Influential points

Lab 5 Handout Solutions

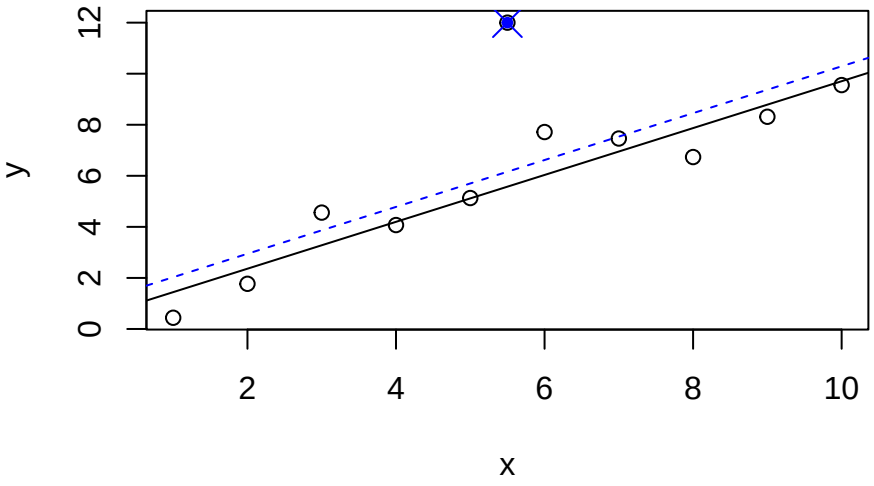
Statistics 139

Important Types of Observations Definitions of types of observations:

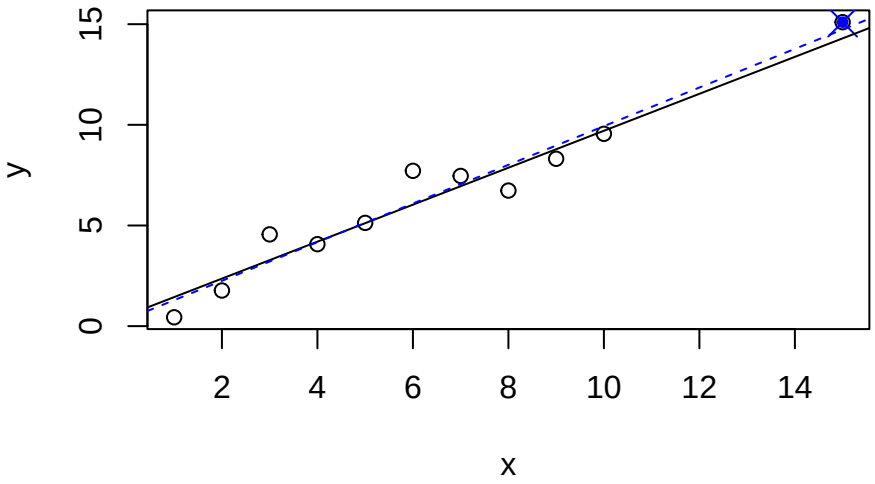
- *Outlier*: An *outlier* is a point that doesn't fit the model well. An outlier may or may not affect the model fit substantially.
- *Influential observation*: An *influential observation* is one whose removal from the dataset would cause a large change in the model fit.
- *Leverage point*: A *leverage point* is extreme in the predictor space. It has the potential to influence the fit, but doest not necessarily do so.

It is important (and easy) to identify these points, but deciding what to do with them can be difficult.

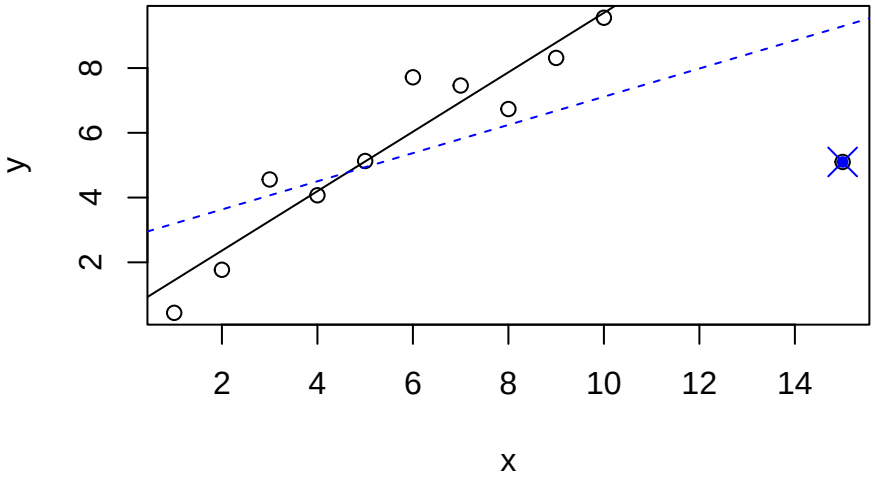
Outlier Here is an outlier that is neither influential nor high leverage.



Leverage point Here is a point of high leverage that is neither influential nor an outlier.



Influential point This is an influential point that is also an outlier.



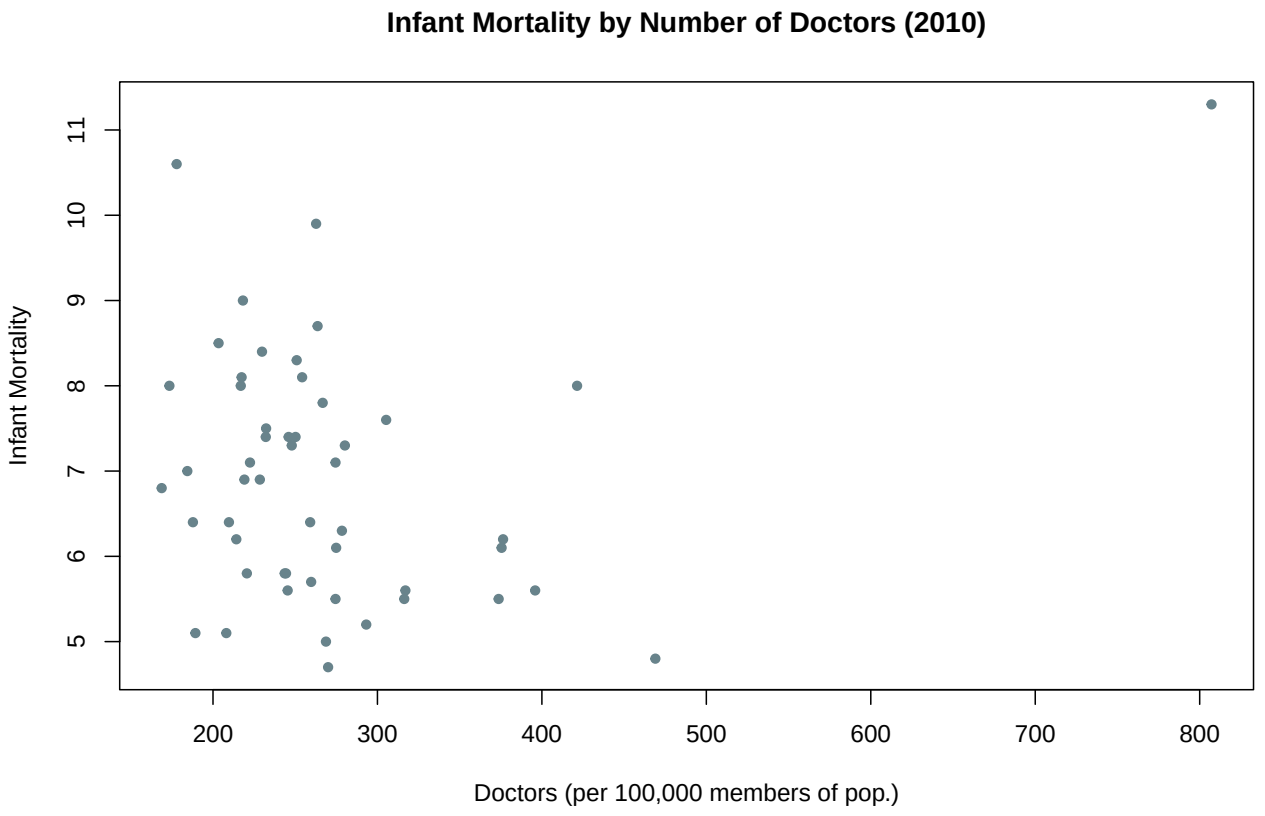
Question 1: Influential points The `census_2010.csv` dataset has data on infant mortality and number of doctors for each of the 50 states including Washington, D.C.

- Infant mortality (`inf.mort`) is measured as number of infant deaths in the first year of life per 1,000 births.
- Number of doctors (`doctors`) is recorded as number of doctors per 100,000 members of the population.

Suppose we are interested in modeling infant mortality rate from number of doctors.

a) Plot the data. Describe what you see—specifically with regards to unusual points? Identify this unusual point.

```
census.2010 = read.csv("data/census_2010.csv")

plot(census.2010$inf.mort ~ census.2010$doctors,
     pch = 20, cex = 1.2, col = "lightblue4", ylab = "Infant Mortality",
     xlab = "Doctors (per 100,000 members of pop.)",
     main = "Infant Mortality by Number of Doctors (2010)")
```




```
#identify the influential point
census.2010$state[census.2010$doctors > 700]
```

```
## [1] "District of Columbia"
```

Washington DC is the severe outlier in the plot: since it is an outlier in the predictor space (X = number of doctors), it can heavily influence the estimates of the β coefficients in regression, expecially the slope.

b) Fit a model predicting infant mortality rate from number of doctors using the complete data, then fit the same model but excluding the influential observation. On a single scatterplot, illustrate the effect of the influential point on the estimated model coefficients.

The model using the complete data has $\hat{\beta}_1 = 0.00205$ with $p = 0.33$, while the model excluding DC has $\hat{\beta}_1 = -0.00680$ with $p = 0.021$. The influential observation pulls up the model slope, obscuring the negative association among the 50 states.

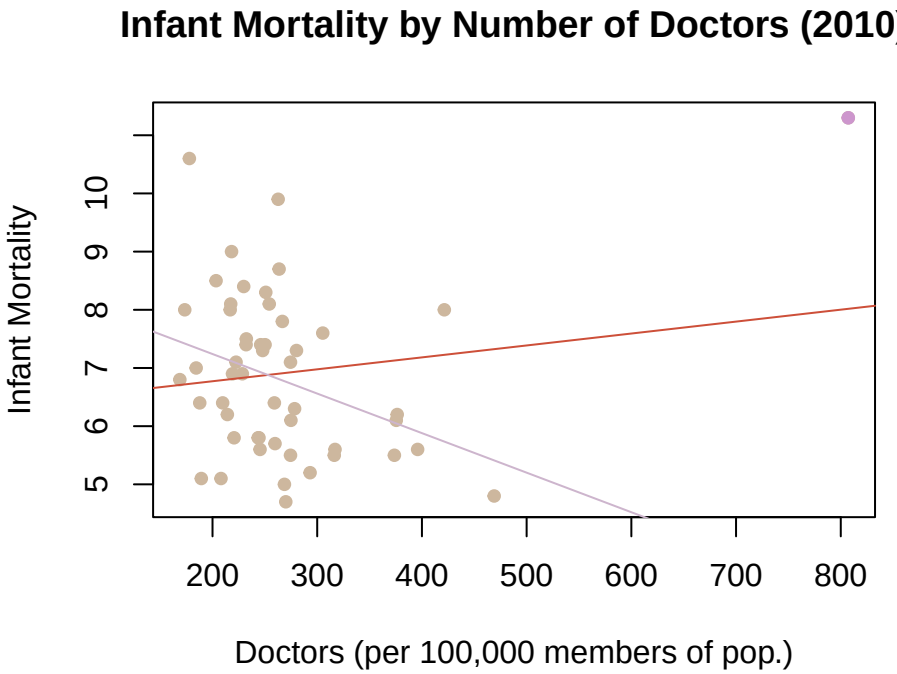
```
dc = (census.2010$state == "District of Columbia")

#fit model with outlier
model1 = lm(inf.mort ~ doctors, data = census.2010)
summary(model1)$coef
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 6.36220274 0.60022400 10.5997140 2.810514e-14
## doctors     0.00204961 0.00208126  0.9847929 3.295637e-01
```

```
#fit model without outlier
model2 = lm(inf.mort ~ doctors, data = census.2010[dc == F, ])
summary(model2)$coef
```

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  8.599060636 0.760333367 11.309593 3.873027e-15
## doctors     -0.006796864 0.002837472 -2.395394 2.055159e-02
```

```
#plot the models
plot(census.2010$inf.mort ~ census.2010$doctors,
     pch = 20, cex = 1.2, col = "bisque3",
     xlab = "Doctors (per 100,000 members of pop.)",
     ylab = "Infant Mortality",
     main = "Infant Mortality by Number of Doctors (2010)")
points(census.2010$inf.mort[dc] ~ census.2010$doctors[dc],
       pch = 20, col = "plum3", cex = 1.2)

abline(lm(census.2010$inf.mort ~ census.2010$doctors),
       col = "tomato3")
abline(lm(census.2010$inf.mort[dc == F] ~ census.2010$doctors[dc == F]),
       col = "thistle3")
```

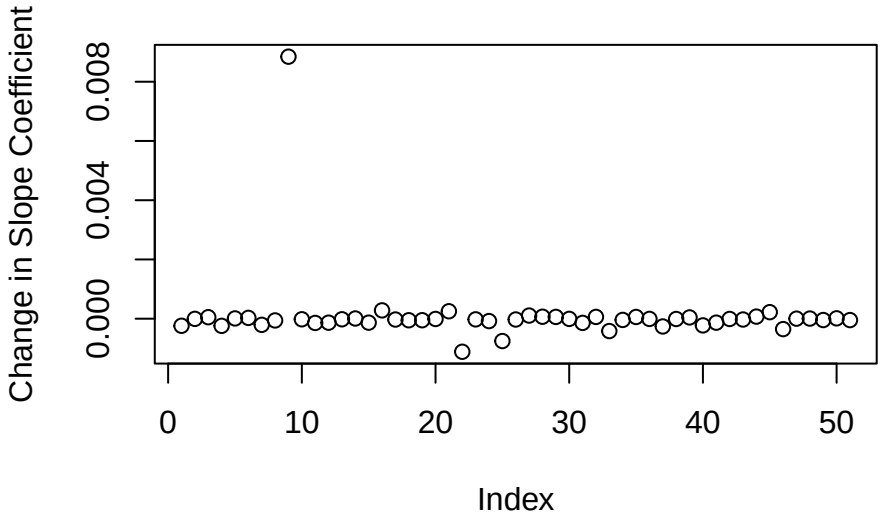


This difference in the cofficient estimate is called DFBETA, and can actually be computed without refitting the model:

$$DFBETA_i \equiv \mathbf{b} - \mathbf{b}_{(i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' e_i}{1 - h_{ii}}$$

where $\mathbf{b}$ denotes the estimated coefficients using all the data and $\mathbf{b}_{(i)}$ denotes the estimated coefficients with the ith subject removed, and h_{ii} denotes the ith diagonal element of the hat matrix, $X(X'X)^{-1}X'$.

```
dfbeta <- as.data.frame(dfbeta(model1))
plot(dfbeta[,2], ylab="Change in Slope Coefficient")
```




```
census.2010[which(dfbeta(model1)[,2] == max(dfbeta(model1)[,2])),]
```

```
##                  state inf.mort doctors
```

```
## 9 District of Columbia     11.3   807.2
```

c) What happens to the sampling distribution of $\hat{\beta}_1$ in the presence of an influential point? Apply a bootstrapping approach to the pairs of observations in the complete dataset and describe what you see.

The sampling distribution of $\hat{\beta}_1$ shows bimodality, rather than being unimodal and symmetric as typically expected. The mode a tad under 0.005 appears from the bootstrap samples that include one (or more) instances of Washington, DC, while the mode just below -0.005 occurs when DC is not selected in the bootstrap sample (see boxplot).

Note that DC should be included roughly 63% of the time ($n = 51$ here, but recall the limit):

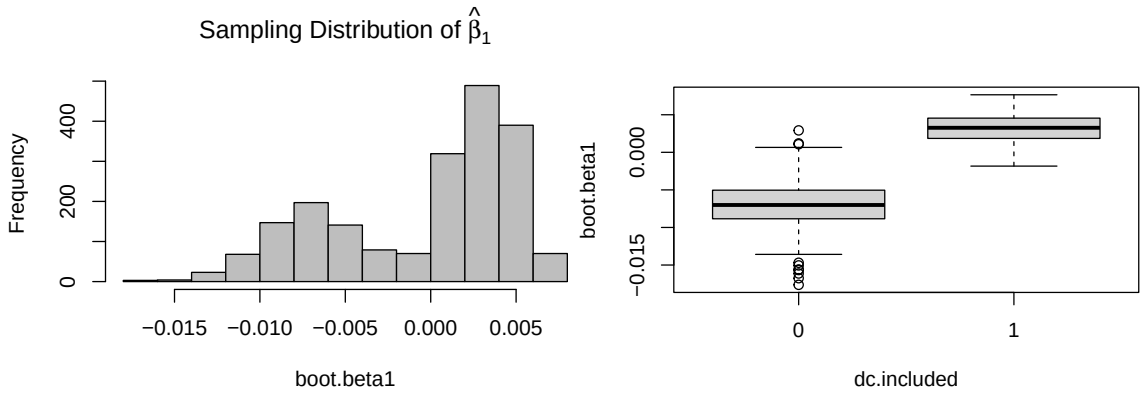
$$\lim_{n\to\infty} (1 - (1 - 1/n)^n) = e^{-1} \approx 0.632$$

We can derive this result. A bootstrap sample is generated by sampling with replacement from the data, and the probability that a particular observation is not chosen from a set of n observations is $1 - 1/n$, so the probability that the observation is not chosen n times is $(1 - 1/n)^n$. This is the probability that the observation does not appear in a bootstrap sample. The limit of this is $1/e$.

```
#set parameters
num.iterations = 2000
n = nrow(census.2010)

boot.beta1 = rep(NA, num.iterations)
dc.included = rep(NA, num.iterations)

#set seed
set.seed(139)

#resample
for(i in 1:num.iterations){

  boot.indices = sample(n, replace = TRUE)
  boot.sample = as.data.frame(census.2010[boot.indices, ])
  boot.lm = lm(inf.mort ~ doctors, data = boot.sample)

  boot.beta1[i] = coef(boot.lm)['doctors']
  dc.included[i] = 1*(sum(boot.indices==which(census.2010$doctors > 700))>0)
}
#how often is DC included
mean(dc.included)
```

```
## [1] 0.653
```

```
#plot sampling distribution
hist(boot.beta1, col = "gray",
     main = expression(paste("Sampling Distribution of ", hat(beta)[1])))

#compare DC when it is included vs. not included
boxplot(boot.beta1~dc.included)
```



d) From a model interpretation perspective, why might it be reasonable to exclude Washington, DC from an analysis of infant mortality and number of doctors based on this data?

If the goal of the analysis is to understand the relationship between infant mortality and number of doctors on a state level, then it is reasonable to exclude Washington, DC on the basis that it is quite unlike a state, which consists of a mix of urban centers and rural areas. It would make sense to include Washington, DC in an analysis comparing the infant mortality rate and number of doctors in each major urban center of the 50 states, for example.