Practice problems–Resampling Applications

Problem 1

Consider the `nuclear dataset` from the `boot` package:

```
The nuclear data frame has 32 rows and 11 columns.

The data relate to the construction of 32 light water reactor (LWR) plants constructed
in the U.S.A in the late 1960's and early 1970's. The data was collected with the aim
of predicting the cost of construction of further LWR plants. 6 of the power plants had
partial turnkey guarantees and it is possible that, for these plants, some manufacturers'
subsidies may be hidden in the quoted capital costs.
```

For more details about each variable, have a look at `?boot::nuclear`.

We will look at the following model for the log-cost of a nuclear station:

```
library(boot)
nuclear_fit <- lm(log(cost) ~ date + log(t1) + log(t2) + log(cap) + pr +
                      ne + ct + bw + log(cum.n) + pt,
                  data = nuclear)

summary(nuclear_fit)
```

```
##
## Call:
## lm(formula = log(cost) ~ date + log(t1) + log(t2) + log(cap) +
##     pr + ne + ct + bw + log(cum.n) + pt, data = nuclear)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28574 -0.10408  0.02784  0.09512  0.25031
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.24198    4.22880  -3.368  0.00291 **
## date          0.20922    0.06526   3.206  0.00425 **
## log(t1)       0.09187    0.24396   0.377  0.71025
## log(t2)       0.28553    0.27289   1.046  0.30731
## log(cap)      0.69373    0.13605   5.099 4.75e-05 ***
## pr           -0.09237    0.07730  -1.195  0.24542
## ne            0.25807    0.07693   3.355  0.00300 **
## ct            0.12040    0.06632   1.815  0.08376 .
## bw            0.03303    0.10112   0.327  0.74715
## log(cum.n)   -0.08020    0.04596  -1.745  0.09562 .
## pt           -0.22429    0.12246  -1.832  0.08125 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1645 on 21 degrees of freedom
## Multiple R-squared:  0.8717, Adjusted R-squared:  0.8106
## F-statistic: 14.27 on 10 and 21 DF,  p-value: 3.081e-07
```

a. Perform a residual analysis for this linear regression model fit.
b. Use bootstrap to construct a 95% confidence interval for the regression coefficient corresponding to date. Be explicit about the type of resampling you chose, the type of confidence interval you chose, and the reason why you made those choices.
c. Use bootstrap to build a 95% confidence interval for the expected cost (i.e. on the original scale) for the nuclear plant corresponding to the 32nd observation (i.e. row 32 in boot::nuclear.)
d. Repeat the above, but for a new observation that has the same covariate values as the 32nd observation, except for date, which should be equal to 73.

Problem 2

Consider the `salinity` dataset in the `boot` package:

```
The salinity data frame has 28 rows and 4 columns.

Biweekly averages of the water salinity and river discharge in Pamlico Sound, North Carolina
were recorded between the years 1972 and 1977. The data in this set consists only of those
measurements in March, April and May.
```

For more details about each variable, have a look at `?boot::salinity`.

We will look at the following model for the salinity level:

```
library(boot)
sal_fit <- lm(sal ~ lag + trend + dis,
              data = salinity)

summary(sal_fit)
```

```
##
## Call:
## lm(formula = sal ~ lag + trend + dis, data = salinity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6613 -0.8242  0.2222  0.6459  2.7537
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.69427    3.13026   3.097  0.00492 **
## lag          0.77692    0.08597   9.038 3.41e-09 ***
## trend       -0.02835    0.16087  -0.176  0.86160
## dis         -0.29903    0.10712  -2.792  0.01013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.327 on 24 degrees of freedom
## Multiple R-squared:  0.8273, Adjusted R-squared:  0.8057
## F-statistic: 38.32 on 3 and 24 DF,  p-value: 2.605e-09
```

a. Perform a residual analysis for this linear regression model fit.
b. Use bootstrap to construct a 95% confidence interval for the regression coefficient corresponding to dis. Be explicit about the type of resampling you chose, the type of confidence interval you chose, and the reason why you made those choices.

Problem 3

For this problem, use the `catsM` data in the `boot` package. This dataset contains weight data for 97 adult domestic cats.

a. Display a fitted line plot for the simple linear regression model where body weight (`Bwt`) is the outcome variable, and heart weight (`Hwt`) is the covariate.
b. Display a plot of residuals vs. fitted values.
c. Comment on the fit of this model. Are there any outliers? If so, identify these points by observation number.
d. Based on your analysis above, to analyze the fit using bootstrap, choose a resampling method (i.e. resampling cases or errors) and explain your reasoning.
e. Bootstrap the slopes of this model and obtain a bootstrap estimate the standard error of $\hat{\beta}_1$.
f. Compute a 95% bootstrap confidence interval using the method of your choice.