STAT 824 hw 05

Influence functions, Edgeworth expansions, bootstrap, residual and wild bootstrap in regression

1. Consider the functional $\sigma^2 = T(F) = \int(x - \int t dF(t))^2 dF(x)$.
   (a) Find the influence function φ_F.
   (b) Identify the remainder R_n in the expansion $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = n^{-1/2}\sum_{i=1}^n \varphi_F(X_i) + \sqrt{n}R(\hat{F}_n - F)$ and argue that it converges in probability to 0 as $n \to \infty$.
   (c) Give the asymptotic behavior of $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2)$ as $n \to \infty$.

2. (a) For a Normal random variable:
      i. Give μ_3, the third central moment.
      ii. Give μ_4/σ^4, where μ_4 is the fourth central moment and σ^2 is the variance.
   (b) Suppose $X_1, \dots, X_n \overset{\text{ind}}{\sim} F$ with $\mathbb{E}X_1 = \mu$ and $\operatorname{Var} X_1 = \sigma^2 < \infty$.
      i. Explain why we expect fast convergence to Normality of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ when F is symmetric.
      ii. Suppose F has the same skewness (μ_3/σ^3) and kurtosis (μ_4/σ^4) as the Normal distribution. Explain why we expect the convergence to Normality of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ to be super fast.

3. Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} F$ with $\operatorname{Var} X_1 < \infty$. Let T be the statistical functional $T(F) = g(\int x dF(x))$, for a differentiable function $g : \mathbb{R} \to \mathbb{R}$ with derivative g' satisfying $|g'(x) - g'(x')| \le L|x - x'|$ for all $x, x' \in \mathbb{R}$. In the von Mises expansion

$$\sqrt{n}(T(\hat{F}_n) - T(F)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_F(X_i) + \sqrt{n}R(\hat{F}_n - F),$$

   give a detailed argument for why $\sqrt{n}R(\hat{F}_n - F)$ becomes negligible.

4. Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} F$ with density f. Let $\xi_{\tau_1}, \xi_{\tau_2}$ be the τ_1, τ_2 quantiles of F, with $0 < \tau_1 < \tau_2 < 1$.
   (a) Use von Mises expansions to find ϑ such that

$$\sqrt{n}[(\hat{\xi}_{\tau_2} - \hat{\xi}_{\tau_1}) - (\xi_{\tau_2} - \xi_{\tau_1})] \to \text{Normal}(0, \vartheta)$$

   in distribution as $n \to \infty$.
   (b) Note that evaluations of the density f appear in the expression for ϑ. Consider replacing f with a kernel density estimate $\hat{f}_n$ (under some choice of bandwidth h) and replacing $F^{-1}(\tau_1)$ and $F^{-1}(\tau_2)$ with $X_{(\lceil \tau_1 n \rceil)}$ and $X_{(\lceil \tau_2 n \rceil)}$, respectively, thereby constructing an estimator $\hat{\vartheta}$ of ϑ. Give the form of an asymptotic $(1 - \alpha)100\%$ CI for $\xi_{\tau_2} - \xi_{\tau_1}$, assuming that your estimator $\hat{\vartheta}$ is a consistent estimator of ϑ.
   (c) Choose a distribution F which is highly non-Normal. Then, for some small sample size n (say between 20 and 50), draw 500 data sets, and with each data set construct the confidence interval you described in part (b) for the IQR (the 0.75 quantile minus the 0.25 quantile). Specifically:
      i. Make a plot showing the true density f corresponding to your distribution F. Indicate the 0.25 and 0.75 quantiles.

ii. Report the proportion of times the confidence interval contained the true IQR.

iii. Include your R code.

Note that you must choose a bandwidth for estimating f. You may simply use the `density()` *function in R to do this if you wish, which has a default way of selecting a bandwidth.*

(d) Now consider using the bootstrap to estimate the sampling distributions of the quantities

$$\sqrt{n}[(\hat{\xi}_{\tau_2} - \hat{\xi}_{\tau_1}) - (\xi_{\tau_2} - \xi_{\tau_1})] \quad \text{and} \quad \sqrt{n}\frac{[(\hat{\xi}_{\tau_2} - \hat{\xi}_{\tau_1}) - (\xi_{\tau_2} - \xi_{\tau_1})]}{\sqrt{\hat{\vartheta}}},$$

that is, of the unstudentized and studentized pivots. Do the following:

i. Describe how you construct the bootstrap version of each pivot quantity.

ii. Generate 100 data sets (under the same settings as before) and compute bootstrap-based CIs for the IQR using based on the unstudentized and studentized pivots. Use 500 Monte Carlo draws to approximate the bootstrap distributions.

iii. Report the coverages of these intervals and compare them to that of the non-bootstrap confidence interval from the previous part. Turn in code.

5. (a) Give a set of residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$, consider generating wild bootstrap residuals $\varepsilon_n^*, \dots, \varepsilon_n^*$ as follows: First generate $U_1^*, \dots, U_n^*$ as independent Beta(1/2,3/2) rvs. Then set $\varepsilon_i^* = \hat{\varepsilon}_i \cdot 4(U_i^* - 1/4)$ for $i = 1, \dots, n$. Show that

$$\begin{aligned} \mathbb{E}_*[\varepsilon_i^*] &= 0 \\ \mathbb{E}_*[(\varepsilon_i^*)^2] &= \hat{\varepsilon}_i^2 \\ \mathbb{E}_*[(\varepsilon_i^*)^3] &= \hat{\varepsilon}_i^3, \end{aligned}$$

which are the moment conditions prescribed for the wild bootstrap in [1].

(b) Consider the linear regression model $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ for $i = 1, \dots, n$, where $\varepsilon_1, \dots, \varepsilon_n$ are independent rvs such that with $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}\varepsilon_i^2 = \sigma_i^2$ for $i = 1, \dots, n$. More precisely, let $(X_{11}, X_{12}, Y_1), \dots, (X_{n1}, X_{n2}, Y_n)$ come from the data generating process in the following code:

```
n <- 30
rho <- .5
X1 <- rnorm(n)
X2 <- rnorm(n,rho*X1,1 - rho^2)
X <- cbind(X1 - mean(X1),X2 - mean(X2))
error <- rgamma(n,.5,scale = 1/(1 + exp(-X[,1]))) - .5 * 1/(1 + exp(-X[,1]))
beta <- c(1,2)
Y <- as.numeric(X %*% beta) + error - mean(error)
```

Letting $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ and $\hat{\boldsymbol{\beta}}_n$ be the least squares estimator of $\boldsymbol{\beta}$, consider, for a vector $\mathbf{c} \in \mathbb{R}^2$ the pivot

$$\sqrt{n}\mathbf{c}^T(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)/\hat{\sigma}_c, \tag{1}$$

where, letting $\mathbf{X}$ be the $n \times 2$ matrix with rows (X_{i1}, X_{i2}), $i = 1, \dots, n$, and $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ be the least-squares residuals,

$$\hat{\sigma}_c^2 = n \cdot \mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2) \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}.$$

i. Give an expression for the variance of $\sqrt{n}\mathbf{c}^T\hat{\boldsymbol{\beta}}_n$.

ii. Give the form of an asymptotic $(1-\alpha)\times 100\%$ CI for $\mathbf{c}^T\boldsymbol{\beta}$, assuming that the pivot in (1) converges in distribution to a standard Normal random variable.

iii. Generate 500 data sets as in the R code above, and for a vector $\mathbf{c}$ of your choosing, build 95% confidence intervals for $\mathbf{c}^T\boldsymbol{\beta}$ with each of the 500 data sets. Record the proportion of times the confidence interval captured the true value of $\mathbf{c}^T\boldsymbol{\beta}$.

iv. Do the same thing, but this time build a confidence interval using the wild bootstrap pivot $\sqrt{n}\mathbf{c}^T(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n)/\hat{\sigma}_c^*$ with the bootstrap residuals considered in the first part; use

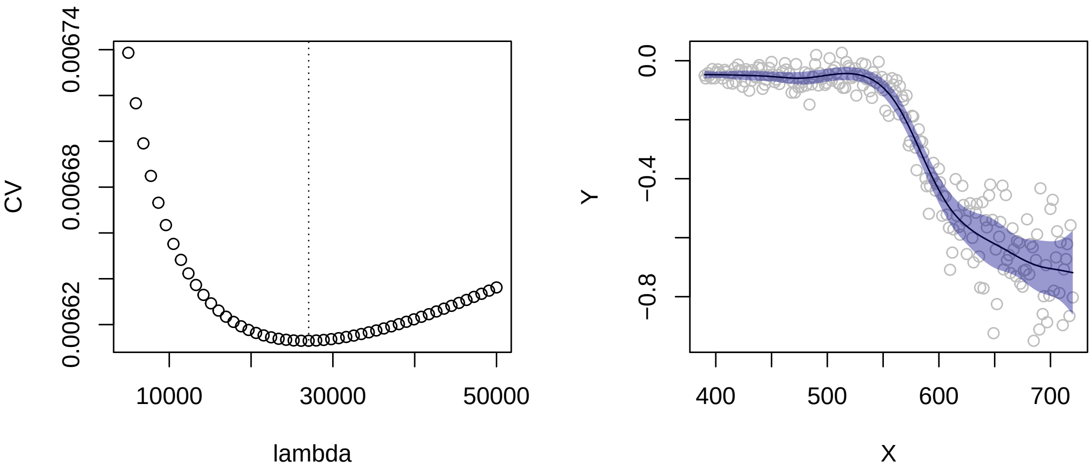
$$(\hat{\sigma}_c^*)^2 = n\cdot\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\cdot\operatorname{diag}((\hat{\varepsilon}_1^*)^2,\ldots,(\hat{\varepsilon}_n^*)^2)\cdot\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}.$$

On each data set, draw 500 bootstrap samples. Record the coverage of the wild bootstrap confidence interval.

6. Import the `lidar` data set from the R package `SemiPar` and consider a nonparametric regression of `logratio` on `range`.

(a) Fit a penalized splines estimator using leave-one-out crossvalidation to choose the level of penalization towards smoothness. Use a large number of spline functions in the basis so that the penalized spline estimator will be approximately the same as the smoothing spline estimator.

(b) Use the tube-formula approach described in Lecture 11 to build a 95% confidence band for the true function, allowing for heterscedastic error term variances. Make plots similar to the ones below, showing the CV output and the fitted function with the confidence band (these are the plots from my analysis; yours should look similar).



References

[1] Enno Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, pages 255–285, 1993.