Lecture 13: Regression Problems

Pratheepa Jeganathan

05/01/2019

Recall

- One sample sign test, Wilcoxon signed rank test, large-sample approximation, median, Hodges-Lehman estimator, distribution-free confidence interval.
- Jackknife for bias and standard error of an estimator.
- Bootstrap samples, bootstrap replicates.
- Bootstrap standard error of an estimator.
- Bootstrap percentile confidence interval.
- Hypothesis testing with the bootstrap (one-sample problem.)
- Assessing the error in bootstrap estimates.
- Example: inference on ratio of heart attack rates in the aspirin-intake group to the placebo group.
- The exhaustive bootstrap distribution.

- Discrete data problems (one-sample, two-sample proportion tests, test of homogeneity, test of independence).
- Two-sample problems (location problem - equal variance, unequal variance, exact test or Monte Carlo, large-sample approximation, H-L estimator, dispersion problem, general distribution).
- Permutation tests (permutation test for continuous data, different test statistic, accuracy of permutation tests).
- Permutation tests (discrete data problems, exchangeability).

The independence problem

Introduction

- Correlation: measures the degree of which two variables are related.
- Regression: measures the stochastic relationship between response variable and one or more predictor variables.
  - Regression relationship: simple linear regression, multiple linear regression, nonlinear regression.

Correlation

- Consider random pairs (X, Y). The strength of the relationship or association between X and Y is of our main interest.
- If X and Y are discrete, we can use odds ratio to measure the association and χ^2 goodness-of-fit test for testing the association.
  - If X and Y are independent $P(X = x, Y = y) = P(X = x)\,P(Y = y)$ for all levels of X and Y.
- If X and Y are continuous, from random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ we can use Pearson correlation coefficient or nonparametric Kendall or Spearman statistics to measure the strength of the association.

Pearson's correlation coefficient

- Let X and Y be continuous random variables with mean μ_X, μ_Y and standard deviation σ_X, σ_Y.
- Pearson's correlation coefficient is

$$\rho = \frac{\mathbb{E}\left(X-\mu_X\right)\left(Y-\mu_Y\right)}{\sigma_X\sigma_Y} = \frac{\mathbb{E}\left(XY\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)}{\sigma_X\sigma_Y}.$$

- If X and Y are independent, $\mathbb{E}\left(XY\right)=\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)$. Thus, $\rho = 0$, converse is not true.
  - If X and Y are bivariate normal, converse is also true.
- If X and Y are dependent, $\rho \neq 0$.
- Pearson correlation coefficient measures the linear association between X and Y.

Estimate Pearson's correlation coefficient

- Sample Pearson's correlation coefficient:

$$\hat{\rho} = r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}.$$

- Slope in simple linear regression is related to sample Pearson's correlation coefficient.

$$\hat{\beta} = r\left(\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\right),$$

where $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are sample standard deviations of X and Y, respectively and $\hat{beta}$ is the least squares estimate of slope in a simple regression of Y on X.

Pearson's correlation coefficient

- If X and Y have a bivariate normal distribution, testing Pearson's correlation coefficient using student's t-distribution.
- Testing using permutation method (assume $\left(X_i, Y_{\Pi(i)}\right)$ is exchangeable):
  - Under the null hypothesis of independence, define $\left(X_i, Y_{\Pi(i)}\right)$, where $\Pi(i)$ is any permutation of $\{1, \cdots, n\}$.
- Construct confidence interval using bootstrap method.
  - Use nonparametric bootstrap: sample with replacement (X_i, Y_i).
- Note:
  - If the range of the distribution is bounded, ρ is always defined.
  - ρ is not defined for Cauchy distribution (it has undefined variance).
  - Caution should be given for heavy-tailed distributions.
  - ρ is high-sensitive to outliers and distribution assumption.

The independence problem (tests based on signs - Kendall)

- Let $(X_i, Y_i), i = 1, \cdots, n$ be IID bivariate observations from a joint distribution $F_{X,Y}(x, y)$.
- Testing independence
  - $\mathrm{H}_0 : F_{X,Y}(x, y) = F_X(x) F_Y(x)$ for all pairs (x, y) versus H_A : X and Y are dependent.
- Kendall population correlation coefficient τ

$$\tau = 2P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - 1.$$

- τ measures the monotonicity between X and Y.
- If X and Y are independent, $\tau = 0$, converse is not true.
- If $\tau \neq 0$, X and Y are dependent.

The independence problem (tests based on signs - Kendall)

- $P\{(Y_2 - Y_2)\,(X_2 - X_1 > 0)\} =$
  $P\,(X_2 > X_1, Y_2 > Y_2)\,P\,(X_2 < X_1, Y_2 < Y_2)$.
- Under H_0, $P\,(X_2 > X_1, Y_2 > Y_2) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{4}\right)$.
- Thus, Under H_0, $\tau = 2\left(\left(\frac{1}{4}\right) + \left(\frac{1}{4}\right)\right) - 1 = 0$.

The independence problem (tests based on signs - Kendall)

- $\mathrm{H}_0 : \tau = 0$ versus $\mathrm{H}_0 : \tau \neq 0$ or $\mathrm{H}_0 : \tau > 0$ or $\mathrm{H}_0 : \tau < 0$.
- Significance level α.
- Test statistic: $\bar{K} = K / \left(n \left(n - 1 \right) / 2 \right)$, where

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q\left(\left(X_i, Y_i \right) \left(X_j, Y_j \right) \right),$$

where

$$Q\left(\left(X_i, Y_i \right) \left(X_j, Y_j \right) \right) = \begin{cases} 1 & ; \left(Y_j - Y_i \right) \left(X_j - X_i \right) > 0 \\ -1 & ; \left(Y_j - Y_i \right) \left(X_j - X_i \right) < 0. \end{cases}$$

- $\left(Y_j - Y_i \right) \left(X_j - X_i \right) > 0$ concordant.
- $\left(Y_j - Y_i \right) \left(X_j - X_i \right) < 0$ discordant.

The independence problem (tests based on signs - Kendall)

```
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         exact = NULL, conf.level = 0.95,
  continuity = FALSE, ...)
```

The independence problem (tests based on signs - Kendall)

- The large-sample approximation:
  - $K^* = \dfrac{K}{\{n(n-1)(2n+5)/18\}^{1/2}} \sim \mathrm{N}\,(0,1)\,.$
- Ties: if there are tied X values and or Y values, assign zero to Q.
  - Approximate test.

Example (tests based on signs - Kendall)

- Hunter L measure of lightness X, along with panel scores Y for nine lots of canned tuna $n = 9$.
- It is suspected that the Hunter L value is positively associated with the panel score.

```
Table8.1 = data.frame(x = c(44.4, 45.9, 41.9, 53.3,
  44.7, 44.1, 50.7, 45.2, 60.1),
  y = c( 2.6,  3.1,  2.5,  5.0,  3.6,
    4.0,  5.2,  2.8,  3.8))
```

Example (tests based on signs - Kendall)

```r
cor.test(x = Table8.1$x, y = Table8.1$y,
  method = "kendall", alternative = "greater")
```

```
##
##  Kendall's rank correlation tau
##
## data:  Table8.1$x and Table8.1$y
## T = 26, p-value = 0.05972
## alternative hypothesis: true tau is greater than 0
## sample estimates:
##       tau
## 0.4444444
```

- T is sum of positive Q's.
- $K = 2T - n(n-1)/2$. Now, we can use this for large-sample approximation.

Kendall' s sample rank correlation coefficient

- $\hat{\tau} = \dfrac{2K}{n(n-1)}$.
- For the example

```
T = 26
n = length(Table8.1$x)
K = 2*T-n*(n-1)/2; K
```

```
## [1] 16
```

```
tau.hat = 2*K/(n*(n-1)); tau.hat
```

```
## [1] 0.4444444
```

```
cor(Table8.1$x, Table8.1$y, method = "kendall")
```

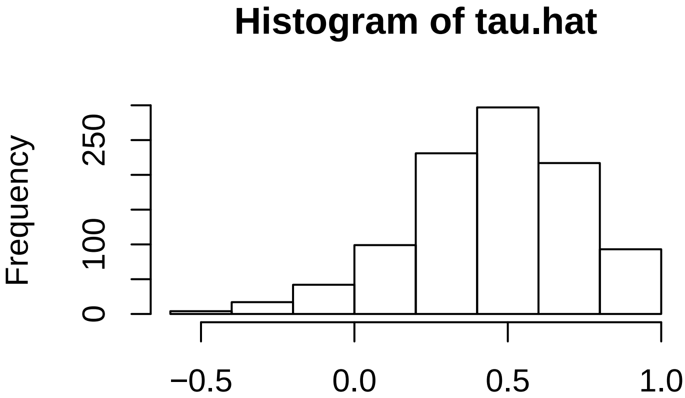
```
## [1] 0.4444444
```

Bootstrap confidence interval (Kendall's correlation coefficient)

- Sample with replacement (X_i, Y_i) to obtain bootstrap sample (X_i^*, Y_i^*).
- Compute bootstrap replicate value of $\hat{\tau}^*$.
  - Necessary to use $Q = 0$ for ties.
- From bootstrap replicates, $\hat{\tau}^{*1}, \hat{\tau}^{*2}, \cdots, \hat{\tau}^{*B}$, construct $(1 - \alpha)\,100\%$ confidence interval for τ.

Bootstrap confidence interval (Kendall's correlation coefficient)

```
library(NSM3)
kendall.ci(Table8.1$x, Table8.1$y, alpha=.05,
  type="t", bootstrap = T, B = 1000)
```



The indepndence problem (tests based on ranks - Spearman)

- Spearman rank correlation coefficient ρ_s.
- ρ_s measures monotonic relationships (whether linear or not).
- Spearman's sample rank correlation coefficient r_s.
- Rank X_i's, denote by R_i's and rank Y_i's, denote by S_i's.
- r_s is Pearson product moment sample correlation of R_i and S_i.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)},$$

$D_i = S_i - R_i, i = 1, \cdots, n.$

Example (tests based on ranks - Spearman)

```r
cor.test(x = Table8.1$x, y = Table8.1$y,
  method = "spearman", alternative = "greater")
```

```
##
##  Spearman's rank correlation rho
##
## data:  Table8.1$x and Table8.1$y
## S = 48, p-value = 0.0484
## alternative hypothesis: true rho is greater than 0
## sample estimates:
## rho
## 0.6
```

Example (tests based on ranks - Spearman)

```r
cor(x = Table8.1$x, y = Table8.1$y,
  method = "spearman")
```

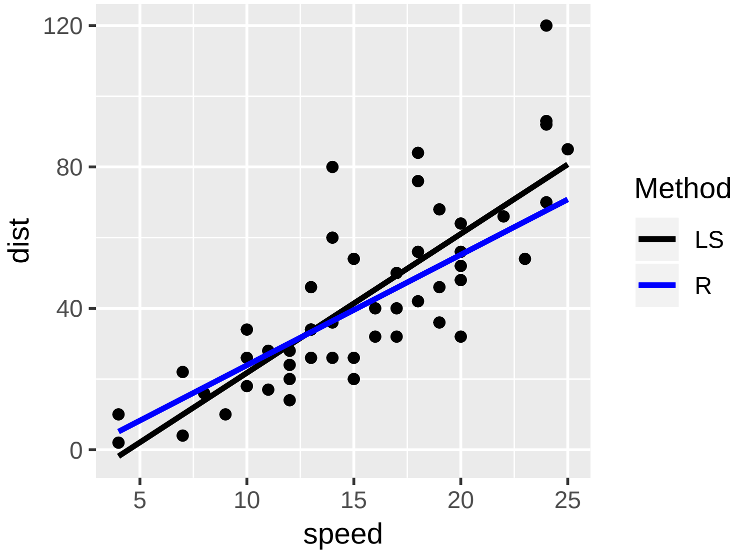
```
## [1] 0.6
```

Rank-based regression analysis

Simple linear regression

- Linear regression in two-sample problem:
  - Combine $X_i; i = 1, \cdots, m$ and $Y_j; j = 1, \cdots, n$.
  - $Z = (X_1, \cdots, X_m, Y_1, \cdots, Y_n)^T$, $N = n + m$.
  - Let $\boldsymbol{g} = (1, \cdots, 1, 0, \cdots, 0)^T$, 1's in first m position and rest is 0's.
  - Two-sample problem as a linear model: $Z_i = \beta_0 + \Delta g_i + \epsilon_i, i = 1, \cdots, N$, where $e_1, \cdots, e_N \sim F(\cdot)$.
  - Estimate Δ and test for Δ.

Rank-based linear regression



Test for slope (based on signs)

- Simple linear model: $Y_i = \alpha + \beta X_i + \epsilon_i$.
  - α - intercept
  - β - slope
  - $\epsilon_1, \cdots, \epsilon_n \sim F(\cdot)$ with median 0.
- β measures every unit increase in the value of the independent (predictor) variable X, expected increase (or decrease, depending on the sign) of the dependent (response) variable Y.

Test for slope (based on signs - Theil (1950))

- $\mathrm{H}_0 : \beta = \beta_0$.
- Test statistic: $C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{Sign}\left(D_j - D_i\right)$, where $D_i = Y_i - \beta_0 x_i$.
- Motivation for the test statistic:
  - $D_j - D_i = Y_j - \beta_0 x_j - Y_i + \beta_0 x_i = Y_j - Y_i + \beta_0\left(x_i - x_j\right)$.
  - Median of $Y_j - Y_i = \beta\left(x_j - x_i\right)$.
  - Thus, under H_0, median of $D_i - D_j = \beta\left(x_j - x_i\right) + \beta_0\left(x_i - x_j\right) = \left(\beta - \beta_0\right)\left(x_j - x_i\right)$.
  - When $\beta > \beta_0$, $D_i - D_j$ is positive and leads to larger C values.
- C is the Kendall's correlation statistics, and can be interpreted as a test for correlation between X and Y.
- Slope estimator associated with Theil statistic $\hat{\beta} = \mathrm{median}\{S_{ij}; 1 \leq i, j \leq n\}$, where $S_{ij} = \dfrac{Y_j - Y_i}{x_j - x_i}; 1 \leq i, j \leq n$.

Rank-based intercept estimator

- Define $A_i = Y_i - \hat{\beta} x_i, i = 1, \cdots, n$.
- A point estimator for α is

$$\hat{\alpha} = \text{median}\{A_1, \cdots, A_n\}.$$

Example (Testing slope)

- Effect of Cloud Seeding on Rainfall.
- Smith (1967) described experiment in Australia on cloud seeding.
- Investigate the effects of a particular method of cloud seeding on the amount of rainfall.
- Data
  - Two area of mountains served as target and control.
  - Effect of seeding was measured by the double ratio: [T/Q (seeded)]/[T/Q (unseeded)].
- The slope parameter β represents the rate of change in Y per unit change in x.
- Test $\mathrm{H}_0 : \beta = 0$ versus $\mathrm{H}_A : \beta < 0$.

Example (Testing slope)

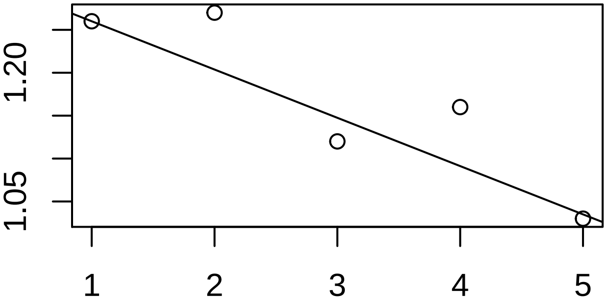
```
Table9.1 = data.frame(x.years.seeded = c(1,2,3,4,5),
  Y.double.ratio = c(1.26,1.27,1.12,1.16,1.03))

theil.fit = theil (Table9.1$x.years.seeded,
  Table9.1$Y.double.ratio,
  beta.0 = 0 ,
  slopes=TRUE,
  type = "l",
  doplot = FALSE)
```

Example (Testing slope)

```
theil.fit
```

```
## Alternative: beta less than 0
## C = -6, C.bar = -0.6, P = 0.117
## beta.hat = -0.056
## alpha.hat = 1.316
##
## All slopes:
##  i j         S.ij
##  1 2  0.01000000
##  1 3 -0.07000000
##  1 4 -0.03333333
##  1 5 -0.05750000
##  2 3 -0.15000000
##  2 4 -0.05500000
##  2 5 -0.08000000
##  3 4  0.04000000
```

Example (Testing slope)



Example (confidence interval for slope)

```r
theil.output = theil(Table9.1$x.years.seeded,
  Table9.1$Y.double.ratio,
  beta.0 = 0 ,
  slopes=TRUE,
  type = "t", doplot = FALSE, alpha = .05)
c(theil.output$L, theil.output$U)
```

```
## [1] -0.15  0.04
```

General multiple linear regression

- Interest in the regression relationship between several (p) independent (predictor) variables and one response variable.
- $Y_i = \zeta + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i, i = 1, \cdots, n$.
- Let $\boldsymbol{\beta}_q = [\beta_1, \cdots, \beta_q]^T$ and $\boldsymbol{\beta}_{p-q} = [\beta_{q+1}, \cdots, \beta_p]^T$.
- $\mathrm{H}_0 : \boldsymbol{\beta}_q = 0$ versus $\mathrm{H}_A : \boldsymbol{\beta}_q \neq 0$.
- Read **HWC** Chapter 9.5
- Use `rfit()` command in R.

The geometry of rank-based linear models

Overview

- Reference (Hettmansperger and McKean 2010, Chapter 3)(hettmansperger2010) and HWC Chapter 9, page 484, comments 24, 25, and 26.
- Analysis (estimation, testing, diagnostic, outlier detection, detection of influential cases) can be based on either signs or ranks.
  - Error distribution could be either asymmetric (use sign) or symmetric (use rank).

The geometry of rank-based linear models

- The model: $Y_i = \alpha + \boldsymbol{x}_i^T \beta + \epsilon_i, i = 1, \cdots, n$.
  - The location parameter of the distribution of ϵ_i is zero.
  - β - $p \times 1$ vector of unknown parameters of interest.
  - α - intercept.
- The model in matrix form: $\boldsymbol{Y} = \mathbf{1}\alpha + \mathbf{X}\beta + \boldsymbol{\epsilon}$.
  - $\mathbf{X}$ has full column rank p.
  - Let Ω_F denote the column space spanned by the columns of $\mathbf{X}$.
  - $\boldsymbol{Y} = \mathbf{1}\alpha + \boldsymbol{\eta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\eta} \in \Omega_F$.
    - Coordinate-free model.
- Estimating $\boldsymbol{\eta}$.
- Testing a general linear hypotheses $\mathrm{H}_0 : \mathbf{M}\beta = 0$ versus $\mathrm{H}_A : \mathbf{M}\beta \neq 0$, where $\mathbf{M}$ is a $q \times p$ matrix of full row rank.

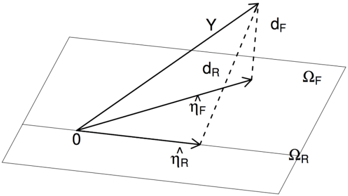
The geometry of rank-based linear models estimation

- Estimate $\boldsymbol{\eta}$ by minimizing the distance between $\boldsymbol{Y}$ and the subspace Ω_F.
  - Define distance in terms of norms or pseudo-norms: $\|\boldsymbol{v}\| = \sum_{i=1}^{n} a\left(R\left(v_i\right)\right) v_i$, $a(1) \leq a(2) \leq \cdots a(n)$ a set of scores generated as $a(i) = \phi\left(\frac{i}{n+1}\right)$ and $\phi\left(u\right) \in (0,1)$.
- Rank estimate of $\boldsymbol{\eta}$ is a vector $\hat{\boldsymbol{Y}}_\phi$ such that

$$D_\phi\left(\boldsymbol{Y}, \Omega_F\right) = \left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\phi\right\|_\phi = \min_{\boldsymbol{\eta} \in \Omega_F} \|\boldsymbol{Y} - \boldsymbol{\eta}\|_\phi .$$

- $\hat{\boldsymbol{\beta}}_\phi = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\boldsymbol{Y}}_\phi$ and $\hat{\alpha} = \text{median}\{Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_\phi\}$.

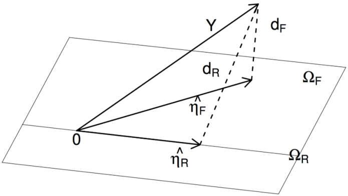
The geometry of rank-based linear models estimation



Source: Hettmansperger & McKean (2011)

Figure 1: Geometry of Estimation

The geometry of rank-based linear models testing



Source: Hettmansperger & McKean (2011)

Figure 2: Geometry of Testing

- Ω_F column space of full model design matrix $\mathbf{X}$.
- Ω_R reduced model subspace $\Omega_R \subset \Omega_F$.
  - $\Omega_R = \{\boldsymbol{\eta} \in \Omega_F : \boldsymbol{\eta} = \mathbf{X}\beta, \text{for some } \beta \text{ such that } \mathbf{M}\beta = 0\}$.
- $\hat{\boldsymbol{Y}}_{\phi,\Omega_R}$ estimate of $\boldsymbol{\eta}$ when the reduced model is fit.
- $D_\phi(\boldsymbol{Y}, \Omega_R) = \left\| \boldsymbol{Y} - \hat{\boldsymbol{Y}}_{\phi,\Omega_R} \right\|_\phi$ denote the distance between $\boldsymbol{Y}$ and Ω_R.

The geometry of rank-based linear models testing

- $RD_\phi = D_\phi(\boldsymbol{Y}, \Omega_R) - D_\phi(\boldsymbol{Y}, \Omega_F)$ reduction in residual dispersion when we pass from reduced model to the full model.
  - Large value of RD_ϕ indicates H_A.

References for this lecture

HWC Chapter 8.

HWC Chapter 9.1-9.4, 9.6.

Hettmansperger, Thomas P, and Joseph W McKean. 2010. *Robust Nonparametric Statistical Methods*. CRC Press.