STAT 420: Methods of Applied Statistics

Midterm I review

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

University of Illinois at Urbana-Champaign
February 28, 2017

Outline

- Midterm 1 is testing your mathematical understanding of the simple and multiple linear regressions, including:
    - Estimating the coefficients (by direct formula or manipulating matrices)
    - Model fitting assessments
    - Distributions of parameter estimates, confidence intervals and hypothesis testing
    - Prediction intervals
    - Testing multiple parameters
    - R implementation
- A calculator (cannot directly fit linear regression), and a one-sided cheat-sheet (A4 size) are allowed.

Parameter estimates

- For simple linear regression, one predictor x with intercept:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r_{xy} \frac{s_y}{s_x}$$

  an alternative formula

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x^2}$$

- For multiple linear regression, any type of design matrix $\mathbf{X}$, as long as its full rank,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

  This is derived from the normal equation.

Matrix operations

- Matrix operations are very important in learning linear regressions, especially for MLR, where no direct formula is given for each individual parameter. They have to be solved jointly using the $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ matrix.
- Another important result is the linear transformation of MVN random variable. If $X \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $Z = \mathbf{A}_{q\times p}X + \mathbf{b}_{q\times 1}$,

$$Z \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}).$$

- This is the foundation for deriving the distribution of $\widehat{\beta}$

Model fitting assessments

- Variance breakdown: $\text{SST} = \text{SSR} + \text{SSE}$.
- The goodness-of-fit statistic: coefficient of determination R^2

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- Estimating the error variance:

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{\text{degrees of freedom}}$$

What is the degrees of freedom (if intercept is included)?

Distributions of parameter estimates

- The key result:

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$$

- So the marginal variances of each $\widehat{\beta}$ are just the diagonal elements of $\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$.
- This can be calculated directly in SLR:

$$\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} & -\frac{\bar{x}^2}{(n-1)s_x^2} \\ -\frac{\bar{x}^2}{(n-1)s_x^2} & \frac{1}{(n-1)s_x^2} \end{pmatrix}$$

CI and Hypothesis testing

- To test a single parameter in SLR, replace σ^2 with $\widehat{\sigma}^2$:

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}} \sim t(n-2)$$

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}/(\sqrt{(n-1)}s_x)} \sim t(n-2)$$

- These results can be used to build CI:

$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2}\, \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\, \frac{\widehat{\sigma}}{\sqrt{(n-1)}s_x}$$

CI and Hypothesis testing

- For testing a single parameter MLR, find the corresponding variance estimate for the parameter in the $\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, and do similar constructions of the t distribution.
- CI's are constructed in a similar fashion.

Prediction intervals

- We will only test prediction intervals of SLR. There are two types of predictions. Predicting the mean response

$$\widehat{\mu}_{\text{new}} \sim \mathcal{N}\left(\beta_0 + \beta_1 x_{\text{new}}, \sigma^2 \left(\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{(n-1)s_x^2}\right)\right)$$

  with CI

$$\widehat{\mu}_{\text{new}} \pm t_{\alpha/2, n-2}\, \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{(n-1)s_x^2}}$$

- CI for a future observed value Y_{new}

$$\widehat{\mu}_{\text{new}} \pm t_{\alpha/2, n-2}\, \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{(n-1)s_x^2}}$$

Testing multiple parameters

- The F test statistic for testing multiple parameters is given by

$$F = \frac{(\mathsf{SSE_R} - \mathsf{SSE_F})/q}{\mathsf{SSE_F}/(n-p-1)}$$

  where q is the number of restrictions in the hypothesis test.
- The R lm() summary output gives the overall F test for all predictors.
- Then this overall F-statistic for testing all predictors is essentially

$$\frac{\mathsf{SSR}/p}{\mathsf{SSE}/(n-p-1)} = \frac{R^2}{1-R^2}\frac{n-p-1}{p}$$

Testing linear constraints

- A linear constraint is specified by

$$\mathsf{H}_0 : \mathbf{A}\boldsymbol{\beta} = c$$

- Apply the linear transformation theorem

$$\mathbf{A}\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\beta}, \sigma^2 \mathbf{A}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{A})$$

- Under the Null, we have $\mathbf{A}\boldsymbol{\beta} = c$, so a t-statistic for testing $\mathbf{A}\boldsymbol{\beta} = c$ is

$$\frac{\mathbf{A}\widehat{\boldsymbol{\beta}} - c}{\sqrt{\widehat{\sigma}^2 \mathbf{A}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{A}}}$$

R output

```
> summary(fit)
Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  7.068221  19.154198    0.369  0.715351
Area        -0.023938   0.022422   -1.068  0.296318
Elevation    0.319465   0.053663    5.953  3.82e-06 ***
Nearest      0.009144   1.054136    0.009  0.993151
Scruz       -0.240524   0.215402   -1.117  0.275208
Adjacent    -0.074805   0.017700   -4.226  0.000297 ***
---
Signif. codes:0  ***  0.001  **  0.01  *  0.05  .  0.1
      1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Good luck!

- Questions?