Goodness of fit tests involving nonparametric function estimation

- Regression. Suppose that we observe (X_i, Y_i): $1 \le i \le n$, where

$$Y_i = f(X_i) + \varepsilon_i, \tag{1}$$

and ε_i's are IID errors with mean zero and variance σ^2. Suppose that The problem of interest is to test whether

$$H_0 : f \in S_0,$$

where S_0 is a known collection of regression functions. For example, S_0 can be the collection of linear functions. A reasonable test statistic for testing H_0 is

$$W = \frac{\sum_{i=1}^n (Y_i - \hat{f}_0(X_i))^2}{\sum_{i=1}^n (Y_i - \hat{f}(X_i))^2},$$

where $\hat{f}_0$ is an estimator of f under H_0 and $\hat{f}$ is an estimator of f. We should reject H_0 when W is large.

- Parametric estimation of f with normal error. Suppose that $f = f_\theta$ for some $\theta \in \Theta$, where $\Theta \subset R^d$ and Θ contains an open set in R^d. Suppose that

$$S_0 = \{f_\theta : \theta \in \Theta_0\},$$

where $\Theta_0 \subset R^{d_0}$ and Θ_0 contains an open set in R^{d_0}. Suppose that $\hat{\theta}_0$ and $\hat{\theta}$ are the least square estimator of θ under the constraints $\theta \in \Theta_0$ and $\theta \in \Theta$ respectively, and $\hat{f}_0 = f_{\hat{\theta}_0}$ and $\hat{f} = f_{\hat{\theta}}$. Suppose that $\varepsilon_i \sim N(0, \sigma^2)$. Then, under H_0,

$$n \log(W) \approx \chi^2(d - d_0) \text{ for large } n,$$

where $\chi^2(d - d_0)$ denotes the chi-square distribution with $d - d_0$ degrees of freedom. Thus we can reject H_0 at level α if $n \log(W) > q_{1-\alpha}$, where $q_{1-\alpha}$ is the $(1 - \alpha)$ quantile for $\chi^2(d - d_0)$. The p-value is the probability that a $\chi^2(d - d_0)$ variable exceeds the observed $n \log(W)$.

- We can also approximate the $(1-\alpha)$ quantile of the distribution of $n \log(W)$ under H_0 using bootstrap data. Let m be the number of bootstrap trials.
  - Compute the residuals $Y_i - \hat{f}(X_i)$: $i = 1, \ldots, n$.
  - For $j = 1, \ldots, m$,
    (i) sample from the residuals and obtain $\hat{\varepsilon}_1^{(j)}, \ldots, \hat{\varepsilon}_n^{(j)}$,
    (ii) compute $Y_i^{(j)} = \hat{f}_0(X_i) + \hat{\varepsilon}_i^{(j)}$ for $i = 1, \ldots, n$,
    (iii) compute $n \log(W)$ based on $Y_i^{(j)}$: $i = 1, \ldots, n$ and X_i: $i = 1, \ldots, n$, and denote the $n \log(W)$ value by $n \log(W^{(j)})$.
  - We can then use the $(1 - \alpha)$ quantile of $2 \log(W^{(j)})$: $j = 1, \ldots, m$ to approximate the $(1-\alpha)$ quantile of the distribution of $n \log(W)$ under H_0. Then the approximate p-value is the proportion of $n \log(W^{(j)})$s that exceed the observed $n \log(W)$.

- Example. Suppose that we have observations (X_i, Y_i): $1 \le i \le n$ generated from (1). Approximate $f(x)$ using

$$a_0 + \sum_{k=1}^{5} (a_k cos(2\pi k x) + b_k \sin(2\pi k x))$$

for $x \in [0, 1]$, where the coefficients a_0, a_1, …, a_5, b_1, …, b_5 are to be estimated using least squared estimation. Consider the testing problem

$$H_0 : \; f \text{ is a constant.} \tag{2}$$

Suppose that our test statistic is $n \log(W)$. Suppose that data are generated as follows.

```
n <- 150
set.seed(1)
x <- runif(n)
y <- sin(2*x) + rnorm(n)
```

Find the p-value of our test for H_0 in (2) using

(a) chi-square approximation of the distribution of $n \log(W)$ under H_0 and

(b) bootstrap.

Sol for (a).

```
w.fun <- function(x,y, resid=F){
 n <- length(y)
 Z <- matrix(0, n, 10)
 for (k in 1:5){
   Z[ ,k] <- cos(2*pi*k*x)
   Z[ ,k+5] <- sin(2*pi*k*x)
 }
 Z <- cbind(rep(1,n), Z)
 mean.y <- mean(y)
 rss0 <- sum((y - mean.y)^2)
 y.lm <- lm(y~Z-1)
 rss <- sum(y.lm$resid^2)
 w <- n*log(rss0/rss)
 if (resid) { ans <- list(w, y.lm$resid); return(ans) } else { return(w) }
}
1-pchisq(w.fun(x,y), 10)  #p-value 0.01920332
```

Sol for (b).

```
pv.fun <- function(x,y,m){
 ans <- w.fun(x,y, resid=T)
 w.obs <- ans[[1]]
 resid <- ans[[2]]
 n <- length(y)
```

```
 w <- rep(0, m)
 mean.y <- mean(y)
 for (j in 1:m){
  e <- sample(resid, n)
  ynew <- mean.y + e
  w[j] <- w.fun(x,ynew)
 }
 return(length(w[w>w.obs])/m)
}
pv.fun(x,y,100) #p-value 0.02
```

Check the distribution of p-value under H_0.

```
ans <- rep(0, 5000)
#ans2 <- ans
n <- 150
for (i in 1:5000){
set.seed(i)
x <- runif(n)
y <- rnorm(n)
ans[i] <- 1-pchisq(w.fun(x,y), 10)
#ans2[i] <- pv.fun(x,y,100)
}
hist(ans)
ks.test(ans, punif)$p.value         #2.976472e-10;
                                    #H0: ans is a random sample from U(0,1)
length(ans[ans<0.05])/length(ans)   #0.0656

#ks.test(ans2, punif)$p.value         #0.3667264
#length(ans2[ans2<0.05])/length(ans2) #0.05
```

- Exercise 1. Suppose that we have observations (X_i, Y_i): $1 \leq i \leq n$ generated from (1). Approximate f using cubic B-spline basis functions on $[0, 1]$ with one knot at 0.5, where the coefficients for B-spline basis functions are to be estimated using least squared estimation. Consider the testing problem

$$H_0: \ f \text{ is a linear function.} \tag{3}$$

Suppose that our test statistic is $n \log(W)$ and data are generated as follows.

```
n <- 150
set.seed(1)
x <- runif(n)
y <- sin(2*x) + runif(n, -0.1, 0.1)*10
```

Find the p-value of our test for H_0 in (3) using

(a) chi-square approximation of the distribution of $n \log(W)$ under H_0 and

(b) bootstrap with 200 bootstrap trials.

- Exercise 2.

  (a) Generate 500 data sets, where the i-th data set is generated as follows.

```
set.seed(i)
n <- 150
x <- runif(n)
y <-  1+x+runif(n, -0.1, 0.1)*5
```

  For each of the 500 generated data sets, perform the test in Exercise 1 for testing (3), where the p-value is to be computed using chi-square approximation of the distribution of $n \log(W)$ under H_0. Use `ks.test` to determine whether there is a strong evidence that the 500 p-vlalues are not from the uniform distribution on $(0, 1)$.

  (b) Do Part (a) again with `n <- 150` replaced by `n <- 5000`.