PSTAT220A Homework 3

2025-11-16

Problem 1

Consider the data on teen gambling in `teengamb` from the `faraway` package. Fit a simple linear regression model with the expenditure on gambling as the response and income as the covariate:

(a) Check the constant variance assumption for errors. What do you conclude?
(b) Assume that constant variance assumption is violated and that instead the residual variance grows linearly with income. Fit a weighted least squares to account for this fact. Report the coefficient and confidence interval.
(c) Construct a bootstrap confidence interval of the regression coefficients and of the residual variance (assuming an income of 1).
(d) Make a plot showing the prediction intervals (again assuming residual variance grows linearly with income). The plot should have `income` on the x-axis, `gamble` on the y-axis, and prediction bands for each point.
(e) Fit a robust regression model using `rlm`. Compare the coefficients from your weighted least squares fit.
(f) Use `optim` to a fit a t-regression. That is, fit the linear model assuming iid residuals with a t_3 distribution. Construct a bootstrap confidence interval of the regression coefficients.

Problem 2

Standard multiple testing guarantees are given assuming independent tests. In general, test statistics for regression coefficients are correlated.

(a) Assume you have $p = 300$ predictors and $n = 1000$ observations. Let $y = X\beta + \epsilon$ (no intercept), where the first 10 coefficients in β equal 0.2 and the last 90 are 0. Generate $X \sim N(0, 1)$ for each feature. Run a simulation where you generate data from this model many times. Fit the linear model and use both Benjamini-Hochberg and the q-value correction to select discoveries controlling the FDR at 0.05. For each simulated dataset compute 1) the false discovery proportion and 2) the sensitivity (the fraction of the 10 significant predictors that were correctly declared successes).

Plot the distribution FDP and sensitivity for both BH and Q-value approaches using histograms or a beeswarm plot or something similar.

(b) Now consider generating data assuming correlated covariates drawn from a multivariate normal, $X \sim N(\mu, \Sigma)$ with Σ an equicorrelation matrix:

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \tag{1}$$

To investigate how problematic positive equicorrelation is for these procedures, repeat part (a) for each of $\rho \in \{0.25, 0.5, 0.75\}$. Plot the distributions as a function of ρ. What can you conclude about how positive equicorrelation amongst the predictors impacts the false discovery proportions? How does it impact sensitivity?

Problem 3

Run a regression with `prostate` data in `faraway` library with `lpsa` as the response variable and the other variables as covariates.

(a) Select the best model using each of (1) AIC (either forward or backward selection); (2) best subset selection with cross-validation; and (3) LASSO with cross-validation. How much agreement is there between approaches?

(b) Can you improve the model by adding higher order terms? Can you improve the model by adding multiplicative interaction terms? Use any selection approache of choice to settle on a "final" model.