Stat 462/862 Assignment 3

(Due on Nov 14, 2019)

1. Consider the data set *Boston* in the *R* package *MASS*. We treat *dis* as the predictor and *nox* as the response variable.

   (a) Fit a cubic polynomial regression to predict *nox* using *dis*. Report the regression output, and plot the resulting data and polynomial fits.

   (b) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares.

   (c) Perform 5-fold cross-validation to select the optimal degree for the polynomial, and explain your results.

   (d) Fit a regression spline to predict *nox* using *dis*. Report the output for the fit using four degrees of freedom. How do you choose the knots? Plot the resulting fit.

   (e) Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits and report the resulting residual sum of squares. Describe the results you obtain.

   (f) Perform 5-fold cross-validation to select the best degrees of freedom for a regression spline on this data. Describe your results.

2. (For graduate students only) Analyze the "motor cycle data" (use "library(MASS)", then load "data(mcycle)", the data are x=times, y=accel). Use smoothing splines to fit the data. Try different df's in $[5, 20]$. Find the optimal df in

   (a) The observation points and the optimal smoothing spline fit.

   (b) The observation points and the three smoothing splines with df=5, 10, 15 (three different colored curves). Then you should also add a "legend" to denote these lines.

   (c) Plot the cross validation errors against different df's from 5 to 20 (show both points and lines). The step of df's is 0.5. (Hint: from this plot you can find the optimal df.)

3. Consider the *Boston* dataset in the R package $MASS$. Let μ be the population mean of nox and μ_{med} be the population median of nox, where nox represents the nitrous oxide level.

   (a) Provide an estimate of μ.

   (b) Estimate the variance of $\hat{\mu}$ using bootstrap.

   (c) Based on the result in (b), provide a 95% confidence interval for μ.

   (c) Provide an estimate of μ_{med}.

   (d) Estimate the variance of $\hat{\mu}_{med}$ using bootstrap.

   (e) Based on the result in (d), provide a 95% confidence interval for μ_{med}.

4. We will now perform cross-validation on a simulated data set.

   (a) Generate a simulated data set as follows.

   ```
   set.seed(100)
   x=rnorm(100)
   y = x-2*x^2+rnorm(100)
   ```

   In this data set, what is the number of data points n and what is the number of predictors, p? Write out the model used to generate the data in equation form.

   (b) Create a scatter plot of X against Y. Comment on what you find.

   (c) Set a random seed, and compute the leave-one-out cross-validation (LOOCV) errors that result from fitting the following four models using least squares:

      i. $Y = \beta_0 + \beta_1 X + \epsilon$

      ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

      iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

      iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$.

   (d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?

   (e) Which of the models in (c) has the smallest LOOCV error? Is this what you expected? Explain your answer.

(f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

5. Let X be a random variable normally distributed with mean $\mu = 2$ and variance $\sigma^2 = 5$. The following simulation studies the property of certain confidence interval of μ.

   (a) Let n be the sample size and $nsim$ be the number of simulations. Let α be significance level. Write a function in which the inputs are $\mu, \sigma, n, nsim, \alpha$, the output is $(1 - \alpha)100$ percent confidence interval for μ at in each simulation.

   (b) Consider the function in part (a). Let w be the coverage of the $(1-\alpha)100$ percent confidence interval. Define

   $w =$ (the frequency that the confidence interval contains the true mean $\mu)/nsim$.

   Compute the coverage for the following setting (i) $n = 10, nsim = 1000, \alpha = 0.05$; (ii) $n = 10, nsim = 1000, \alpha = 0.025$; (iii) $n = 100, nsim = 1000, \alpha = 0.05$; (iv) $n = 100, nsim = 1000, \alpha = 0.025$. Compare these four coverage.

   (c) (For graduate students 862 only) Suppose that σ^2 is unknown but can be estimated based on the sample. Repeat parts (a) and (b).