36-401 Class Notes and Exercise for 12/7/00
Computer Intensive Methods II (RwG 303-326)
Howard Seltman

The Bootstrap

Main Idea Re-sample the data and re-analyze to get a sampling distribution.

1. Bradley Efron (1982) showed that we can "pull ourselves up by our bootstraps" and get information about the sampling distribution of a statistic by re-calculating the statistic many times using pseudo-datasets generated by resampling the data *with* replacement. A typical pseudo-sample has missing and duplicate observations.

2. General calculations: Let b be an estimator (statistic) based on the original, observed sample. Repeatedly create a new sample from the observed sample, calculate the same statistic and call in b_1^* through b_B^* where B is e.g. 1000 or more. Then

   - the distribution of the b^*'s approximates the true sampling distribution of b,
   - the "bootstrap bias", $\bar{b}^* - b$, estimates the true bias, $E(b) - \beta$,
   - and the "bootstrap standard deviation", s_{b^*}, estimates the true standard error of b, σ_b

   if the appropriate bootstrap assumptions are met,

3. Regression residual resampling: Re-sample the n residuals n times; calculate new Y_i's as $\hat{Y}_i + e_i^*$; analyze the new sample. Repeat B times. Assumptions are that X's are fixed and errors are i.i.d. (but not-necessarily normal).

4. Regression data resampling: Re-sample n entire cases; analyze the new sample; repeat B times. Assumes original sample is "representative". Makes more sense than residual resampling for cases where X is random.

5. Example

```
x_leads$air
rslt_lm(blood~air, leads)
leadres_rslt$res
boot_matrix(NA, 1000, 2)
for (i in 1:1000) {
  y_rslt$fit+sample(leadres, length(leadres), rep=T)
  boot[i,]_lm(y~x)$coef
}

summary(rslt)$coef
#                Value Std. Error   t value     Pr(>|t|)
# (Intercept) 4.963372  0.3353571 14.800256 4.529883e-09
#         air 8.975998  2.6637468  3.369689 5.573415e-03
cat("Boot b0 mean=", round(mean(boot[,1]),3), "  sd=",round(sd(boot[,1]),3),"\n")
# Boot b0 mean= 4.984   sd= 0.309
cat("Boot b1 mean=", round(mean(boot[,2]),3), "  sd=",round(sd(boot[,2]),3),"\n")
# Boot b1 mean= 8.884   sd= 2.466
```

6. Confidence intervals are more complicated

Breakout Session for 12/7/00

```
x_leads$air
rslt_lm(blood~air, leads)
leadres_rslt$res
boot_matrix(NA, 1000, 2)
n_nrow(leads);  Seq_1:n
for (i in 1:1000)
  boot[i,]_lm(blood~air, leads, subset=sample(Seq, n, rep=T))$coef

summary(rslt)$coef
#                Value Std. Error   t value     Pr(>|t|)
# (Intercept) 4.963372  0.3353571 14.800256 4.529883e-09
#         air 8.975998  2.6637468  3.369689 5.573415e-03
cat("Boot b0 mean=", round(mean(boot[,1]),3), "  sd=",round(sd(boot[,1]),3),"\n")
# Boot b0 mean= 4.996   sd= 0.338
cat("Boot b1 mean=", round(mean(boot[,2]),3), "  sd=",round(sd(boot[,2]),3),"\n")
# Boot b1 mean= 8.732   sd= 2.739
```

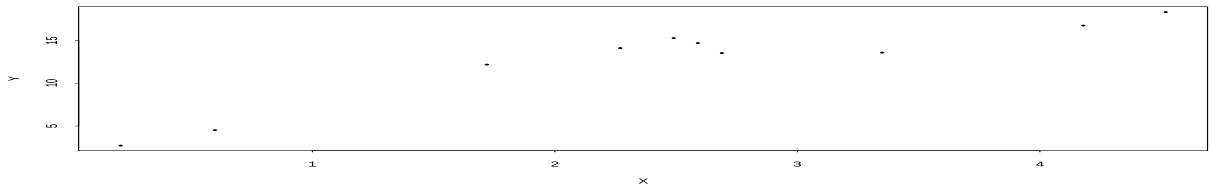
1) What is the nature of the bootstrap performed here? What are the bootstrap bias and standard error? What are the conclusions? How do they compare to the bootstrap above? Why?

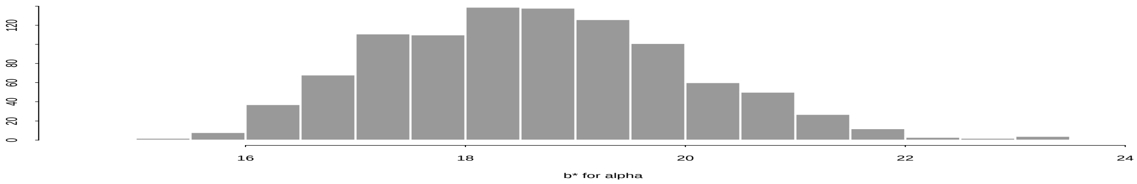
```
plot(dtf$X,dtf$Y, xlab="X", ylab="Y")
parameters(dtf)_list(alpha=20,beta=0.4)
rslt_nls(Y~alpha*(1-exp(-beta*X)), dtf)
summary(rslt)$param
#            Value Std. Error   t value
# alpha 18.4831470  1.5138283 12.209540
#  beta  0.5783129  0.1192827  4.848254

tval_qt(.975, 8);  parm_summary(rslt)$param
cat("[", round(parm[1,1]-tval*parm[1,2],2), ", ", round(parm[1,1]+tval*parm[1,2],2), "]\n", sep="")
#   [14.99, 21.97]

boot _ matrix(NA, 1000, 2); b_rslt$param[1]; myres_rslt$res
for (i in 1:1000) {
  this_data.frame(X=X,Y=rslt$fit+sample(myres, n, rep=T))
  parameters(this)_list(alpha=20,beta=0.4)
  parm_summary(nls(Y~alpha*(1-exp(-beta*X)), this))$param
  bstar_parm[1,1]                 # b* for alpha
  tstar_(parm[1,1]-b)/parm[1,2]   # t* for alpha
  boot[i,]_c(bstar,tstar)
}

hist(boot[,1], xlab="b* for alpha", 25)
qnt_quantile(boot[,2], c(0.05,0.95));   sdBootB_sqrt(var(boot[,1]))   # 1.376369
CIL_b-qnt[2]*sdBootB;   CIH_b-qnt[1]*sdBootB
cat("[", round(CIL,2), ", ", round(CIH,2), "]\n", sep="")
# [16.51, 21.64]
```





2) What types of bootstrap is performed here? What are the conclusions? How do the CI's compare?