POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

Outline:

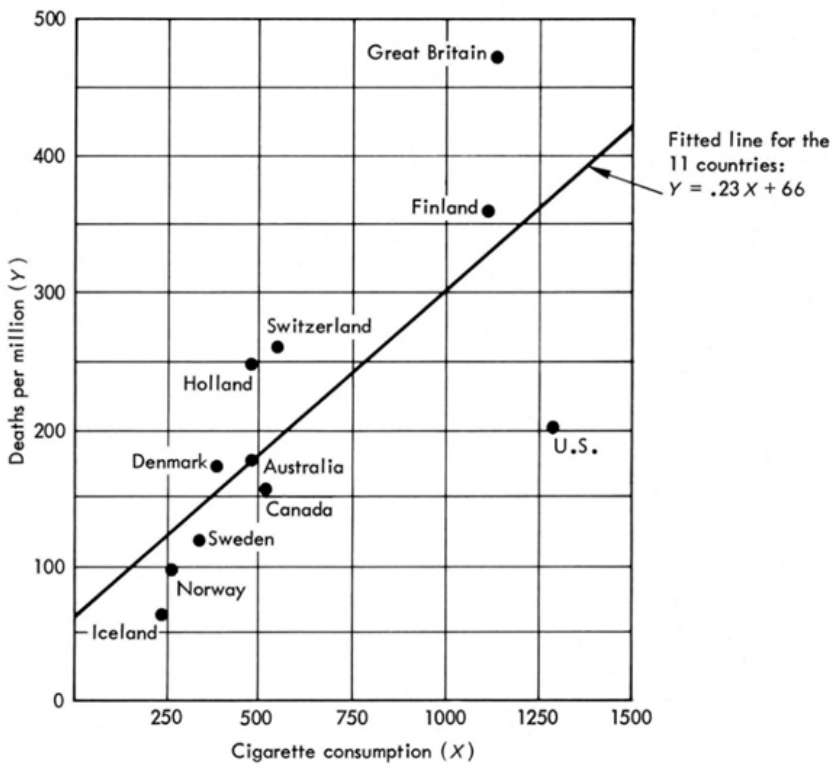
1. Definition of Influential Observations
2. Potential Causes of Influential Observations
3. Five Measures of (Potential) Influence
4. Detecting Influential Observations
5. Example: *cigarettes and deaths from lung cancer*

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

0. Motivation: *cigarettes and deaths from lung cancer*

source = Tufte, *Data Analysis for Politics and Policy*, p. 78-84

sample = 11 countries (9 western/northern Europe, Australia, United States)



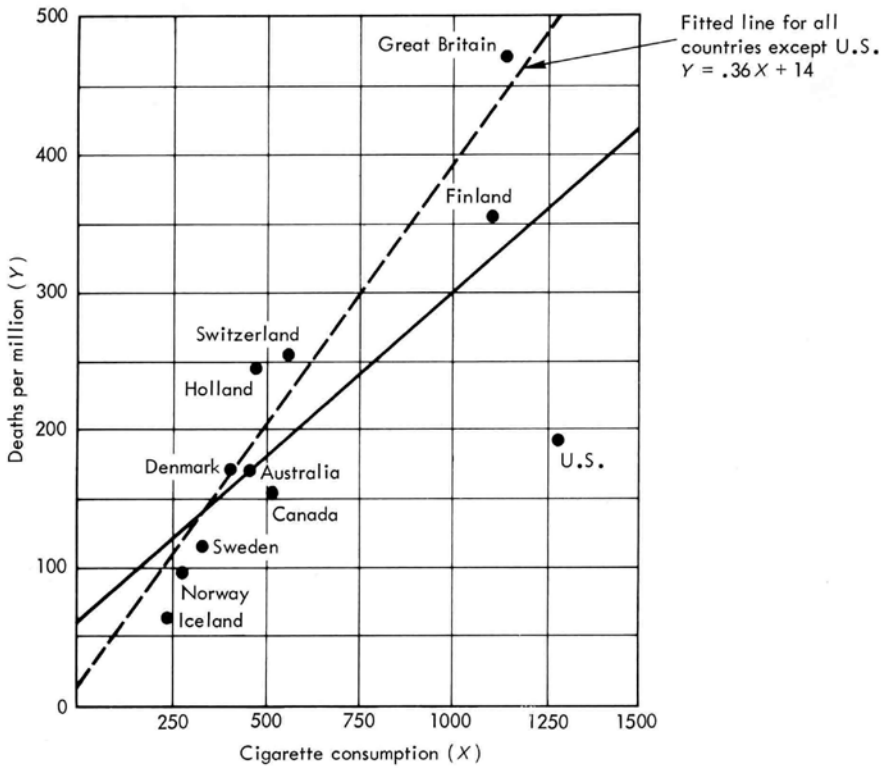


FIGURE 3-7 Lung cancer and cigarette consumption: fitted line for ten countries, omitting the United States

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

1. Definition of Influential Observations

"Loosely speaking, an observation is *influential* if dropping it from the analysis changes the key OLS estimates by a practically 'large' amount. …

"OLS is susceptible to outlying observations because it minimizes the sum of squared residuals: large residuals receive a lot of weight in the least squares minimization problem. If the estimates change by a practically 'large' amount when we slightly modify our sample, we should be concerned. "

Possible Consequences

1. Masking: failing to identify influential observations

   "[influential] observations tend to have small residuals… because these observations can force the regression [line] to be close to them"

2. Swamping: mistaking acceptable observations for unacceptable ones

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

2. Potential Causes of Influential Observations

1. Data entry errors
2. Small population
3. Disparate populations – i.e., naïvete (example of measuring room temperatures – where you measure in the *kitchen* matters a lot!)
4. Random chance (with normally distributed data, 1 in 22 obs. will be $\geq 2\sigma$ away from μ)

… and Possible Solutions (assuming there isn't a data coding or measurement error)

1. *Caveat emptor* (Let the buyer beware!) i.e., do nothing to address the possible problem
2. Exclude observations
3. Include observations but add dummy variables
4. Multivariate regression
5. Robust regression

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

3. Five Measures of Potential Influence

A. Leverage (p. 331 in passing)(Fox p. 24–25, 33)

- 'hat values' are found on the diagonal of the 'Hat matrix'
- sum of hat values (across all obs.) equals number of model parameters (p) estimated
- average hat value must equal $^{p}/_{n}$

In simple regression, $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$

Method 1. direct computation using this equation (for small datasets)
Method 2. in R, hatvalues(*Model*)
Method 3. in R, see ?lm.influence

Rule of thumb: consider correction for observation i if $h_i > 2 \times {}^{p}/_{n}$
(for simple regression, $p = 2$, so use $h_i > {}^{4}/_{n}$)

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

3. Five Measures of Potential Influence

B. Discrepancy (p. 328-329)(Fox p. 25–29, 33)

- *internally* studentized residuals (*i* included in estimate of *sigma*, i.e., $\widehat{\sigma_i}$)
- *externally* studentized residuals (*i* excluded from estimate of *sigma*, i.e., $\widehat{\sigma_{\sim i}}$)

In regression, $t_i = \frac{e_i}{\widehat{\sigma_{\sim i}} \cdot \sqrt{1-h_i}}$ where e_i is residual for i^{th} observation and h_i is hat value

Method 1. augment regression of *y* on *x* with dummy variable equal to 1 for one observation
its coefficient equals residual if regression line had been computed without it
its t-statistic equals its studentized residual (follows *t* distribution with $n-p-1$ d.f.)

Method 2. in R, rstandard(*Model*) gives you the internally studentized residuals (worse)
while rstudent(*Model*) gives you the externally studentized residuals (better)

Rule of thumb: consider correction for observation *i* if $|t_i| > t_{.025,\, n-p-1}$
(i.e., $|t_i| > 2$ in large enough samples)

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

3. Five measures of potential influence

C. Cook's Distance (Fox p. 30, 34)

- numerator is sum of squared <u>changes</u> in all observations' predicted values between two regressions:
  - first regression includes all obs., $j = 1, \ldots, n$ yields predictions $\widehat{y_j}$
  - second regression excludes one particular obs., i yields predictions $\widehat{y_{j(\sim i)}}$
- denominator is p+1 times <u>first</u> regression's Mean Squared Error: $\widehat{\sigma^2} = \frac{\sum(y_j - \widehat{y_j})^2}{n-p}$
- combines leverage and discrepancy

In regression, $D_i = \frac{\sum(\widehat{y_j} - \widehat{y_{j(\sim i)}})^2}{(p+1)\cdot\widehat{\sigma^2}}$

Alternatively: $D_i = \frac{e_i'^2}{p} \cdot \frac{h_i}{(1-h_i)^2}$ (e_i' is defined on Fox p. 25… similar to t_i but includes i)

Method 1. in R, cooks.distance(*Model*) – see ?influence.measures

Rule of thumb: consider correction for observation i if $D_i > \frac{4}{n-p}$

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

3. Five measures of potential influence

D. DfBETA (Fox p. 29–30, 34)

- examines whether the coefficient on x_k changes when a case i is deleted

In regression, $d_{ik} = \hat{\beta}_k - \hat{\beta}_{k(\sim i)}$

Method 1. run model with observation i and model without it; compare slopes, intercepts
Method 2. in R, dfbeta(*Model*) – see ?influence.measures

Rule of thumb: consider correction for observation i if $d_{ik} > \text{se}(\hat{\beta}_{k(\sim i)})$ or if $d_{ik} > 2 \cdot \text{se}(\hat{\beta}_{k(\sim i)})$

3. Five measures of potential influence

E. DfFITS (Fox p. 30, 34)

- combines leverage and influence, but looks at changes in predicted values of *y*

In regression, $DFFITS_i = t_i \times \sqrt{\frac{h_i}{1-h_i}}$ (where h_i is hat value, t_i is studentized residual)

$$DFFITS_i = \frac{\widehat{y_i} - \widehat{y_{i(\sim i)}}}{\widehat{\sigma_{\sim i}} \cdot \sqrt{h_i}}$$

Method 1. in R, dffits(*Model*) – see ?influence.measures

Rule of thumb: consider correction for observation *i* if $DFFITS_i > 2\sqrt{\frac{p}{n-p}}$

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

4. Detecting Influential Observations

A. Visual inspection

1. Scatterplot leverage vs. studentized residuals (the two elements of Cook's distance)
   in R, influencePlot(*Model*, …) (requires the *car* package)
2. Scatterplot residuals vs. fitted values
3. Normal quantile plot
4. Half-normal plot (requires the *faraway* package)
5. Histogram and/or stemplot of residuals

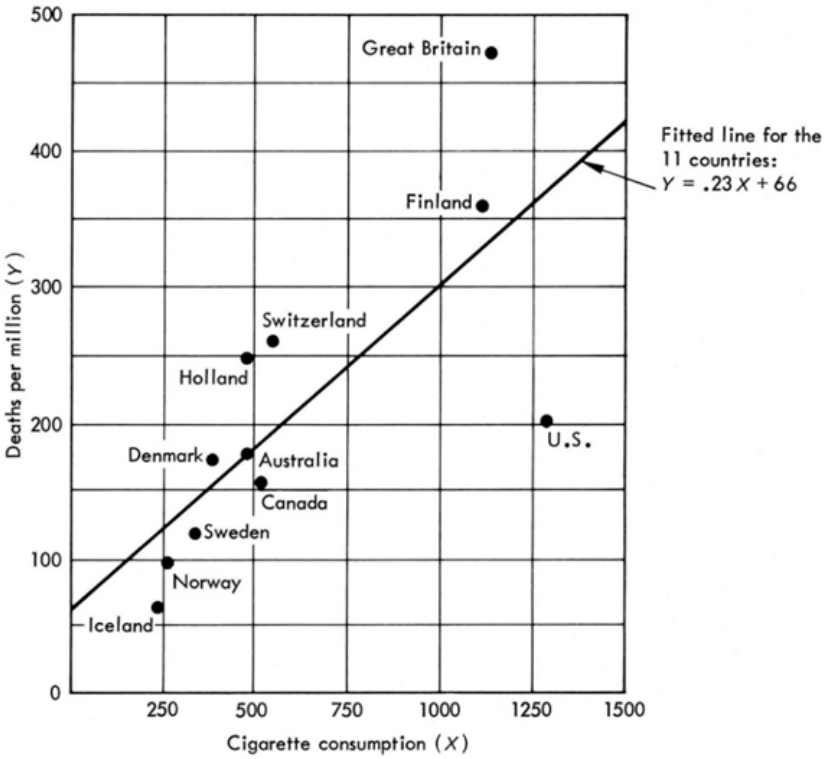
B. Global tests of normality of residuals
(We'll return to this in a few weeks; foreshadowing: the Jarque-Bera test checks skewness and kurtosis)

5. Example: *cigarettes and deaths from lung cancer*

source = Tufte, *Data Analysis for Politics and Policy*, p. 78-84

sample = 11 countries (9 western/northern Europe, Australia, United States)

Country	Y_i = observed lung cancer deaths per million in 1950	X_i = cigarettes consumed per capita in 1930
Iceland	58	220
Norway	90	250
Sweden	115	310
Canada	150	510
Denmark	165	380
Australia	170	455
United States	190	1280
Holland	245	460
Switzerland	250	530
Finland	350	1115
Great Britain	465	1145



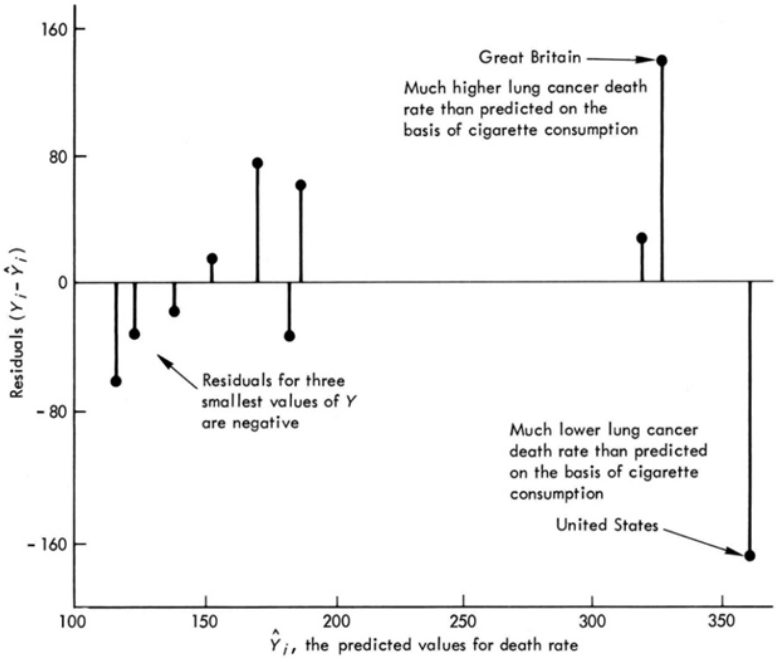
POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

5. Example: *cigarettes and deaths from lung cancer*

Calculate fitted values, calculate residuals, and plot them against each other

Residual Analysis

Y_i = observed lung cancer deaths per million in 1950	X_i = cigarettes consumed per capita in 1930	$\hat{Y}_i$ = predicted lung cancer death rate for a given X_i, $\hat{Y}_i = .23X_i + 66$	Residual = observed − predicted = $Y_i - \hat{Y}_i$
58	220	.23(220) + 66 = 116	58 − 116 = −58
90	250	.23(250) + 66 = 123	90 − 123 = −33
115	310	.23(310) + 66 = 137	115 − 137 = −22
150	510	.23(510) + 66 = 183	150 − 183 = −33
165	380	.23(380) + 66 = 153	165 − 153 = 12
170	455	.23(455) + 66 = 170	170 − 170 = 0
190	1280	.23(1280) + 66 = 359	190 − 359 = −169
245	460	.23(460) + 66 = 171	245 − 171 = 74
250	530	.23(530) + 66 = 187	250 − 187 = 63
350	1115	.23(1115) + 66 = 321	350 − 321 = 29
465	1145	.23(1145) + 66 = 328	465 − 328 = 137



POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

5. Example: *cigarettes and deaths from lung cancer*

R code to estimate models:

```
library(foreign)
lung <- read.dta("C:\Users\sjbasing\TufteTable3-3.dta")

Model11 <- lm(deaths ~ cigarettes, lung)
Model11$coefficients

lung$dummy <- ifelse(lung$Country=="United States",c(1),c(0))
Model11D <- lm(deaths ~ cigarettes + dummy, lung) # option 1: include dummy for US
Model11D$coefficients

remove = -c(11)
Model10<- lm (deaths ~ cigarettes, lung, subset =remove) # option 2: omit US
Model10$coefficients
```

model estimated with 11 countries: slope = .229; intercept = 65.75

model estimated with 10 countries (US deleted/dummied): slope = .358; intercept = 13.55

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

5. Example: *cigarettes and deaths from lung cancer*

R code to graph data points and regression lines

```
plot(cigarettes, deaths, pch = 16)
abline(lm(deaths~cigarettes))
text(cigarettes, deaths, labels=Country, pos=3, cex=.6, xpd=TRUE)

abline(lm(deaths~cigarettes, subset=remove), col="blue")
```

POLS 6481. Research Design and Quantitative Methods II
Lecture 4. Influential Data Points
Readings: Wooldridge, *Introductory Econometrics 5e*, 9.5; *Regression Diagnostics*, p. 21-40

5. Example: *cigarettes and deaths from lung cancer*

R code to evaluate diagnostics

```
influence <- lm.influence(Model11)
leverage <- influence$hat
discrep <- rstudent(Model11) # externally studentized
cooks <- cooks.distance(Model11)
dffits <- dffits(Model11)

library(car)
influencePlot (Model11, main="Influence Plot")
```

* You could replace the third line above with:
discrep <- rstandard(Model11) # internally studentized

Highest leverage: United States (0.399, vs. threshold of .364 ($=4/n$))
Highest studentized residual: United States (-4.85, vs. threshold of 2.31 ($= t_{.025, n-3}$))

5. Example: *cigarettes and deaths from lung cancer*