Biostatistics 753
Problem Set Two Solutions

March 2, 2010
(Should $X$ be included in a regression model?) Building on the example in the lecture stream, for $i = 1, \ldots, n; j = 1, 2$ let

$$
\begin{align*}
[Y_{ij} \mid \alpha, \beta, \sigma] &= \alpha + \beta X_j + \epsilon_{ij} \\
\epsilon_{ij} &\sim \mathcal{N}(0, 1) \\
X_1 &= -1 \\
X_2 &= +1
\end{align*}
$$

(a) Let $(\hat{\alpha}, \hat{\beta})$ be the least squares estimates. Show that irrespective of whether $X$ is included in the model:

$$
\hat{\alpha} = \bar{Y} = \bar{Y}_- = \frac{Y_{++}}{2n}, \text{ the overall mean}
$$

and if $X$ is included,

$$
\hat{\beta} = (\bar{Y}_2 - \bar{Y}_1)/2 = (Y_2 - Y_1)/2.
$$

Solution:

When $X$ is not in the model, $\epsilon_{ij} = Y_{ij} - \alpha$ and $\hat{\alpha} = \arg \min_\alpha \sum \sum \epsilon_{ij}^2$, so we compute:

$$
\frac{\partial S(\alpha)}{\partial \alpha} = 2 \sum_i \sum_j (Y_{ij} - \alpha) = 0
$$

$$
\Rightarrow \hat{\alpha} = \frac{Y_{++}}{2n} = \bar{Y}
$$
When $X$ is in the model, $\varepsilon_{ij} = Y_{ij} - \alpha - \beta X_j$, and we have

$$\frac{\partial S(\alpha, \beta)}{\partial \alpha} = 2 \sum_i \sum_j (Y_{ij} - \alpha - \beta X_j) = 0$$

$$\implies Y_{++} - 2n\hat{\alpha} - \hat{\beta}n \sum_j X_j = 0$$

$$\implies \hat{\alpha} = \frac{Y_{++}}{2n} = \bar{Y}$$

$$\frac{\partial S(\alpha, \beta)}{\partial \beta} = -2 \sum_i \sum_j (Y_{ij} - \alpha - \beta X_j) X_j = 0$$

$$\implies \sum_i \sum_j Y_{ij} X_j - \hat{\alpha}n \sum_j X_j - \hat{\beta}n \sum_j X_j^2 = 0$$

$$\implies \sum_i (-Y_{i1} + Y_{i2}) - 2n\hat{\beta} = 0$$

$$\implies Y_{+2} - Y_{+1} - 2n\hat{\beta} = 0$$

$$\implies \hat{\beta} = \frac{Y_{+2} - Y_{+1}}{2n} = \frac{\bar{Y}_2 - \bar{Y}_1}{2}$$

(b) Separately, for the cases when $X$ is/ is not included in the model, compute $\bar{Y}_{ij}$.

Solution:

When $X$ is not included, $\bar{Y}_{ij} = \hat{\alpha} = \bar{Y}$.

When $X$ is included:

$$\hat{Y}_{ij} = \hat{\alpha} + \hat{\beta} X_j$$

$$\hat{Y}_{i1} = \hat{\alpha} + \hat{\beta} X_1 = \hat{\alpha} - \hat{\beta} = \bar{Y} - \frac{\bar{Y}_2 - \bar{Y}_1}{2} = \bar{Y}_1$$

$$\hat{Y}_{i2} = \hat{\alpha} + \hat{\beta} X_2 = \hat{\alpha} + \hat{\beta} = \bar{Y} + \frac{\bar{Y}_2 - \bar{Y}_1}{2} = \bar{Y}_2$$

(c) Consider using the prediction model that uses $a\hat{\beta}$ rather than $\hat{\beta}$. That is, use $\hat{Y}_{ij} = \hat{\alpha} + a\hat{\beta} X_j$.

The MSE of $\hat{\alpha} + a\hat{\beta}$ for estimating $(\alpha + \beta)$ (and since $|X| = 1$ for predicting the $Y_{ij}$) is:

$$\text{MSE}(\hat{\alpha} + a\hat{\beta}) = (1 + a^2) \frac{\sigma^2}{2n} + (1 - a)^2 \beta^2$$

If you are choosing between not including $X$ ($a = 0$) or including $X$ ($a = 1$) show that you should use $a = 1$ if $\beta^2 > \frac{\sigma^2}{2n}$. 

11
Solution:

If \( a = 1 \) \( \implies \) \( \text{MSE}(\hat{\alpha} + \hat{\beta}) = \frac{\sigma^2}{n} \). If \( a = 0 \) \( \implies \) \( \text{MSE}(\hat{\alpha}) = \frac{\sigma^2}{2n} + \beta^2 = \frac{1}{2} \text{MSE}(\hat{\alpha} + \hat{\beta}) + \beta^2 \).

We would use \( a = 1 \) instead of \( a = 0 \) \( \iff \) \( \text{MSE}(\hat{\alpha} + \hat{\beta}) < \text{MSE}(\hat{\alpha}) \iff \frac{\sigma^2}{n} < \frac{\sigma^2}{2n} + \beta^2 \iff \beta^2 > \frac{\sigma^2}{2n} \).

(d) More generally, show that the optimal \( a \) is:

\[
\begin{align*}
    a(\sigma^2, \beta, n) &= \frac{\beta^2}{\frac{\sigma^2}{2n} + \beta^2} = \frac{2n \beta^2}{\sigma^2 + 2n \beta^2} = \frac{t^2}{1 + t^2} \\
    t^2 &= \frac{(2n)\beta^2}{\sigma^2}
\end{align*}
\]

Note that \( t^2 \) has the form of the square of the t-statistic testing \( \beta = 0 \).

Solution:

The optimal \( a \) is the one that minimizes:

\[
\text{MSE}(\hat{\alpha} + a\hat{\beta}) = (1 + a^2)\frac{\sigma^2}{2n} + (1 - a)^2 \beta^2,
\]

then we compute

\[
\frac{\partial \text{MSE}(\hat{\alpha} + a\hat{\beta})}{\partial a} = 2a\frac{\sigma^2}{2n} - 2(1 - a)\beta^2
\]

\[
\implies \quad \frac{a_s \sigma^2}{n} - 2\beta^2 + 2a_s \beta^2 = 0
\]

\[
\implies \quad a_s = \frac{\beta^2}{\frac{\sigma^2}{2n} + \beta^2} = \frac{t^2}{t^2 + 1}
\]

So that \( a_s \) is the optimal \( a \).
Part v: Simulation

1 Introduction

The objective of this exercise was to examine the performance of several different decision rules for the inclusion or partial inclusion of a covariate. These criteria are based on several different measures of model performance and some details about their theory can be found in the solutions to the first assignment.

2 Results

The analysis was done in R [1]. Code can be found in the attached appendix. The results of the simulations described in the assignment document were as follows:

Table 1: Results from simulations with 500 samples and $\beta = 0$

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Bias ($\beta = 0$)</th>
<th>Variance ($\beta = 0$)</th>
<th>MSE ($\beta = 0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least-Squares</td>
<td>0.0103</td>
<td>0.0526</td>
<td>0.0526</td>
</tr>
<tr>
<td>P-value Based ($p = 0.00$)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P-value Based ($p = 0.05$)</td>
<td>0.0045</td>
<td>0.0166</td>
<td>0.0166</td>
</tr>
<tr>
<td>P-value Based ($p = 0.10$)</td>
<td>0.0088</td>
<td>0.0247</td>
<td>0.0248</td>
</tr>
<tr>
<td>P-value Based ($p = 0.32$)</td>
<td>0.0134</td>
<td>0.0426</td>
<td>0.0427</td>
</tr>
<tr>
<td>P-value Based ($p = 0.50$)</td>
<td>0.0094</td>
<td>0.0489</td>
<td>0.0488</td>
</tr>
<tr>
<td>P-value Based ($p = 1.00$)</td>
<td>0.0103</td>
<td>0.0526</td>
<td>0.0526</td>
</tr>
<tr>
<td>MSE Based</td>
<td>0.0133</td>
<td>0.0428</td>
<td>0.0429</td>
</tr>
<tr>
<td>PI (semi-full knowledge)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PI (fully empirical)</td>
<td>0.0086</td>
<td>0.0266</td>
<td>0.0266</td>
</tr>
</tbody>
</table>
Table 2: Results from simulations with 500 samples and $\beta = 0.1291$

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Bias ($\beta = 0.1291$)</th>
<th>Variance ($\beta = 0.1291$)</th>
<th>MSE ($\beta = 0.1291$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least-Squares</td>
<td>0.0103</td>
<td>0.0526</td>
<td>0.0526</td>
</tr>
<tr>
<td>P-value Based ($p = 0.00$)</td>
<td>-0.1291</td>
<td>0</td>
<td>0.0167</td>
</tr>
<tr>
<td>P-value Based ($p = 0.05$)</td>
<td>-0.0773</td>
<td>0.0275</td>
<td>0.0334</td>
</tr>
<tr>
<td>P-value Based ($p = 0.10$)</td>
<td>-0.0606</td>
<td>0.0373</td>
<td>0.0409</td>
</tr>
<tr>
<td>P-value Based ($p = 0.32$)</td>
<td>-0.0200</td>
<td>0.0491</td>
<td>0.0494</td>
</tr>
<tr>
<td>P-value Based ($p = 0.50$)</td>
<td>0.0005</td>
<td>0.0523</td>
<td>0.0522*</td>
</tr>
<tr>
<td>P-value Based ($p = 1.00$)</td>
<td>0.0103</td>
<td>0.0526</td>
<td>0.0526</td>
</tr>
<tr>
<td>MSE Based</td>
<td>-0.0179</td>
<td>0.0498</td>
<td>0.0501</td>
</tr>
<tr>
<td>PI (semi-full knowledge)</td>
<td>-0.0943</td>
<td>0.0033</td>
<td>0.0122</td>
</tr>
<tr>
<td>PI (fully empirical)</td>
<td>-0.0329</td>
<td>0.0316</td>
<td>0.0326</td>
</tr>
</tbody>
</table>

Table 3: Results from simulations with 500 samples and $\beta = 0.2236$

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Bias ($\beta = 0.2236$)</th>
<th>Variance ($\beta = 0.2236$)</th>
<th>MSE ($\beta = 0.2236$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least-Squares</td>
<td>0.0103</td>
<td>0.0526</td>
<td>0.0526</td>
</tr>
<tr>
<td>P-value Based ($p = 0.00$)</td>
<td>-0.2236</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>P-value Based ($p = 0.05$)</td>
<td>-0.1298</td>
<td>0.0477</td>
<td>0.0634</td>
</tr>
<tr>
<td>P-value Based ($p = 0.10$)</td>
<td>-0.0906</td>
<td>0.0553</td>
<td>0.0590</td>
</tr>
<tr>
<td>P-value Based ($p = 0.32$)</td>
<td>-0.0214</td>
<td>0.0586</td>
<td>0.0427</td>
</tr>
<tr>
<td>P-value Based ($p = 0.50$)</td>
<td>-0.0003</td>
<td>0.0548</td>
<td>0.0547*</td>
</tr>
<tr>
<td>P-value Based ($p = 1.00$)</td>
<td>0.0103</td>
<td>0.0526</td>
<td>0.0526</td>
</tr>
<tr>
<td>MSE Based</td>
<td>-0.0205</td>
<td>0.0585</td>
<td>0.0588</td>
</tr>
<tr>
<td>PI (semi-full knowledge)</td>
<td>-0.1067</td>
<td>0.0132</td>
<td>0.0245</td>
</tr>
<tr>
<td>PI (fully empirical)</td>
<td>-0.0549</td>
<td>0.0390</td>
<td>0.0419</td>
</tr>
</tbody>
</table>
Table 4: Results from simulations with 500 samples and $\beta = 0.6708$

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Bias ($\beta = 0.6708$)</th>
<th>Variance ($\beta = 0.6708$)</th>
<th>MSE ($\beta = 0.6708$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least-Squares</td>
<td>0.0103</td>
<td>0.0526</td>
<td>0.0526</td>
</tr>
<tr>
<td>P-value Based ($p = 0.00$)</td>
<td>-0.6708</td>
<td>0</td>
<td>0.45</td>
</tr>
<tr>
<td>P-value Based ($p = 0.05$)</td>
<td>-0.0541</td>
<td>0.1102</td>
<td>0.1129</td>
</tr>
<tr>
<td>P-value Based ($p = 0.10$)</td>
<td>-0.0175</td>
<td>0.0807</td>
<td>0.0809</td>
</tr>
<tr>
<td>P-value Based ($p = 0.32$)</td>
<td>0.0066</td>
<td>0.0570</td>
<td>0.0569*</td>
</tr>
<tr>
<td>P-value Based ($p = 0.50$)</td>
<td>0.0090</td>
<td>0.0542</td>
<td>0.0542</td>
</tr>
<tr>
<td>P-value Based ($p = 1.00$)</td>
<td>0.0103</td>
<td>0.0526</td>
<td>0.0526</td>
</tr>
<tr>
<td>MSE Based</td>
<td>0.0066</td>
<td>0.0570</td>
<td>0.0569*</td>
</tr>
<tr>
<td>PI (semi-full knowledge)</td>
<td>-0.0578</td>
<td>0.0426</td>
<td>0.0459</td>
</tr>
<tr>
<td>PI (fully empirical)</td>
<td>-0.0615</td>
<td>0.0610</td>
<td>0.0647</td>
</tr>
</tbody>
</table>

Note that that *“* indicate that $MSE > Var$ as a result of numerical error.

3 Discussion

As was expected, the ordinary least squares estimate was approximately normally distributed (see the top plot of figure (1)). The p-value based estimate consisted of a point mass at zero and a skewed distribution closer to the mean. The empirical partial inclusion method was also skewed, as can be seen in the bottom plot of figure (1).

![Histogram of OLS Estimate of Beta](image1)

![Histogram of P-value Based Estimate of Beta (p = 0.05)](image2)

![Histogram of Partial Inclusion (fully empirical) Estimate of Beta](image3)

Figure 1: Simulated distribution of the estimates
As can be seen in the above results (tables (1)-(4)), the p-value based methods with $p = 1.00$ yielded exactly the least squares estimate as $p = 1.00$ enforced the ‘always add X’ rule. The ‘$p=0.00$’ case yielded variances of zero for the estimated $\beta$ since the estimates were uniformly zero.

A careful comparison of the p-value based estimates was more revealing. It was found that the bias increased with the p-value cutoff value as can be seen in the left plot of figure (2), eventually converging to zero. Furthermore, for small p-value cutoff values, as beta increases the bias increases (away from zero) whereas for p-value cutoff values larger than approximately 0.05 this pattern is not seen. In fact, the curve associated with $\beta = 0.2236$ maximized the bias. For the smaller two $\beta$, as the p-value cutoff increases, the variances increase and level out as can be seen in the right plot of figure (2). The larger two $\beta$-valued curves are quite different; they increase until the p-value cutoff is approximately 0.05 and then decrease.

![Figure 2](image)

Figure 2: Investigating p-value based methods for different beta

It is interesting to note that in terms of bias, as the cutoff increases the estimator seems to do better. That is, adding X under less strict conditions tends to decrease bias. From a variance-minimizing perspective, the picture is not quite as simple. For smaller values of $\beta$, the smaller cutoff values are desirable, but this is not the case for the larger $\beta$s. In fact, for large values of $\beta$ and small p-value cutoffs, the variance is very high. It is thus advisable in general to use higher p-value cutoffs in terms of bias and variance unless there is prior knowledge about $\beta$ being small.

The other estimators are compared in figure (3). The eternal tradeoff be-
tween bias and variance can be seen easily as the p-value based method is uniformly of low bias but high variance, et cetera. The most surprising result is that although the semi-full knowledge partial inclusion method has lowest variance, it also has a significant bias. One might have suggested that on the contrary that since it uses more information than is available to the other methods it might be superior in both criteria, but the above mentioned tradeoff takes precedence here.

Figure 3: Investigating methods for different beta

4 Recommendation

Based on the above simulations, it would be advisable to use the p-value based methods with cutoff $p < 0.50$ or $p = 1$ in cases where bias is to be avoided, but the partial inclusion methods do have some very desirable low variance properties. A broader recommendation might be the partial inclusion empirically-based method as it appeases the requirements of both lower bias and lower variance.

5 Appendix: R Code

```r
# Set up
alpha<-0
sig<-1
n<-10
nreps<-500
```
a=c(0,0.25,0.50,0.80)
beta<-sqrt(a*sig^2/(2*n*(1-a)))
Y<-rep(0,2*n)
a.fun<-function(sig,bet,m) {
bet^2/(sig^2/(2*m)+bet^2)
}
bias<-function(x.hat,x) {
mean(x.hat-x)
}
mse<-function(x.hat,x) {
mean((x.hat-x)^2)
}

b.hat<-rep(0,nreps)
sig.sq.hat<-rep(0,nreps)
b.hat.full<-rep(beta,nreps)
b.hat.p<-matrix(0,6,nreps)
b.hat.mse<-rep(0,nreps)
b.hat.semi<-rep(0,nreps)
b.hat.emp<-rep(0,nreps)
eps<-matrix(rnorm(2*n*nreps,0,1),nrow=nreps)

#Simulate

for (b in beta) {
  for (i in 1:nreps) {
    X<-c(rep(-1,n),rep(1,n))
    Y<-alpha+X*b+eps[i,]
    model<-lm(Y~X)
    a.hat<-model$coef[1]
b.hat[i]<-model$coef[2]
sig.sq.hat[i]<-sum((model$resid)^2)/(2*n-2)

    #P-value based
    t<-sqrt(2*pi*(b.hat[i])^2/sig.sq.hat[i])
pval<-2*(1-pt(t,2*n-1))
    inclx<-(pval<=c(0,.05,.10,.25,.5,.1))
b.hat.p[i]<-inclx*b.hat[i]

    #MSE based
    small.model<-lm(Y~1)
sig.sq.small.hat<sum((small.model$resid)^2)/(2*n-1)
b.hat.mse[i]<-(sig.sq.hat[i]<sig.sq.small.hat)*b.hat[i]

    #Partial inclusion (semi-full knowledge)
}
b.hat.semi[i]<-a.fun(1,b,n)*b.hat[i]

#Partial inclusion (fully empirical)
b.hat.emp[i]<-a.fun(sig.sq.hat[i],b.hat[i],n)*b.hat[i]

}

#Assess the estimates
print(paste('For the value beta=',b))

print(paste('Least-Square Estimates including X bias:',bias(b.hat,b)))
print(paste('Least-Square Estimates including X var:',var(b.hat)))
print(paste('Least-Square Estimates including X mse:',mse(b.hat,b)))

print(paste('P-value based with p=0 (never add X) bias:',bias(b.hat.p[1,],b)))
print(paste('P-value based with p=0 (never add X) var:',var(b.hat.p[1,])))
print(paste('P-value based with p=0 (never add X) mse:',mse(b.hat.p[1,],b)))
print(paste('P-value based with p=0.05 bias:',bias(b.hat.p[2,],b)))
print(paste('P-value based with p=0.05 var:',var(b.hat.p[2,])))
print(paste('P-value based with p=0.05 mse:',mse(b.hat.p[2,],b)))

print(paste('P-value based with p=0.10 bias:',bias(b.hat.p[3,],b)))
print(paste('P-value based with p=0.10 var:',var(b.hat.p[3,])))
print(paste('P-value based with p=0.10 mse:',mse(b.hat.p[3,],b)))

print(paste('P-value based with p=0.32 bias:',bias(b.hat.p[4,],b)))
print(paste('P-value based with p=0.32 var:',var(b.hat.p[4,])))
print(paste('P-value based with p=0.32 mse:',mse(b.hat.p[4,],b)))

print(paste('P-value based with p=0.50 bias:',bias(b.hat.p[5,],b)))
print(paste('P-value based with p=0.50 var:',var(b.hat.p[5,])))
print(paste('P-value based with p=0.50 mse:',mse(b.hat.p[5,],b)))

print(paste('P-value based with p=1 (always add X) bias:',bias(b.hat.p[6,],b)))
print(paste('P-value based with p=1 (always add X) var:',var(b.hat.p[6,])))
print(paste('P-value based with p=1 (always add X) mse:',mse(b.hat.p[6,],b)))

print(paste('MSE based bias:',bias(b.hat.mse,b)))
print(paste('MSE based var:',var(b.hat.mse)))
print(paste('MSE based mse:',mse(b.hat.mse,b)))

print(paste('Partial inclusion (semi-full knowledge) bias:',bias(b.hat.semi,b)))
print(paste('Partial inclusion (semi-full knowledge) var:',var(b.hat.semi)))
print(paste('Partial inclusion (semi-full knowledge) mse:',mse(b.hat.semi,b)))
print(paste('Partial inclusion (fully empirical) bias:', bias(b.hat.emp,b)))
print(paste('Partial inclusion (fully empirical) var:', var(b.hat.emp)))
print(paste('Partial inclusion (fully empirical) mse:', mse(b.hat.emp,b)))

par(mfrow=c(3,1))
hist(b.hat,xlim=c(0,1.5),main='Histogram of OLS Estimate of Beta',xlab='beta')
hist(b.hat.p[2,],xlim=c(0,1.5),main='Histogram of P-value Based Estimate of Beta (p=0.05)',xlab='beta')
hist(b.hat.emp,xlim=c(0,1.5),main='Histogram of Partial Inclusion (fully empirical) Estimate of Beta',xlab='beta')

References