Confidence Interval and Confidence Region Visualizations for Simple Linear Regression Models

Zhuoyun Wang (zhuoyun2@illinois.edu)

10/13/2019

Abstract

Linear regression is a basic and commonly used type of predictive analysis and the simple linear regression model is first model introduced to students in Statistics and sets the foundation for more complicated models. Confidence intervals (CI) and confidence regions (CR) of the linear regression models are computed from the statistics of the observed data, that might contain the true value of unknown population parameters. Having a clear understanding of them should be a requirement for Statistics students since they have close relationships with other statistics topics, such as significance testing, and are crucial to evaluate the reliability of the estimates.

This project will visualizate confidence bands (surfaces in 3D) for estimated response variable $\hat{y}$ and confidence ellipses (ellipsoids in 3D) for estimated regression coefficients β. The deliverables can hopefully be used as tools to facilitate understanding of CI, CR, and linear regression system as a whole.

Introduction & Backstory

It is well-known that under the assumptions of linearity, multivariate normality, no multicollinearity, and homoscedasticity, the multiple linear regression model has the form $Y = X\beta + \epsilon$, where $Y = (y_1, y_2, \dots y_n)' \in R^n$ is the $n \times 1$ response vector, $X = [1_n, x_1, \dots, x_p] \in R^{n \times (p+1)}$ is the $n \times (p+1)$ design matrix with 1_n be an $n \times 1$ vector of ones and $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})' \in R^n$ is j-th predictor vector $(n \times 1)$, $\beta = (\beta_0, \beta_1, \dots, \beta_p)' \in R^{p+1}$ is $(p+1) \times 1$ vector of coefficients, and $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)' \in R^n$ is the $n \times 1$ error vector. We are often interested in estimating $E[Y|x_*] = x_*^T\beta$ as well as obtaining predictions for future observations at the $((p+1) \times 1$ vector of predictors x_*. The Gauss-Markov theorem tells us that best linear unbiased estimator of $E[Y|x_*]$ is $\hat{y_*} = x_*^T\hat{\beta} = x_*^T(X^TX)^{-1}X^Ty$. The standard error of the fitted value is $sefit(\hat{y_*}) = \hat{\sigma}\sqrt{x_*^T(X^TX)^{-1}x_*}$, where $\hat{\sigma}$ is the estimated value of the standard deviation (σ) of ϵ, and a $100(1-\alpha)\%$ confidence interval for $E[Y|x_*]$ is $(\hat{y_*} - t_{\alpha/2}sefit(\hat{y_*}), \hat{y_*} + t_{\alpha/2}sefit(\hat{y_*}))$. In addition, in MLR we typically want a confidence region, which is similar to a confidence interval but holds for multiple coefficients simultaneously. Given the distribution of $\hat{\beta}$ and some probability theories, we can form a $100(1-\alpha)\%$ confidence region use limits such that $(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \leq (p+1)\hat{\sigma}^2F^{(\alpha)}_{(p+1,n-p-1)}$, where $F^{(\alpha)}_{(p+1,n-p-1)}$ is the critical value for significance level $/alpha$. The inequality defines an ellipsoidal region in the (p+1)-dimensional Cartesian parameter space R^{P+1}. The center of the ellpsoid is at the estimate $\hat{\beta}$.

When p = 1, we only have one intercept and one predictor (X) for the linear regression model, which can be expressed as $Y = \beta_0 + \beta_1X$. We can easily plot the observations, regression line, confidence interval bands for $\hat{y}$, and the confidence region ellipse for $\hat{\beta}$ in 2D settings. However, as p goes larger and more predictors involve in the model, the confidence intervals and confidence regions will become higher-dimensional, so we will need some visulization tools to help present these intervals and regions. This project will primarily

visualize the confidence intervals and confidence regions in 3D settings using the programming language R. There will also be some interactive features added to the 3D graphs so that the users can adjust confidence levels to see how the shapes of confidence intervals and regions change as confidence levels go higher or lower.

Goals

The project deliverables will give students who are newly introduced to linear regression system a better idea about the concept and properties of confidence intervals and regions, and thus set the foundation for further studies in predictive analysis.

Methods

This project will primarily explore and utilize the “rgl” package under the R environment through RStudio. A bunch of functions in this package will be used for achieving the goals: those that can potentially be used for drawing the observations, regression surfaces, confidence surfaces and ellipsoids include rgl.spheres(), rgl.surface(), plot3d(), ellipse3d(), etc, and those for adding user interactive features include togglewidget(), etc. The project will also be open to any other libraries and functions that are useful for doing visualizations. The codes will be written in R Markdown and presented in a .Rmd file, and the interactive 3D WebGL graphic deliverables should be held in a .html file and accessible through any compatible web browser.