Atlas-PS 8

David Atlas

10/21/2018

Problem 1

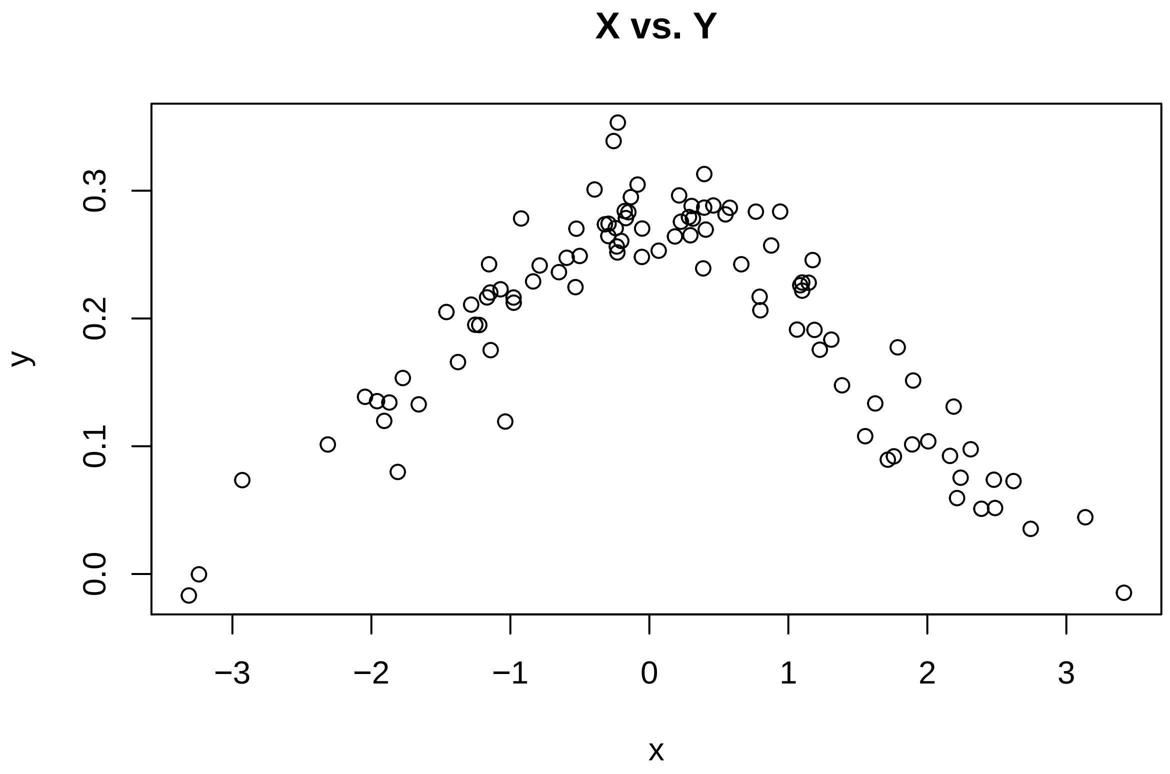
First, we read in the dataset for the problem.

```
x <- scan("xvalues.txt")
y <- scan("yvalues.txt")
```

a)

First, we plot the data.

```
plot(x, y, main="X vs. Y")
```



The MLE of the mean of the distribution is simply the observed mean of X, so we can easily find it.

```
mu <- mean(x)
print(round(mu, 3))
```

```
## [1] 0.173
```

We use the to parameterize g. We can say

$$g(x) = \frac{1}{\sqrt{4\pi}}\mathrm{e}^{-\frac{(\mathrm{x}-.173)^2}{4}},$$

which is the normal density with $\mu = .173$ and $\sigma^2 = 2$.

We can find the apparent error below:

```
mean(abs(y - dnorm(x, mean(x), sqrt(2))))
```

```
## [1] 0.02518254
```

The mean absolute error is .025.

b)

Next, we partition our dataset into 2 halves, and repeat the process.

```
x_1 <- x[1:50]
x_2 <- x[50:100]

y_1 <- y[1:50]
y_2 <- y[50:100]

# Find the predicted values of y
e_1 <- abs(y_1 - dnorm(x_1, mean(x_2), sqrt(2)))
e_2 <- abs(y_2 - dnorm(x_2, mean(x_1), sqrt(2)))

print(round(mean(e_1) + mean(e_2), 3))
```

```
## [1] 0.061
```

The out of sample error is .061

c)

This is what one might expect. The error is lower when calculated on data points that were included in the fit of the function. When we fit on one partition and calculate the error on the other partition, we find that there is a much higher error.

Problem 2

a)

To show the equality in the problem, we establish a few equalities from the text (Gentle).

$$\begin{align} T_j^* &= nT - (n-1)T_{-j} \tag{1} \\ J(T) &= nT - (n-1)\overline{T}_{(\cdot)}. \tag{2} \end{align}$$

Therefore, we can say that

$$\begin{aligned}
\sum_{i=1}^{n}(T_j^* - J(T))^2 &= \sum_{i=1}^{n}(nT - (n-1)T_{-j} - nT + (n-1)\overline{T}_{(\cdot)})^2 \\
&= \sum_{i=1}^{n}(-(n-1)T_{-j} + (n-1)\overline{T}_{(\cdot)})^2 \\
&= \sum_{i=1}^{n}(n-1)^2(-T_{-j} + \overline{T}_{(\cdot)})^2.
\end{aligned}$$

Note that $(a-b)^2 = (b-a)^2$, so we can say that

$$\sum_{i=1}^{n}(n-1)^2(-T_{-j} + \overline{T}_{(\cdot)})^2 = (n-1)^2\sum_{i=1}^{n}(T_{-j} - \overline{T}_{(\cdot)})^2.$$

At this point, we can say that

$$\frac{\sum_{i=1}^{n}(T_j^* - J(T))^2}{n(n-1)} = \frac{n-1}{n}\sum_{i=1}^{n}(T_{-j} - \overline{T}_{(\cdot)})^2.$$

b)

To show that $\widehat{V(T)}_J \le \frac{\sum_{i=1}^{n}(T_j^* - T)^2}{n(n-1)}$, where $\widehat{V(T)}_J = \frac{\sum_{i=1}^{n}(T_j^* - J(T))^2}{n(n-1)}$, we note that the minimization of the function $f(z) = \sum_{i=1}^{n}(x_i - z)^2$, for any set of observations $(x_1, \dots, x_n)$, is equal to $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

This can be shown by taking the first derivative of $f(z)$,

$$f'(z) = -2\sum_{i=1}^{n}(T_j^* - z).$$

Setting it equal to zero and solving,

$$\begin{aligned}
&-2\sum_{i=1}^{n}(T_j^* - z) = 0 \\
\implies &\sum_{i=1}^{n} T_j^* = nz \\
\implies &z = \frac{\sum_{i=1}^{n} T_j^*}{n}.
\end{aligned}$$

Therefore, $\sum_{i=1}^{n}(T_j^* - J(T))^2 \le \sum_{i=1}^{n}(T_j^* - z)^2$, $\forall z \in \mathbb{R}$. By extension, $\frac{\sum_{i=1}^{n}(T_j^* - J(T))^2 \le}{n(n-1)} \le \frac{\sum_{i=1}^{n}(T_j^* - T)^2}{n(n-1)}$.

Problem 3

a)

Let $b_{2,-j}$ be the calculation of b_2 having removed the j^{th} jackknife of the dataset.

The jackknife estimate of the variance (from Gentle) is

$$\widehat{V(b_2)_J} = \frac{\sum_{j=1}^{r}(rb_2 - (r-1)b_{2,j} - \sum_{j=1}^{r}(rb_2 - (r-1)b_{2,j}))^2}{r(r-1)}.$$

The standard deviation is simply the square root of $\widehat{V(b_2)_J}$.

b)

Next, we use the datapoints on Blackboard to find b_2 and the jackknife estimate of the standard deviation for the cases $k = 1$ and $k = 5$.

We read in the data, and define the calculation of b_2.

```
jackknife <- scan("./Module 8 Data Sets/Jackknife.txt")
b2 <- function(y){
  ybar <- mean(y)
  return(sum((y - ybar)^4) / (sum((y - ybar)^2))^2)
}
```

Next, we find the statistic over the whole dataset.

```
round(b2(jackknife), 4)
```

```
## [1] 0.0267
```

The value of b_2 is .0267 over the entire sample.

Next, we find the jackknife estimate for the standard deviation of b_2 for $k = 1$.

```
t_star <- function(data, T, k){
  idx <- split(seq_along(data), ceiling(seq_along(data) /  k))
  r <- length(data) / k
  return(sapply(idx, function(id){
    return(r * T(data) - (r -1) * T(data[-id]))
  }))
}

J <- function(data, T, k){
  mean(t_star(data, T, k))
}

var_t <- function(data, T, k){
  tstar <- t_star(data, T, k)
  j <- J(data, T, k)
  r <- length(data) / k
  return(sum((tstar - j)^2) / (r * (r-1)))
}


sqrt(var_t(jackknife, b2, 1))
```

```
## [1] 0.003714044
```

The jackknife estimate of the standard deviation of b_2 when $k = 1$ is .00371.

Next, we try again with $k = 5$.

```
sqrt(var_t(jackknife, b2, 5))
```

```
## [1] 0.003692711
```

The jackknife estimate of the standard deviation of b_2 when $k = 5$ is .00369.

c)

To get a sense for the actual standard deviation of b_2, we generate several samples from a $N(0, 1)$ distribution.

```
set.seed(73)
sd(sapply(seq(1, 10000), function(i){
  b2_hat <- b2(rnorm(length(jackknife), 0, 1))
}))
```

```
## [1] 0.004568534
```

Using samples of the same size as the dataset that we read in, we see that the standard deviation observed is actually greater than the jackknifed estimate of the standard deviation. The percent error of the estimator relative to that of the sampled data is about 23%.

Therefore, the estimator does not perform well, as it underestimates the variance seen, which can lead to incorrect inference on the data.