictor, ...
- Y: regressand, endogenous, response, measured, **output**, criterion, ...
- Estimate for β_0: b_0. Intercept
- Estimate for β_1: b_1. Slope
- Errors: $e_i = y_i - b_0 - b_1 x_i$. Residuals
- Homoscedasticity = noise have variance constant over x-values.
- Heteroscedasticity = noise do not have constant variance
- Simple regression = only one X variable. Multiple regression = more than one X variable.
- Linear regression = least squares (LS), ordinary least squares (OLS), ℓ_2-regression

R Command: `lm`. (= 'linear model').

Parameter estimation

In least squares, the 'line of best fit ' = one where b_0, b_1 minimizes

$$\sum_{i=1}^{n} (y_i - c_0 - c_1 x_i)^2.$$

- Above = total distance of all residuals
- Least squares = minimize squared Euclidean distance of the residuals
- **Why square**? (and not sum of absolute values, say?)

Parameter estimation

In least squares, the 'line of best fit ' = one where b_0, b_1 minimizes

$$\sum_{i=1}^{n}(y_i - c_0 - c_1 x_i)^2.$$

- Above = total distance of all residuals
- Least squares = minimize squared Euclidean distance of the residuals
- **Why square**? (and not sum of absolute values, say?)
  - Least squares = easy to optimize. Exact formula.
  - Minimize the total variance of the error.
  - Penalize large error: double the error $\Rightarrow$ more than double the penalty!

Interpret the R output

Example: `numchar` vs `linebreak` in emails.
Example: `head` vs `total` length in possums.
Example: `math` vs `read` scores in hsb2.

- Estimates for β_0, β_1
- *p*-values: are β_0, β_1 significantly different from zero?
- R^2: $= 1-$ variance in error / variance in Y. Percentage of variance 'explained' by line.