Jackknife

Max Turgeon

STAT 3150–Statistical Computing

Lecture Objective

- Use jackknife to estimate the bias and standard error of an estimator.

Motivation i

- In the previous module, we saw how we can could perform estimation and hypothesis testing using simulations.
    - **Main idea**: Simulate data from a fixed distribution, compute estimate/test statistic, and repeat the simulation to approximate the sampling distribution.
- This approach can be very powerful when studying the behaviour of estimators, or when comparing multiple testing strategies.
- However, there is a big obstacle in applying these methods for data analysis:
    - *They all assume we know the data generating mechanism.*

Motivation ii

- How can we apply these same principles for data analysis?
  - **Resampling methods**
- We will study resampling methods for the next three modules, and we will see how they can be used for data analysis.

Jackknife i

- The **jackknife** is a method that was first introduced to estimate the *bias* of an estimator.
- We start with a sample $X_1, \ldots, X_n$. From that sample, we compute an estimate $\hat{\theta}$ of a parameter θ.
    - We are interested in estimating $E(\hat{\theta}) - \theta$.
- For each i, we can also create another sample by *omitting* the i-th observation:

$$X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n.$$

- For each of these sample, we can also compute an estimate $\hat{\theta}_{(i)}$ of θ.

Jackknife ii

- E.g compute the sample mean or variance while omitting the i-th observation
- In other words, we now have $n+1$ estimates of θ!
- The jackknife estimate of the bias $E(\hat{\theta})-\theta$ is given by

$$\widehat{\text{bias}}_{jack} = (n-1)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} - \hat{\theta}\right).$$

Example i

- Consider the following two estimate of the variance:

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2, \quad \hat{\sigma}_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

- The only difference is the constant in front of the sum, which implies that $\hat{\sigma}_2^2$ is the unbiased estimate.
- Let's compute the jackknife bias estimate of $\hat{\sigma}_1^2$.

Example ii

```r
# Generate a random sample
n <- 20
xvars <- rgamma(n, shape = 3, rate = 5.5)

# Compute the estimate
theta_hat <- mean((xvars - mean(xvars))^2)
c("estimate" = theta_hat,
  "theoretical" = 3/5.5^2)
```

```
##    estimate theoretical
##  0.06026442  0.09917355
```

Example iii

```r
# Jackknife
theta_i_hat <- numeric(n)

for (i in 1:n) {
  xvars_jack <- xvars[-i]
  mean_i <- mean(xvars_jack)
  theta_i_hat[i] <- mean((xvars_jack - mean_i)^2)
}
```

Example iv

```r
# Estimate of bias
(bias <- (n-1)*(mean(theta_i_hat) - theta_hat))
```

```
## [1] -0.003171811
```

```r
c("De-biased" = theta_hat - bias,
  "Unbiased" = var(xvars))
```

```
##  De-biased   Unbiased
## 0.06343623 0.06343623
```

Example i

- Consider the `patch` dataset in the `bootstrap` package. It contains measurements of a certain hormone on the bloodstream of 8 individuals, after wearing a patch.
- For each individual, we have three measurements: `placebo`, `oldpatch`, and `newpatch`.
- The parameter of interest is a ratio of differences:

$$\theta = \frac{E(\text{newpatch}) - E(\text{oldpatch})}{E(\text{oldpatch}) - E(\text{placebo})}.$$

Example ii

```r
library(bootstrap)
str(patch)
```

```
## 'data.frame':    8 obs. of  6 variables:
##  $ subject : int  1 2 3 4 5 6 7 8
##  $ placebo : num  9243 9671 11792 13357 9055 ...
##  $ oldpatch: num  17649 12013 19979 21816 13850 ...
##  $ newpatch: num  16449 14614 17274 23798 12560 ...
##  $ z       : num  8406 2342 8187 8459 4795 ...
##  $ y       : num  -1200 2601 -2705 1982 -1290 ...
```

Example iii

- `y` is `newpatch - oldpatch`, and `z` is `oldpatch - placebo`.
- Recall that $E(X/Y) \neq E(X)/E(Y)$. So even if we have an unbiased estimate of both the numerator and the denominator of θ, their ratio will generally be **biased**.

```
# Estimate of theta
theta_hat <- mean(patch$y)/mean(patch$z)
```

Example iv

```r
# Jackknife
n <- nrow(patch)
theta_i <- numeric(n)

for (i in 1:n) {
  theta_i[i] <- mean(patch[-i,"y"])/mean(patch[-i,"z"])
}
```

Example v

```r
# Estimate of bias
(bias <- (n-1)*(mean(theta_i) - theta_hat))
```

```
## [1] 0.008002488
```

```r
c("Biased" = theta_hat,
  "De-biased" = theta_hat - bias)
```

```
##      Biased   De-biased
## -0.07130610 -0.07930858
```

Example vi

- The bias is significant: it represents 11% of the estimate.

But be careful:

```
# NOT THE SAME THING
mean(patch$y/patch$z)
```

```
## [1] 0.0379914
```

Estimate of the standard error

- The jackknife can also be used to estimate the standard error of an estimate:

$$\widehat{\text{se}}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}\right)^2}.$$

Example (cont'd)

```r
# Continuing on with the patch dataset
(se <- sqrt((n-1)*mean((theta_i - mean(theta_i))^2)))
```

```
## [1] 0.1055278
```

```r
# 95% CI
c("LB" = theta_hat - bias - 1.96*se,
  "UB" = theta_hat - bias + 1.96*se)
```

```
##         LB         UB
## -0.2861430  0.1275259
```

Example i

- We will consider the `law` dataset in the `bootstrap` package.
- It contains information on average `LSAT` and `GPA` scores for 15 law schools.
- We are interested in the correlation ρ between these two variables

```r
library(bootstrap)
str(law)
```

Example ii

```
## 'data.frame': 15 obs. of 2 variables:
## $ LSAT: num 576 635 558 578 666 580 555 661
651 605 ...
## $ GPA : num 3.39 3.3 2.81 3.03 3.44 3.07 3
3.43 3.36 3.13 ...
```

```r
# Estimate of rho
(rho_hat <- cor(law$LSAT, law$GPA))
```

```
## [1] 0.7763745
```

Example iii

```r
# Jackknife
n <- nrow(law)
rho_i <- numeric(n)

for (i in 1:n) {
  rho_i[i] <- cor(law$LSAT[-i], law$GPA[-i])
}
```

Example iv

```r
# Estimate of bias
(bias <- (n-1)*(mean(rho_i) - rho_hat))
```

```
## [1] -0.006473623
```

```r
c("Biased" = rho_hat,
  "De-biased" = rho_hat - bias)
```

```
##    Biased De-biased
## 0.7763745 0.7828481
```

Example v

```r
(se <- sqrt((n-1)*mean((rho_i - mean(rho_i))^2)))
```

```
## [1] 0.1425186
```

```r
# 95% CI
c("LB" = rho_hat - bias - 1.96*se,
  "UB" = rho_hat - bias + 1.96*se)
```

```
##        LB        UB
## 0.5035116 1.0621846
```

Final remarks i

- The jackknife is a simple resampling technique to estimate bias and standard error.
    - The idea is to remove one observation at a time and recompute the estimate, so that we get a sample from the sampling distribution.
- The theoretical details behind the jackknife are beyond the scope for this course. But two important observations:
    - The “debiased” estimate is generally only asymptotically unbiased. But its bias goes to 0 “more quickly” than the bias of the original estimator.
    - The jackknife only works well for “smooth plug-in estimators”. In particular, the jackknife does **not** work well with the median.

Final remarks ii

- The jackknife was generalized in two important ways:
    - **Bootstrap**: This will be the main topic for next week.
    - **Cross-validation**: This is a method for estimating the prediction error (see STAT 4250).