CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 6
Regression Review, part 2

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

OLS Matrix Formulation

- When we have multivariate data, it is most convenient to formulate the OLS/MLE problem using matrix math
  - i = 1, 2, 3, …, n data point index
  - k = 0, 1, 2, …, p –1 predictor variable index

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$$

$$y_i = i^{th}\ response\ data\ value$$

$$x_{i,k} = i^{th}\ value\ for\ the\ k^{th}\ predictor\ variable$$

OLS Matrix Formulation

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix} = design\ matrix$$

Model in matrix form: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

(each row in $\boldsymbol{X}$ and $\boldsymbol{Y}$ is a "data point")

For a review, see https://onlinecourses.science.psu.edu/stat501/node/382

OLS Matrix Formulation

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Sum of square errors

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X\beta})^T (\boldsymbol{Y} - \boldsymbol{X\beta})$$

Maximum Likelihood estimate (minimum SSE):

$$\frac{\partial \chi^2}{\partial \beta_k} = 0 = \sum_{i=1}^{n} x_{i,k}\varepsilon_i \quad giving \quad \boldsymbol{X^T \varepsilon} = \boldsymbol{0}$$

Using the definition of the residual,

$$\boldsymbol{X^T X \beta} = \boldsymbol{X^T Y}$$

OLS Matrix Formulation

Minimum SSE occurs when the coefficients are estimated as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{HY}$$

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = hat\ matrix$$

Note that $\boldsymbol{H}$ and $\boldsymbol{I} - \boldsymbol{H}$ are symmetric

$\boldsymbol{I}$ = identity matrix

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

(common alternate notation: $b_k = \hat{\beta}_k$)

OLS Matrix Formulation

(for a review of covariance, see lecture 15 at www.lithoguru.com/scientist/statistics/review.html)

We can use our solution to calculate the covariance matrices:

$$cov(\boldsymbol{e}) = s_e^2(\boldsymbol{I} - \boldsymbol{H}), \quad var(e_i) = s_e^2(1 - h_{ii})$$

$$cov(e_i, e_j) = -s_e^2 h_{ij}$$

Elements of the hat matrix

$$cov(\hat{\boldsymbol{Y}}) = s_e^2\boldsymbol{H}, \quad var(\hat{y}_i) = s_e^2 h_{ii}$$

$$cov(\hat{\boldsymbol{\beta}}) = s_e^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Variance of the residuals

MLE Straight-Line Regression

- Our model: $E[Y|X] = \beta_0 + \beta_1 X$

$$\beta_1 = \frac{cov(X,Y)}{var(X)} \qquad \rho = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

$$\beta_0 = E[Y] - \beta_1 E[X]$$

- Ordinary least squares (OLS) estimators:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

Uncertainty of Fit Parameters

- The regression fit is based on a sample of data

Population Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- To create confidence intervals for $b_0, b_1, and\ \hat{y}_i$, we need to know their sampling distributions
  - Given the assumption of $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, the parameter sampling distributions are unbiased and t-distributed (DF = n – 2 = degrees of freedom)

Uncertainty of Fit Parameters

$$var(b_1) = \frac{var(\varepsilon)}{n\ var(X)} \qquad var(b_0) = \frac{var(\varepsilon)}{n}\left(1 + \frac{\bar{x}^2}{var(X)}\right)$$

$$cov(b_0, b_1) = -\bar{x}var(b_1)$$

Estimators:

$$s_e^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$$ (p = number of model parameters)

$$s_{b1}^2 = \frac{s_e^2}{(n-1)s_x^2} \qquad s_{b0}^2 = \frac{s_\varepsilon^2}{n} + s_{b1}^2\bar{x}^2$$

Confidence Intervals

- The sampling distributions are Student's t with DF = n – 2

Slope CI: $b_1 \pm t_{n-2,\alpha} s_{b1}$ $\qquad s_{b1}^2 = \frac{s_\varepsilon^2}{(n-1)s_x^2}$

Intercept CI: $b_0 \pm t_{n-2,\alpha} s_{b0}$ $\qquad s_{b0}^2 = \frac{s_\varepsilon^2}{n} + s_{b1}^2\bar{x}^2$

Critical t-value

α = significance level (e.g., 0.05)

Uncertainty of Predictions

- Uncertainty in $\hat{y}_i$ comes from the spread of the residuals and from uncertainty in the best fit parameters b_1 and b_0

$$var(\hat{y}_i) = \frac{var(\varepsilon)}{n}\left(1 + \frac{(x_i - \bar{x})^2}{var(X)}\right)$$

$$s_{\hat{y}_i}^2 = \frac{s_\varepsilon^2}{n} + s_{b1}^2(x_i - \bar{x})^2$$

- Uncertainty in a predicted new measurement $\hat{y}_{new}$ adds additional uncertainty of a single measurement

$$s_{\hat{y}_{new}}^2 = s_{\hat{y}_i}^2 + s_\varepsilon^2$$

Uncertainty of Correlation Coefficient

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \qquad r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- The sampling distribution for r is about Student's t (DF = n – 2) only for $\rho = 0$. For this case,

$$s_r = \frac{s_\varepsilon}{\sqrt{n-1}s_y}$$

Uncertainty of Correlation Coefficient

- For $\rho \neq 0$, the sampling distribution is complicated
- We'll use the Fisher z-transformation:

$$z = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right)$$

- When n > 25, z is about normal with

$$E[z] = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) \qquad var(z) = \frac{1}{n-3}$$

THE UNIVERSITY OF TEXAS AT AUSTIN
WHAT STARTS HERE CHANGES THE WORLD

Lecture 6: What have we learned?

- Why is the matrix formulation of OLS commonly used?
- What is the hat matrix? Why is it important?
- Be able to use the results of this lecture to calculate standard errors and confidence intervals for model parameters and predicted model values