Regression Diagnostics

Andrew Zieffler

Anscombe's Data (1973)

- Four contrived data sets that produce the EXACT SAME regression output
  - Coefficient estimates (intercept and slope)
  - Coefficient standard errors
  - Coefficient-level *t*- and *p*-values
  - Correlation
  - Model standard error (RMSE)

$n = 11$

$R^2 = 0.67$

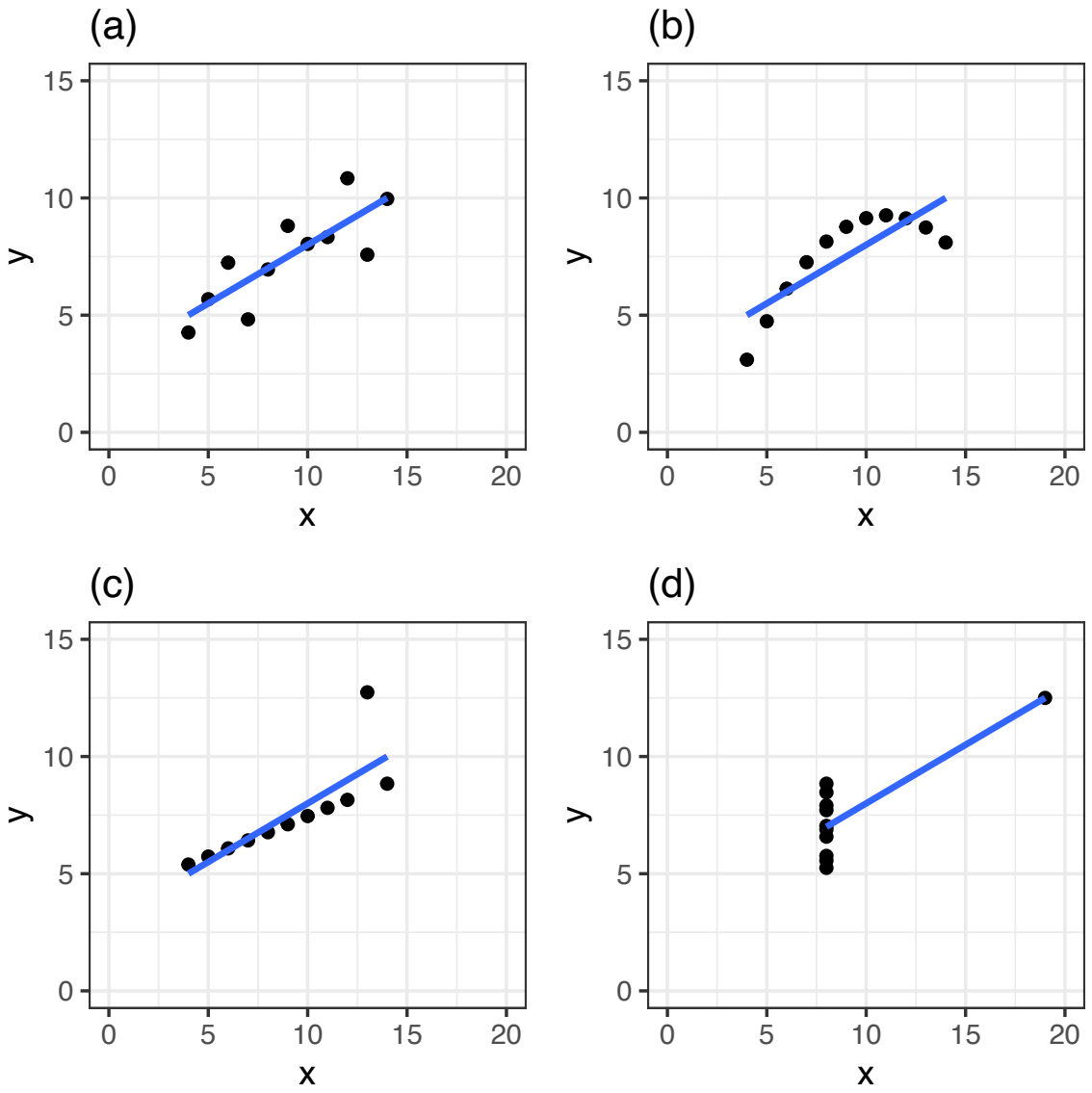
$\text{RMSE} = 1.24$

Coefficient	Estimate	SE	*t*	*p*
B_0	3.00	1.12	2.67	0.026
B_1	0.50	0.12	4.24	0.002

Do these model-and coefficient-level summaries provide an adequate description/summarization of the data?

- Despite having the same regression output,
  - The data (and hence the residuals) are markedly different

- The relationship in (a) seems reasonably linear; regression output adequately describes this.
- The relationship in (b) is curvilinear; regression output does not adequately describe this.
- In (c), there is a single observation that is influencing the regression line; regression output does not adequately describe this.
- In (d), if not for the observation with an *x*-value near 20 we would not be able to fit a line at all; regression output does not adequately describe this



Alberto Cairo has created a newer dataset that emphasizes visualization, the *Datasaurus Dozen*: https://www.autodeskresearch.com/publications/samestats

Key Message

There are several ways in which the linear relationship may fail

- Nonlinearity
- Outlying data
- Influential data
- Collinearity

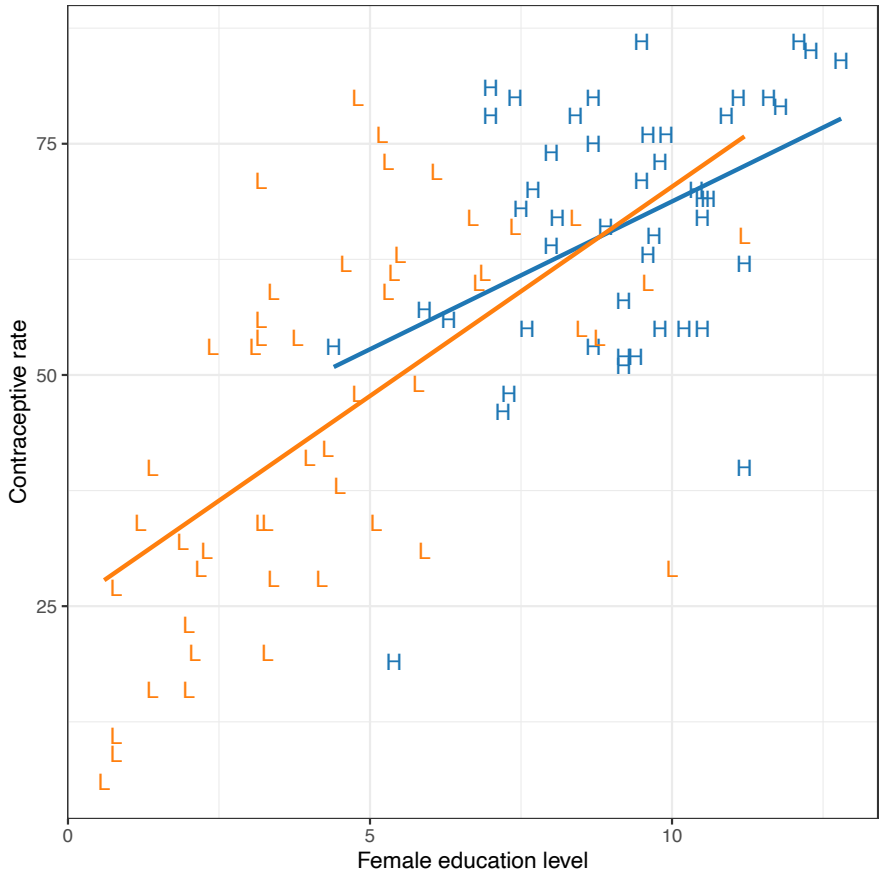
The conventional regression summaries do not tell the entire story. Diagnostic tools such as residual plots and other measures help fill in gaps.

Case Study

- **Data:** *contraception.csv*

- **Goal:** Fit model to estimate effect of female education level variation on contraception rate. Also to examine whether this effect differs for countries with low (L) and high (H) GNI.

Scatterplot suggests a potential interaction between female education level and GNI level on contraceptive rate.



Model-Level Output

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0.497	0.480	14.4	30.6	7.57e-14	3	-394.	799.	812.	19296.	93	97

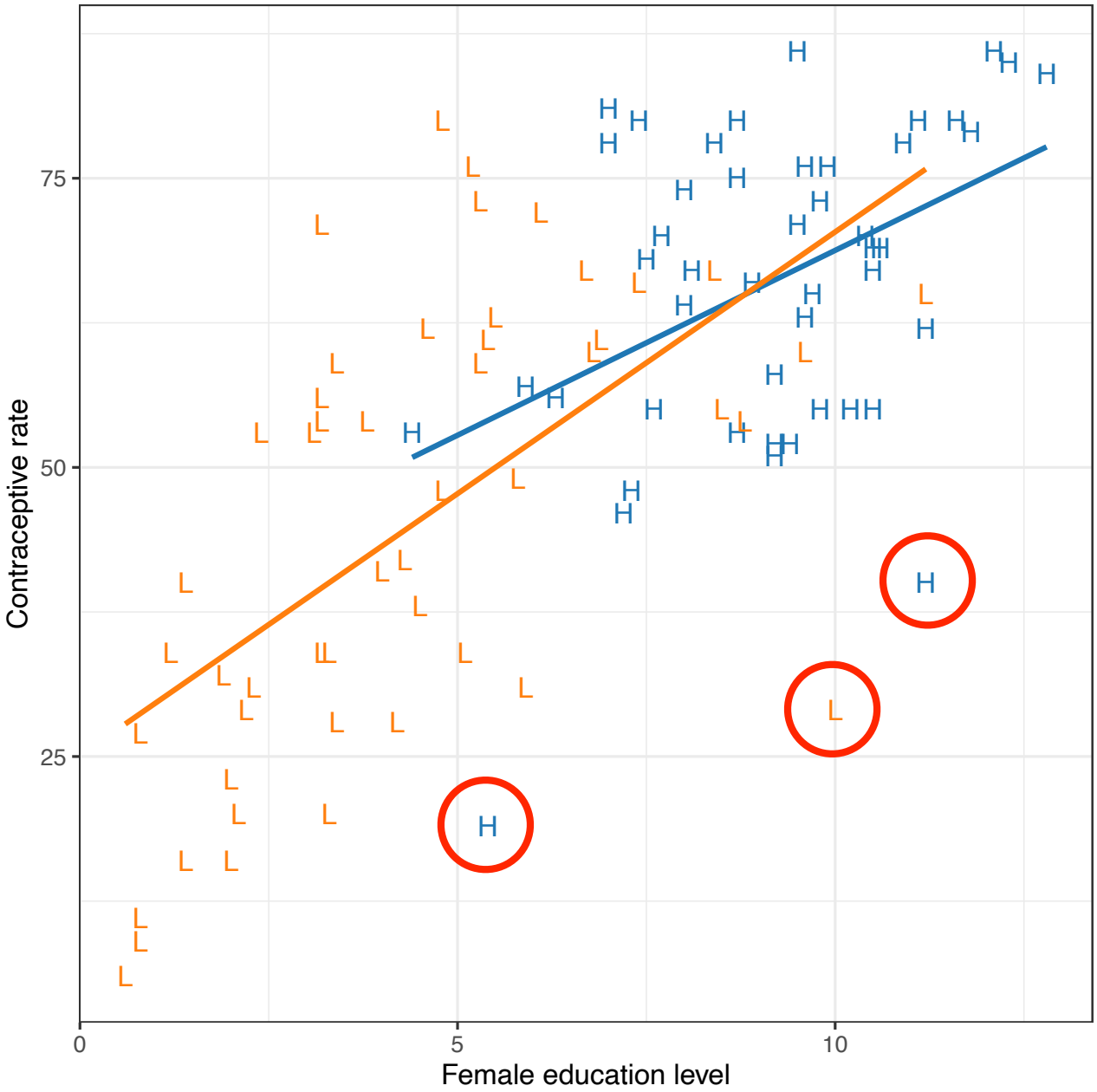
Coefficients

term	estimate	std.error	statistic	p.value	conf.low	conf.high
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	25.1	4.06	6.18	0.0000000168	17.0	33.2
educ_female	4.53	0.798	5.67	0.000000158	2.94	6.11
high_gni	11.7	11.4	1.03	0.307	-10.9	34.3
educ_female:high_gni	-1.33	1.39	-0.956	0.341	-4.09	1.43

If we were to interpret these results, we would conclude that:

- The model explains 49.7% of the variation in contraceptive rates.
- However, there is too much uncertainty to conclude there is an interaction effect.

These results, however, may be influenced by several observations that seem to deviate from the fitted regression lines.



What happens if we remove the three observations in question?

Model-Level Output

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0.586	0.572	12.9	42.4	3.57e-17	3	-372.	753.	766.	14935.	90	94

Coefficients

term	estimate	std.error	statistic	p.value	conf.low	conf.high
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	22.5	3.71	6.06	3.10e- 8	15.1	29.9
educ_female	5.33	0.751	7.09	2.88e-10	3.84	6.82
high_gni	20.2	10.8	1.87	6.46e- 2	-1.25	41.6
educ_female:high_gni	-2.61	1.32	-1.98	5.06e- 2	-5.23	0.00696

If we were to interpret these results, we would conclude that:

- The model explains 58.6% of the variation in contraceptive rates.
- There is now some evidence that there may be an interaction effect.

Comparing the two sets of results:

- The model suggested by the data switched from a main-effects model to an interaction model.
- The R^2 value increased (0.497 —> 0.586) after omitting the three observations.
- The RMSE value decreased (14.4 —> 12.9) after omitting the three observations.
- The coefficient estimates and SEs for important predictors (i.e., focal effects of female education level), changed.
- The interpretation of results also changed dramatically.

This suggests that regression outliers can impact:

- Empirical model specification;
- Model-level fit;
- Coefficient-level estimation; and
- Interpretations.

When this happens, the regression model fails to be a good summary of the data.

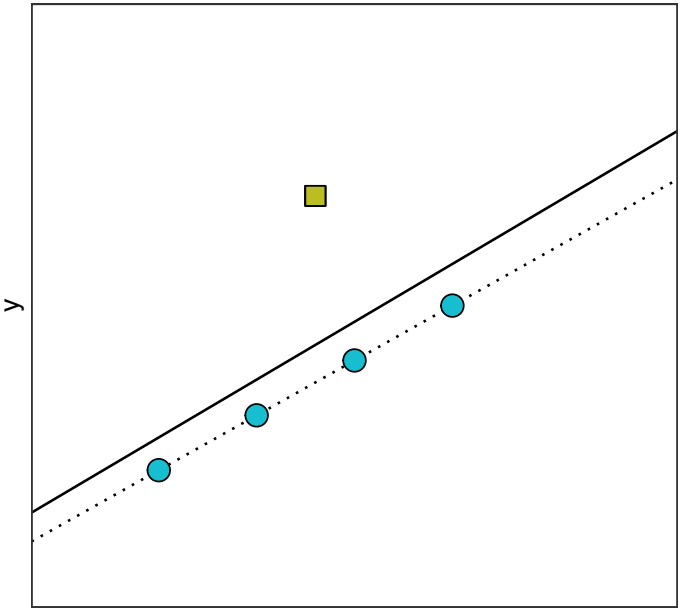
What Properties Lead to Influence?

A **regression outlier** is an observation that has an unusually low/ high residual value given the set of predictor values.

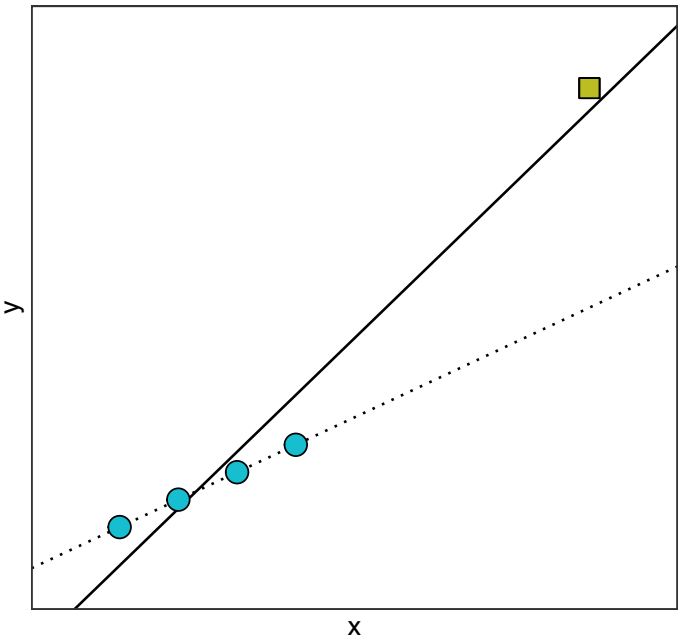
- The yellow (square) observation in (a) is a regression outlier.
  - It is, however, not an outlier in the *x*-distribution
  - Because it is not an outlier in its *x*-value, we say this observation has **low leverage**

- The yellow (square) observation in (b) is a regression outlier.
  - It is also an outlier in the *x*-distribution
  - Because it is an outlier in its *x*-value, we say this observation has **high leverage**

(a)



(b)

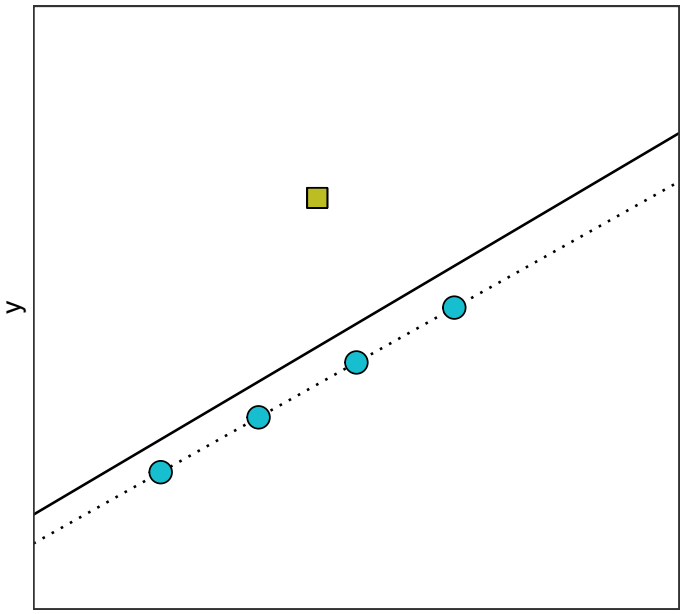


An observation that is both a **regression outlier** and has **high leverage** produces substantial influence on the regression coefficients.

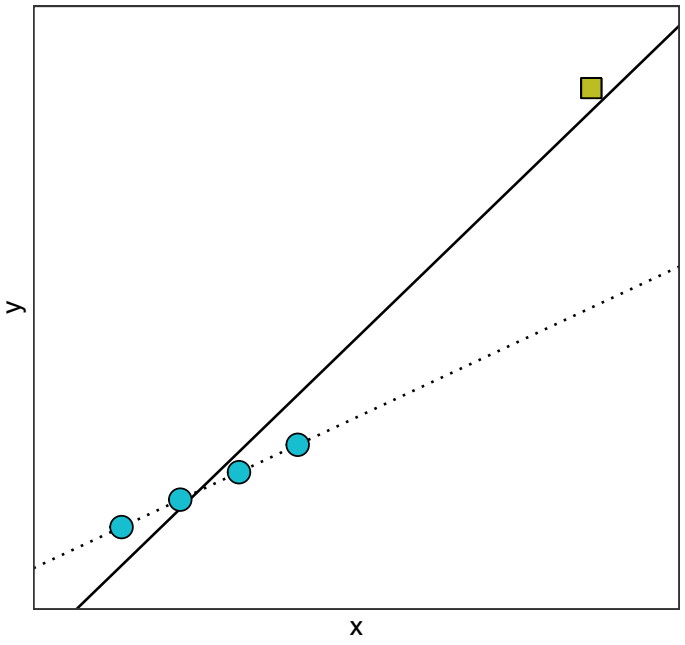
- The yellow (square) observation in (a) is a **regression** outlier, but has **low leverage**.
  - Deleting this outlier will have little to no effect on the slope and intercept

- The yellow (square) observation in (b) is a **regression** outlier *and* has **high leverage**.
  - Deleting this outlier will have a large effect on the slope and intercept

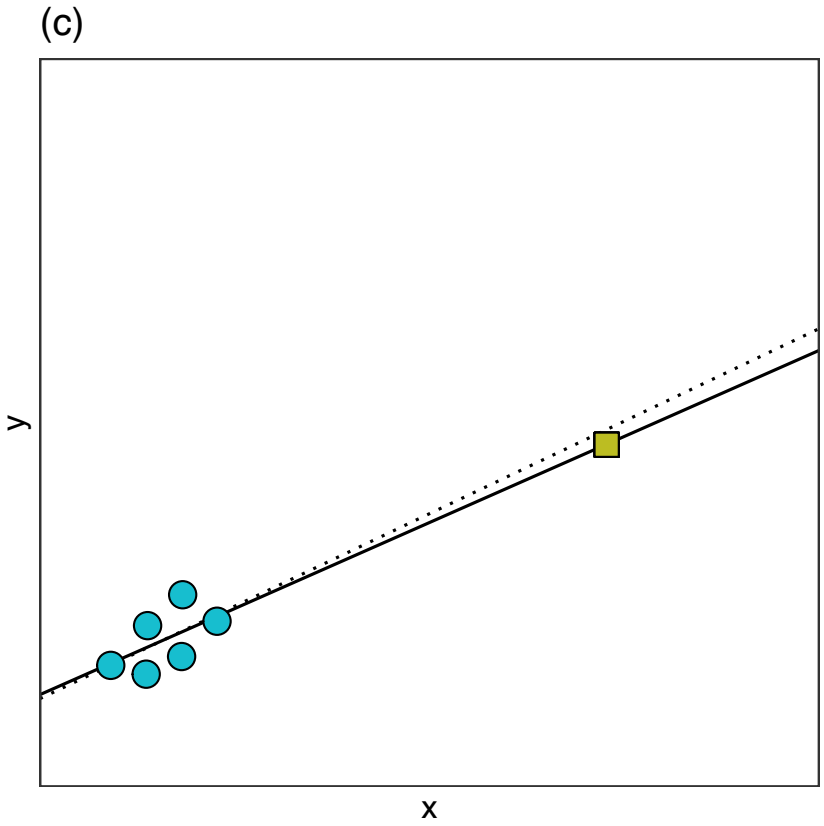
(a)



(b)



- The yellow (square) observation in (c) has **high leverage** *but* is a not a regression outlier.
  - Deleting this outlier will have almost no effect on the slope and intercept



It is the combination of **leverage** with being a **regression outlier** that produces influence on the regression coefficients. We call these observations **influential observations**. Heuristically,

$$\text{Influence} = \text{Leverage} \times \text{Outlyingness}$$

Leverage

Computation of Hat Values

- The hat value is a common measure of leverage
  - Quantified by examining the distances between *x*-values, accounting for correlation
  - Leverage values are the diagonal elements of the hat matrix, **H**
  - Denoted h_{ii}

Recall that

$$\mathbf{\hat{Y} = HY}$$

where

$$\mathbf{H = X(X^T X)^{-1} X^T}$$

Formula for Simple Regression

For simple linear regression (one predictor) we can also use the following computational formula. This is presented to help understand how the hat values measure leverage.

$$
\begin{aligned}
h_{ii} &= \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{1}{n} + \frac{1}{n-1}\left(\frac{x_i - \bar{x}}{s_x}\right)^2
\end{aligned}
$$

The hat-values are defined by the observation's distance to the mean *x*-value. In multiple regression, this analytical formula no longer works, but conceptually, the hat-values would measure the observation's distance to the centroid (multivariate mean) accounting for the correlation between the predictors.

The h_{ii} values are essentially equal to the ratio of the covariance between the observed and predicted values to the variance of the observed values.

$$h_{ii} \approx \frac{\mathrm{Cov}(\hat{Y}_i, Y_i)}{\mathrm{Var}(Y_i)}$$

Because we are normalizing the covariances by dividing by the variance of Y_i, then

$$0 \leq h_{ii} \leq 1$$

Also, note that

$$\sum_{i=1}^{n} h_{ii} = p$$

where *p* is the number of parameters in the model. This implies

$$\bar{h} = \frac{p}{n}$$

Using one of Anscombe's datasets, we can illustrate leverage via the hat-values. There are two *x*-values represented in the data. Observations with the same *x*-value have the same leverage value. The observation with the hat-value of 1 shows a large deviation from the mean *x*-value.

$$0 \leq h_{ii} \leq 1$$

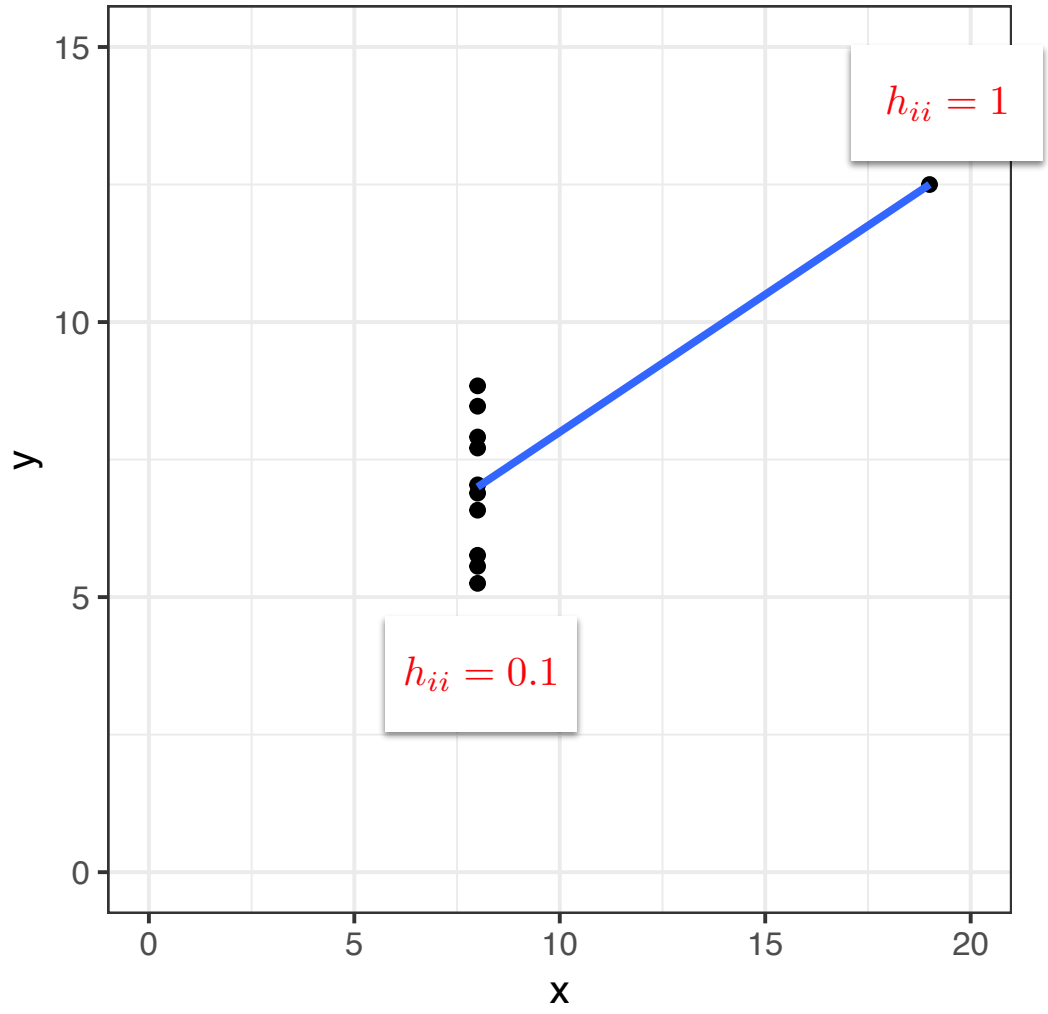
$$\sum_{i=1}^{n} h_{ii} = p$$

$$\sum_{i=1}^{n} h_{ii} = 2$$

$$\bar{h} = \frac{p}{n}$$

$$\bar{h} = \frac{2}{11} = 0.18$$

The hat-value of 1 is roughly 5 times larger than the average hat-value, indicating that the observation has high leverage, whereas a hat-value of 0.1 is close to the average hat-value (low leverage).



Problem with High Leverage

Our regression model is

$$\mathbf{y} = f(\mathbf{x}, \boldsymbol{\beta}) + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon)$$

When we estimate the regression coefficients, the resulting residuals, e_i, do not have a constant variance. Observations with higher leverage values have a lower variance. Why?

The variance of each fitted residual, e_i, is

$$\mathrm{Var}(e_i) = s_e^2(1 - h_{ii})$$

Where s_e^2 is the unbiased estimator of σ^2_ε,

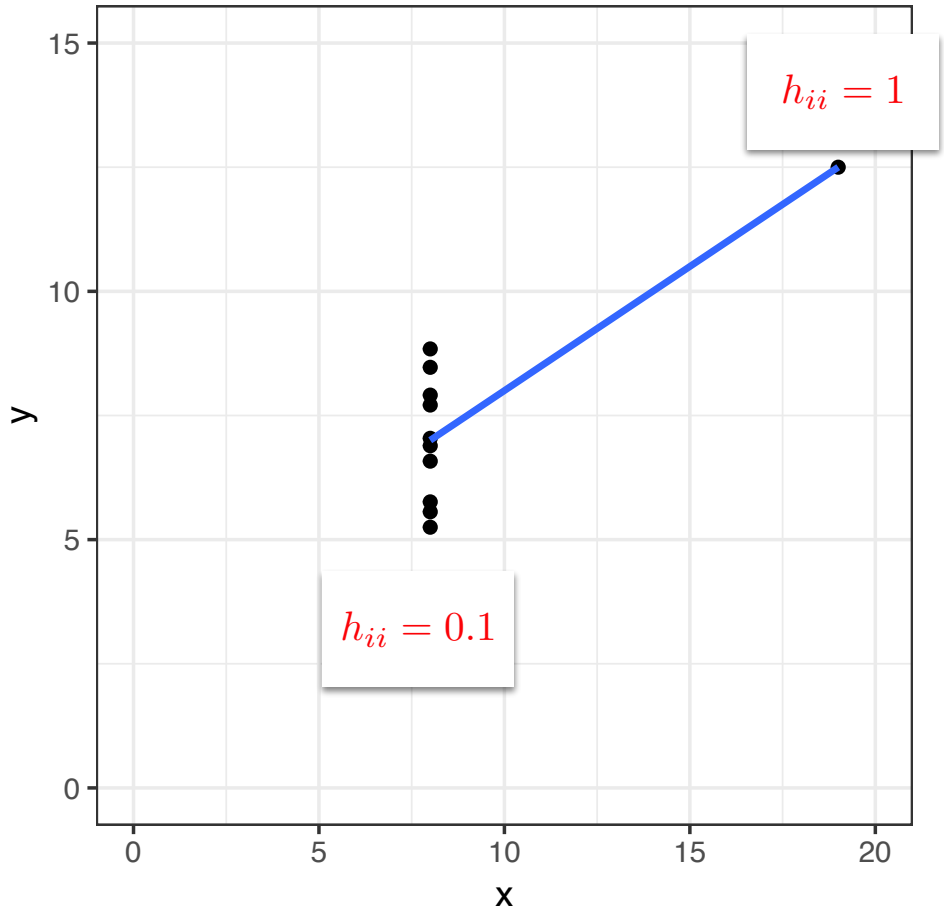
$$s_e^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$$

Looking at this formula,

$$\text{Var}(e_i) = s_e^2(1 - h_{ii})$$

as leverage approaches 1 (high leverage) the estimated variance of the fitted residual approaches zero.

The estimated variance for the observation with a leverage value of 1 is 0. The fitted line passes though the observation.



Leverage (or hat-) values are produced as a column in the output obtained using the `augment()` function from the **broom** package.

```
# Obtaining leverage values
out1 = augment(lm.1)
head(out1)
```

```
  contraceptive educ_female high_gni .fitted  .resid .std.resid   .hat .sigma   .cooksd
          <dbl>       <dbl>    <dbl>   <dbl>   <dbl>      <dbl>  <dbl>  <dbl>     <dbl>
1            57         5.9        1    55.7  -1.34      0.0976 0.0875   14.5 0.000228
2            66         8.9        1    65.2  -0.752     0.0528 0.0217   14.5 0.0000155
3            55        10.5        1    70.4  15.4      -1.08   0.0326   14.4 0.00990
4            62         4.6        0    45.9 -16.1       1.13   0.0201   14.4 0.00653
5            67        10.5        1    70.4   3.36     -0.237  0.0326   14.5 0.000474
6            51         9.2        1    66.2  15.2      -1.07   0.0213   14.4 0.00619
```

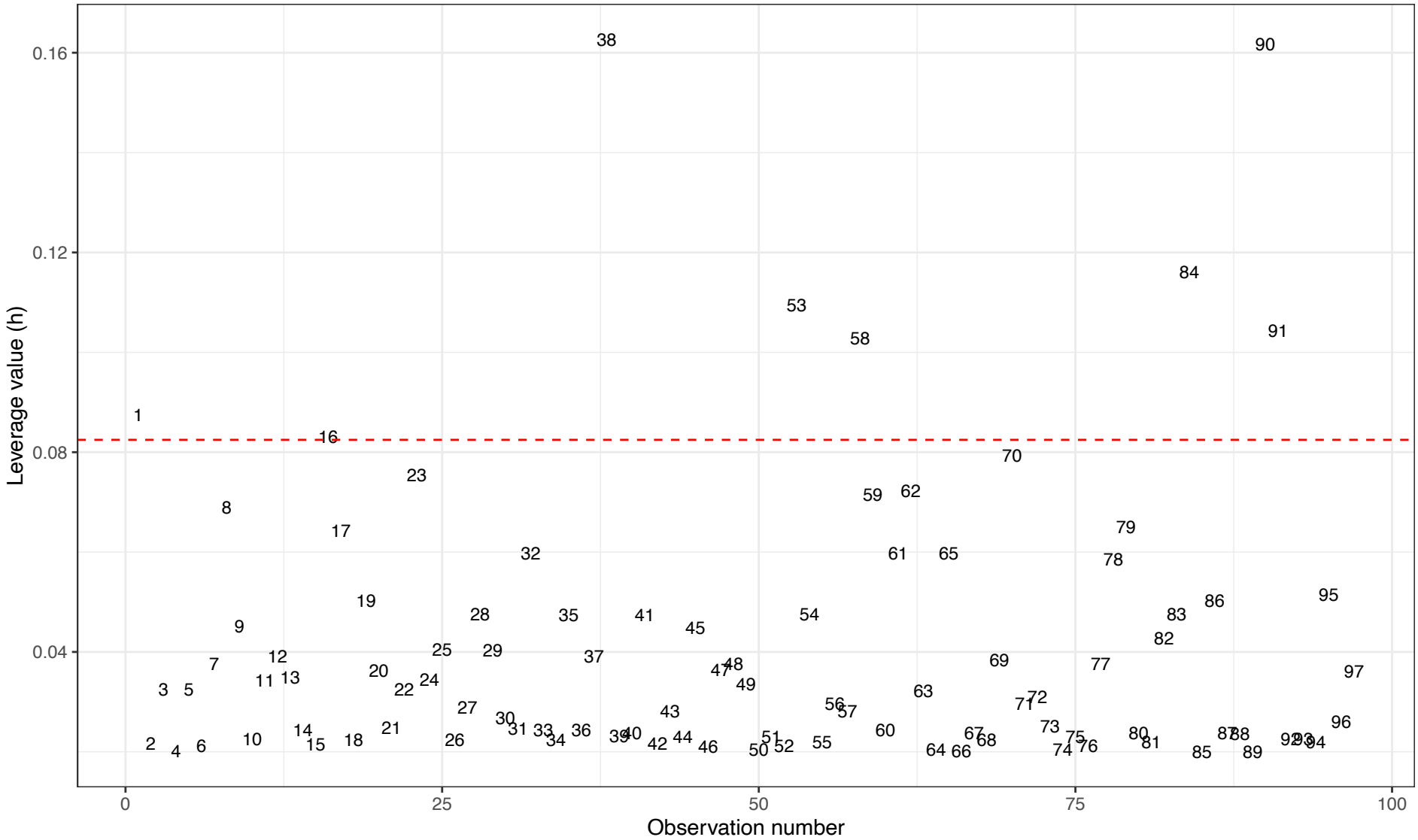
```
# Average leverage value (h-bar)
mean(out1$.hat)
```

```
[1] 0.04123711
```

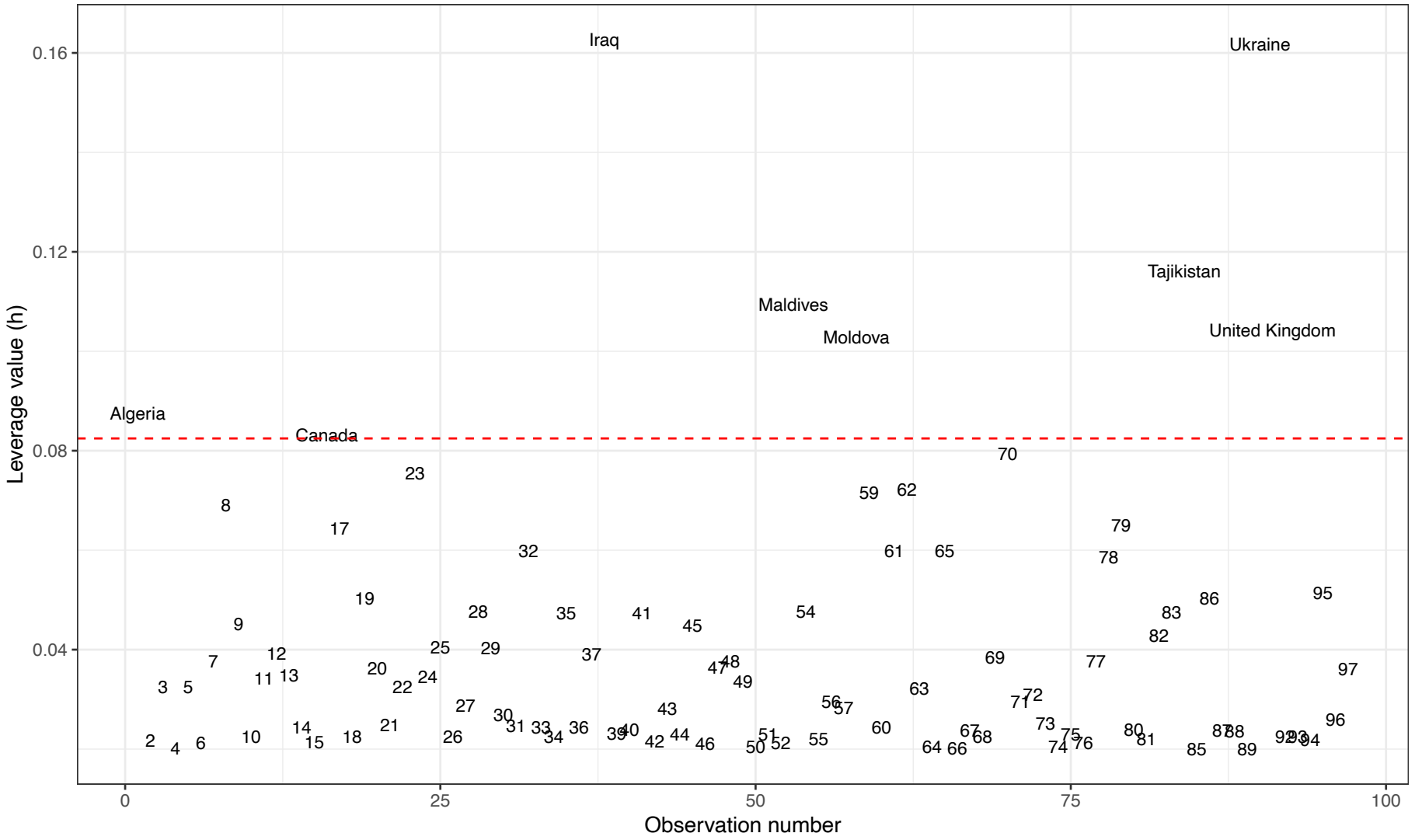
```
# Find large leverage values: h >= 2(h-bar)
out1 %>% filter(.hat >= 2 * 0.04123711) %>% arrange(desc(.hat))
```

```
  contraceptive educ_female high_gni .fitted .resid .std.resid   .hat .sigma  .cooksd
          <dbl>       <dbl>    <dbl>   <dbl>  <dbl>      <dbl>  <dbl>  <dbl>    <dbl>
1            53         4.4        1    50.9  -2.14      0.162  0.163    14.5 0.00128
2            65        11.2        0    75.8  10.8      -0.818  0.162    14.4 0.0323
3            29        10          0    70.4  41.4      -3.05   0.116    13.7 0.307
4            19         5.4        1    54.1  35.1      -2.58   0.109    14.0 0.204
5            84        12.8        1    77.7  -6.29      0.461  0.104    14.5 0.00619
6            60         9.6        0    68.5   8.55     -0.627  0.103    14.5 0.0113
7            57         5.9        1    55.7  -1.34      0.0976 0.0875   14.5 0.000228
8            85        12.3        1    76.1  -8.88      0.644  0.0831   14.5 0.00940
```

You can also look at an index/case plot of the leverage values (y = leverage vs. x = row number). Here, instead of plotting points, I have plotted the text of the row number. There is a solid line drawn at the average hat-value and a dotted line drawn at 2 times the average hat-value.



An alternative index/case plot would indicate the countries with high leverage values.



Regression Outliers

Regression Outliers

We can create a **standardized residual** by dividing each fitted residual by its standard error. Since,

$$\text{Var}(e_i) = s_e^2(1 - h_{ii})$$

$$\text{SE}(e_i) = s_e\sqrt{(1 - h_{ii})}$$

The **standardized residual** (aka, *internally studentized residual*) is

$$e_i' = \frac{e_i}{s_e\sqrt{(1 - h_{ii})}}$$

To evaluate whether a particular standardized residual is extreme, we test whether its value differs from 0 using a hypothesis test. Since we are using the data to estimate s_e and h_{ii}, we would use a *t*-test.

(Externally) Studentized Residuals

The problem with testing whether the standardized residuals differ from 0 is that, since the numerator and denominator are not independent (the term s_e in the denominator is a function of e_i.), the resulting statistic is not *t*-distributed.

To fix this problem, we can compute an estimate of s_e that is computed based on the regression deleting the i^{th} observation.

$$t_i = \frac{e_i}{s_{e(-i)}\sqrt{(1-h_{ii})}}$$

Since the numerator and denominator are now independent, the resulting statistic is *t*-distributed with *n–k–1 df* (where *k* is the number of predictors). This new measure is referred to as a **studentized residual**.

Terminology: Studentized residuals are also sometimes referred to as: *deleted studentized residuals*, *externally studentized residuals,* and in some cases, even as *standardized residuals*. Ugh.

Here is the `augment()` function's output for the first observation.

	contraceptive <dbl>	educ_female <dbl>	high_gni <dbl>	.fitted <dbl>	.resid <dbl>	.std.resid <dbl>	.hat <dbl>	.sigma <dbl>	.cooksd <dbl>	id <int>	country <chr>
1	57	5.9	1	55.7	-1.34	0.0976	0.0875	14.5	0.000228	1	Algeria

Estimate of residual standard deviation when corresponding observation is dropped from model

The residual standard deviation (RMSE) for the model was 14.4. Without Algeria in the model the RMSE would be 14.5.

It is important to note that the `.std.resid` column is the standardized residual value. **The studentized residual is not given.** However, we could compute the studentized residual from other columns.

$$t_i = \frac{e_i}{s_e(-i)\sqrt{1-h_i}}$$

```
# Compute studentized residual
1.34 / (14.5 * sqrt(1 - 0.0875))

[1] 0.09674318
```

We can obtain the standardized and studentized residuals directly using the `rstandard()` and `rstudent()` functions. Below we show this output for the first observation only.

```
# Compute standardized residuals
rstandard(lm.1)

         1
0.09756739

# Compute studentized residuals
rstudent(lm.1)

         1
0.09704638
```

Studentized Residuals: Mean-shift model

We can also compute the studentized residuals by fitting our model and also including a dummy variable *d* that is 1 for observation *i* and 0 otherwise. For our interaction model, the model to compute the studentized residuals is:

$$\text{Cont Rate}_i = \beta_0 + \beta_1(\text{Education}_i) + \beta_2(\text{GNI}_i) + \beta_4(\text{Education}_i)(\text{GNI}_i) + \beta_5(d_i) + \epsilon_i$$

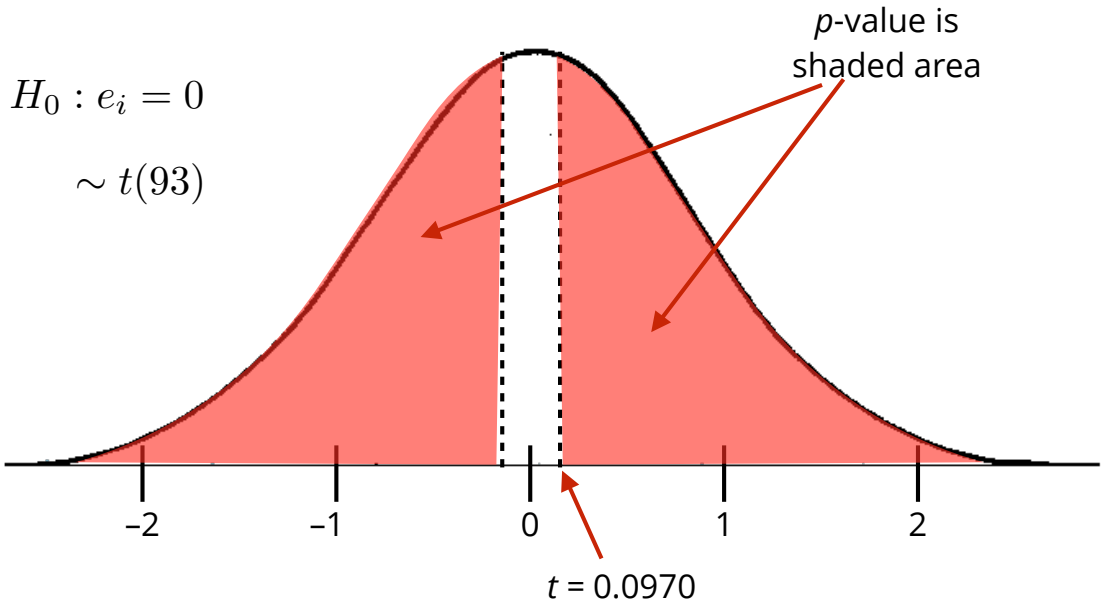
Fitting this model to the contraception data where *d* is 1 for *Observation 1* and 0 otherwise, the coefficient-level output is:

	term	estimate	std.error	statistic	p.value	conf.low	conf.high
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	25.1	4.08	6.15	0.0000000201	17.0	33.2
2	educ_female	4.53	0.802	5.64	0.000000184	2.93	6.12
3	high_gni	11.4	11.9	0.960	0.340	-12.2	35.0
4	obs_1	1.47	15.2	0.0970	0.923	-28.6	31.6
5	educ_female:high_gni	-1.30	1.43	-0.908	0.366	-4.14	1.54

The *t*-statistic for the dummy coefficient *d* is the value of the studentized residual. This model also provides a test of whether this is equal to zero (p = 0.923).

The studentized residual for *Observation 1* is 0.970.

We can test whether or not this value differs significantly from zero by using the *t*-distribution with *df* equal to the original model's residual *df.* In our data, the residual *df* for `lm.1` is 93.




```
2 * (1 - pt(0.0970, df = 93))
```

```
[1] 0.9229351
```

Note that this is the same as testing whether the regression coefficient associated with our dummy variable is equal to zero. The *p*-value from the regression output is the same as the one we just computed.

Adjusting for multiple tests

We would carry out this same test for each of the observations. Since we have 97 observations, this is **97 hypothesis tests**. So we don't artificially inflate our type I error rate, we typically carry out a Bonferroni adjustment on each *p*-value associated with this test.

```
(0.9229351) * 97

[1] 89.5247
 # Anything over 1 is set to 1 (it is a p-value)
```

We typically carry out this test for the observation with the largest absolute studentized residual. Rather than fitting the model with the dummy variable, we can use the `outlierTest()` function from the **car** package to obtain the studentized residual, unadjusted, and Bonferroni adjusted *p*-value for the observation with the largest absolute residual.

```
# Test of the largest absolute studentized residual
outlierTest(lm.1)

No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
84 -3.20261           0.001871      0.18149
```

Influence

Measuring Influence: DFBETAs

Remember that influence on the regression coefficients is a combination of leverage and discrepancy. There are several ways to measure influence.

The most direct way to measure the influence of an observation is to drop that observation from the dataset and measure the difference in regression coefficients. We can define this difference as d_{ij} (the difference between the j^{th} regression coefficient when the i^{th} observation is dropped) as:

$$d_{ij} = B_j - B_{j(-i)}$$

These are referred to as *DFBETAs* following the precedent set by Belsley et al. (1980).

The `dfbeta()` function computes the differences DFBETA values for each coefficient for each observation. The output from `dfbeta()` is shown below for Observation 1.

```
dfbeta(lm.1)

     (Intercept)   educ_female     high_gni educ_female:high_gni
1  -6.003595e-16  1.117983e-16  0.305434363        -0.0299422430
```

The DFBETA values are sometimes scaled, either by dividing the DFBETA value (1) by the deleted estimate of the SE for the B_j coefficient, or (2) by a scaled RMSE estimate. The `dfbetas()` function computes the scaled DFBETAs using (2) above. The output is again only shown for Observation 1.

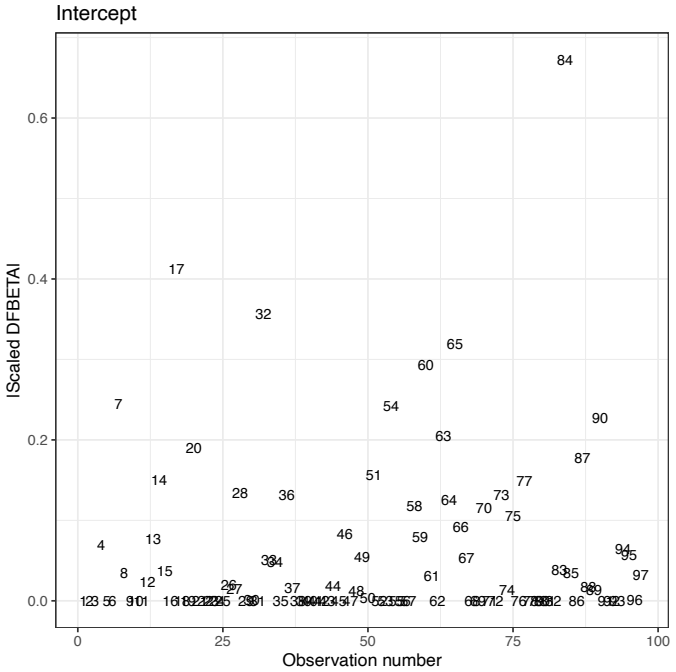
```
dfbetas(lm.1)

        (Intercept)          educ_female            high_gni educ_female:high_gni
      -1.470595e-16         1.393546e-16         2.668321e-02        -2.141606e-02
```

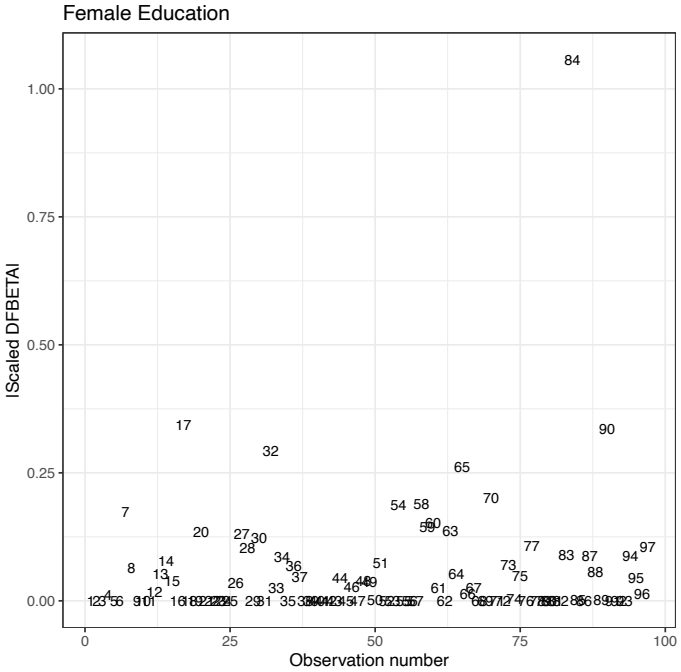
$$d_{ij}^{*} = \frac{d_{ij}}{\text{RMSE}_{(-i)} \times \sqrt{c_{kk}}}$$

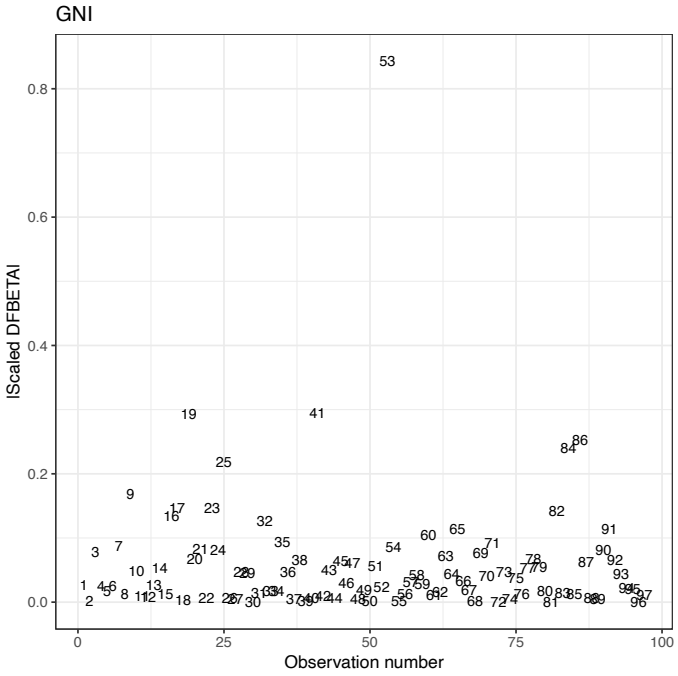
where c_{kk} is the k^{th} diagonal element of the unscaled covariance matrix $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$

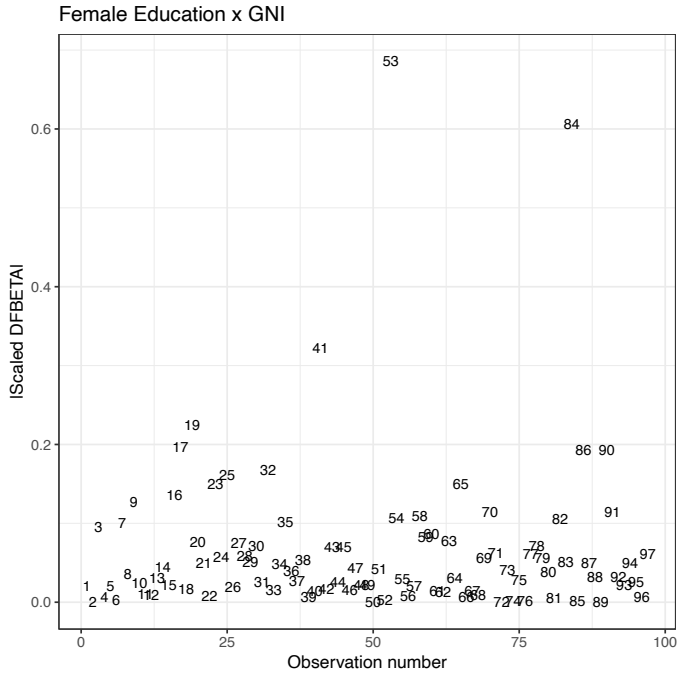
The four index plots show the standardized DFBETA values plotted versus row number. We graphically look for values that are substantially higher than most observations.

Observations 53 (Maldives) and 84 (Tajikistan) seem to have influence on the coefficients of interest (effects involving female education level).









Measuring Influence: Cook's Distance

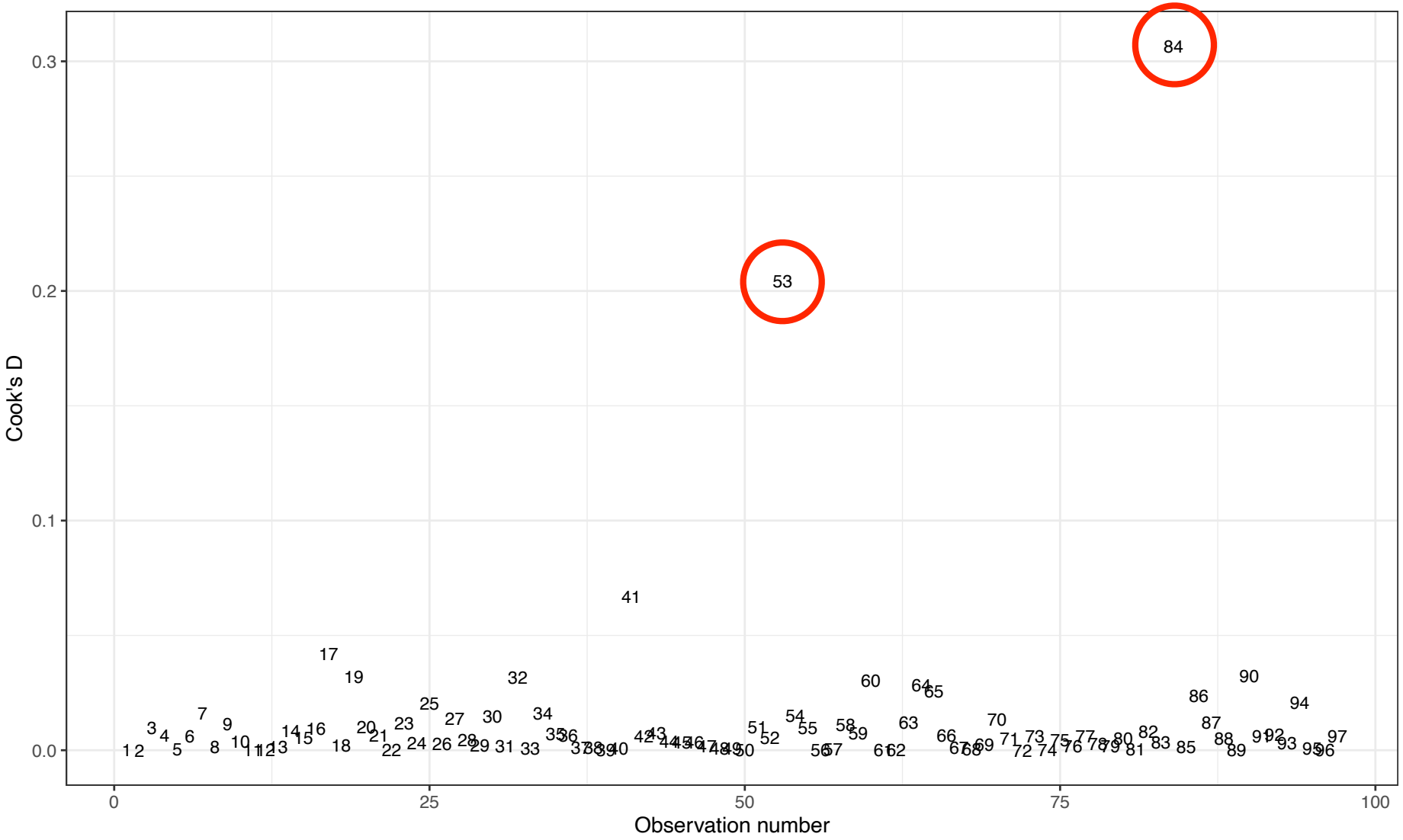
Cook (1977) proposed measuring the "distance" between B_j and $B_{j(-i)}$ by computing a measure analogous to the *F*-statistic that is independent of the scales of the *x*-variables. This measure is based on the standardized residual, number of predictors in the model (model complexity), and leverage value.

$$D_i = \frac{e_i'^2}{k+1} \times \frac{h_{ii}}{1-h_{ii}}$$

The `augment()` function outputs a column called `.cooksd` that gives Cook's distance measure for each observation. Below we show this output for the first observation in the dataset.

	contraceptive	educ_female	high_gni	.fitted	.resid	.std.resid	.hat	.sigma	.cooksd
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	57	5.9	1	55.7	-1.34	0.0976	0.0875	14.5	0.000228

The index plot below plots Cook's distance vs, row number. We graphically look for values that are substantially higher than most observations. Again, observations 53 and 84 have Cook's D values which are markedly higher than the Cook's D values for the other observations.



Measuring Influence: DFFITS

Belsley et al. (1980) proposed a similar measure to Cook's D referred to as DFFITS

$$\text{DFFITS}_i = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

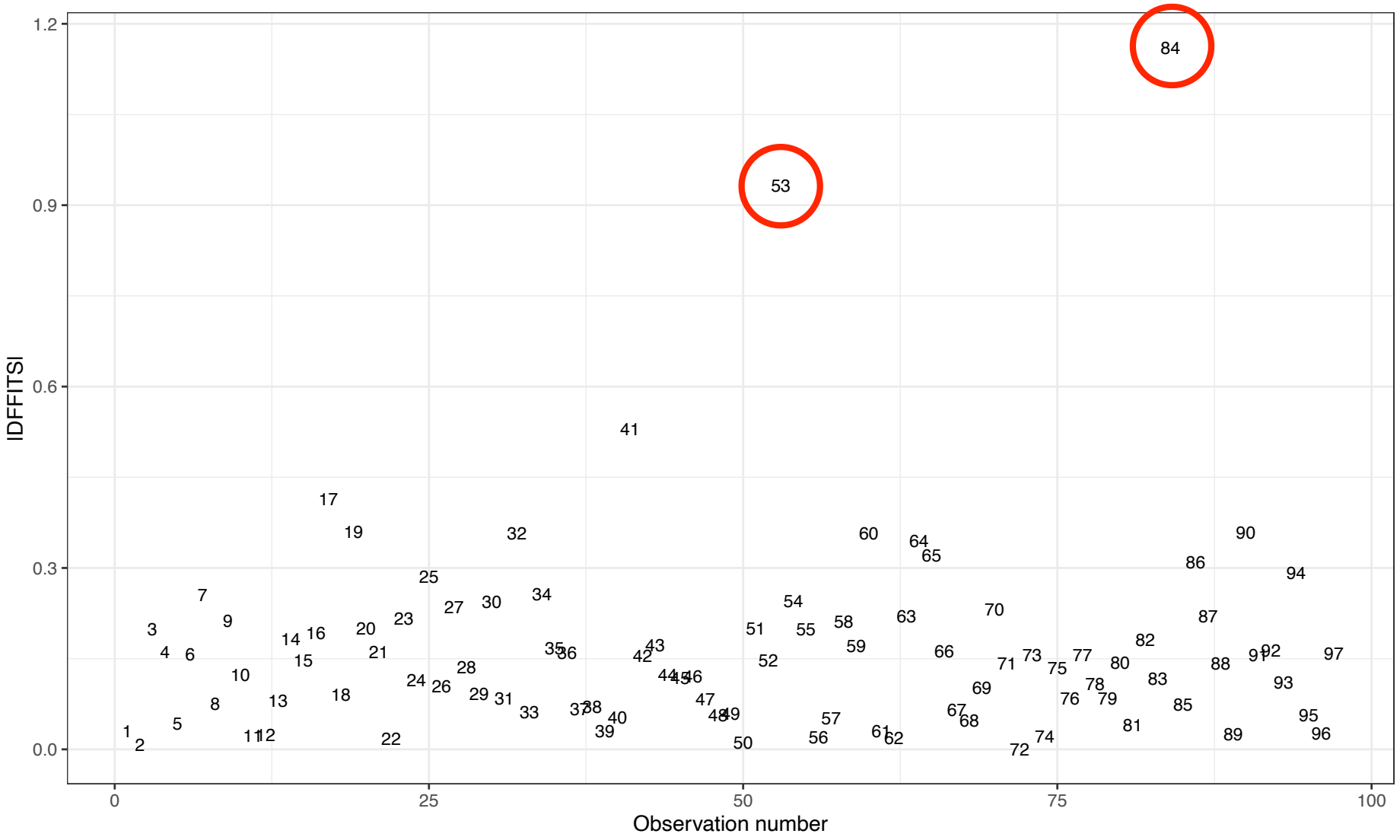
where t_i is the studentized residual. In general, apart from unusual data, the DFFITS values are quite related to the Cook's D values.

$$D_i \approx \frac{\text{DFFITS}_i^2}{k + 1}$$

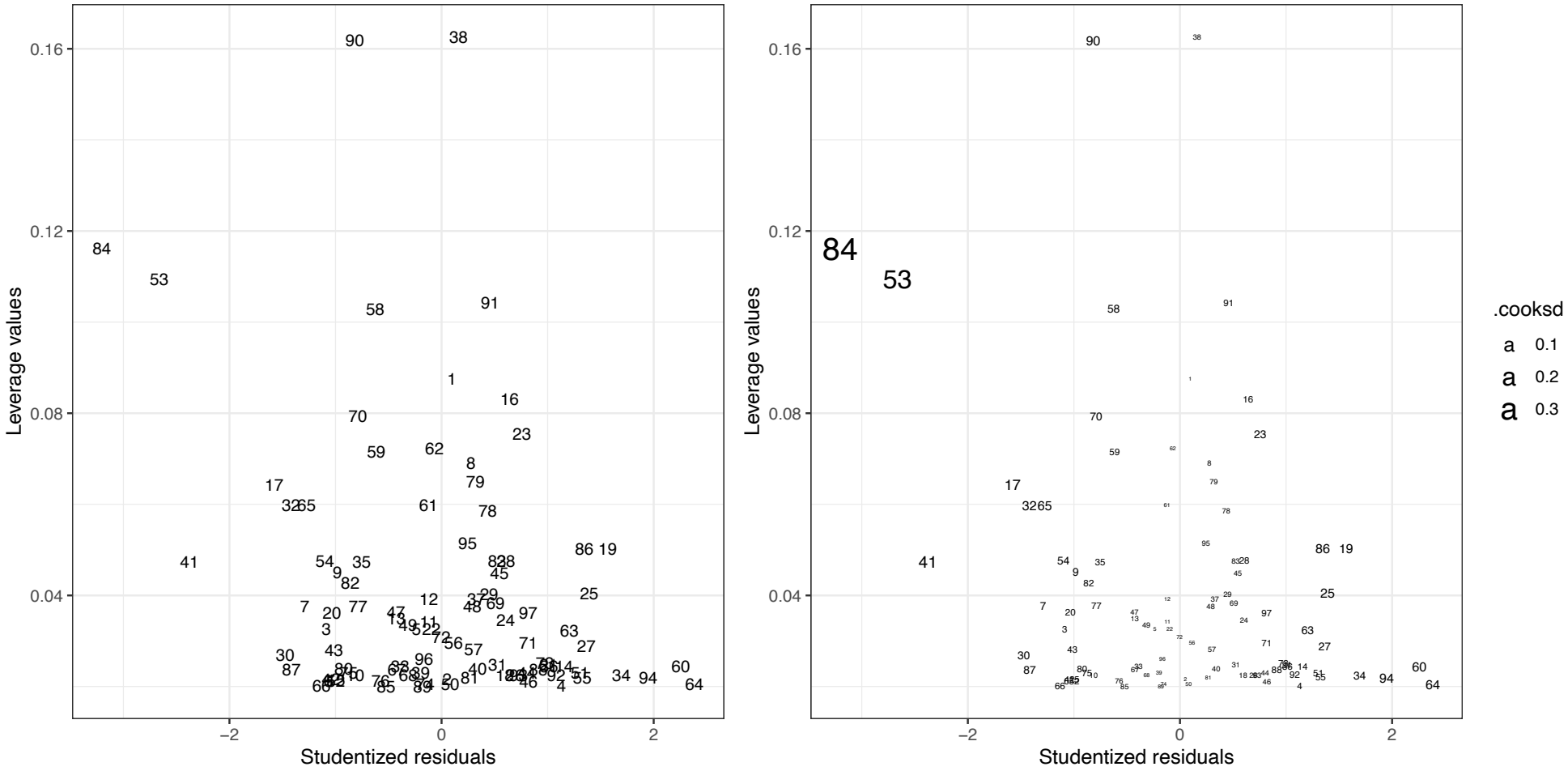
```
dffits(lm.1)

          1
0.03005657
```

Observation 1 has a DFFITS value of 0.03.

The index plot below plots the absolute value of the DFFITS values vs, row number. We graphically look for values that are substantially higher than most observations. Here again, observations 53 and 84 have DFFITS values which are markedly more extreme than the remaining DFFITS values for the other observations.



All of the deletion statistics we have looked at (DFBETA, DFFITS) rely on the studentized residual and the leverage values. An alternative diagnostic is to examine a graphic that plots the **leverage values versus the studentized residuals** (LEFT) and look for large observations. We can also size the observations in this plot based on their Cook's D values (RIGHT).



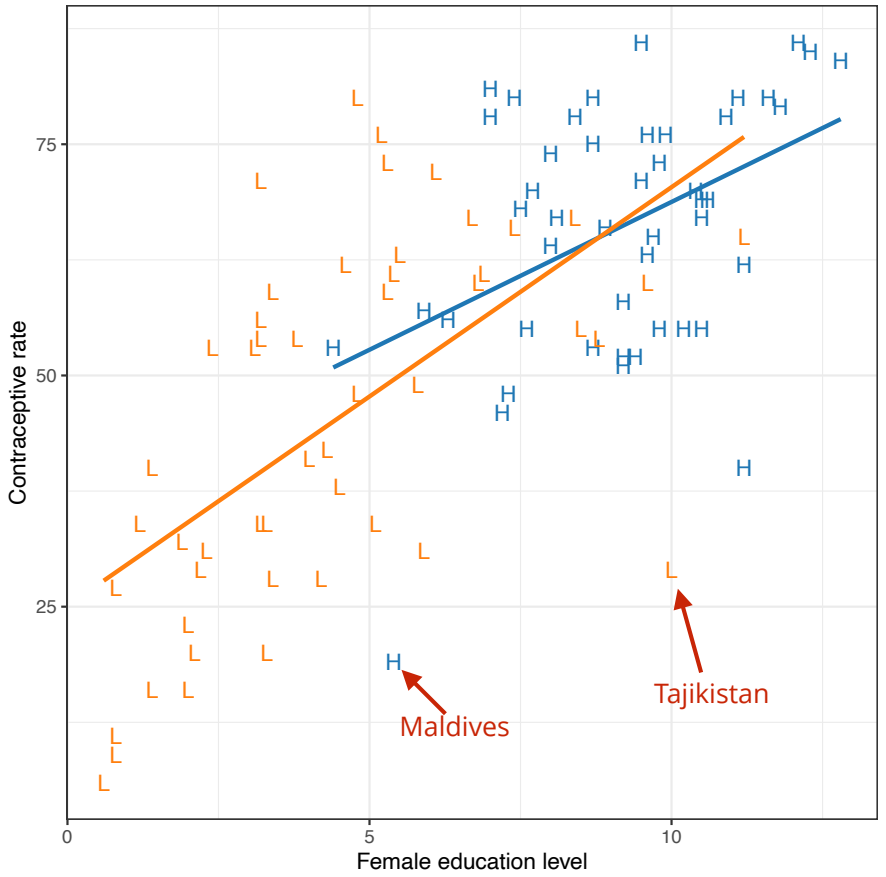
All of the numerical and graphical evidence point to observations 53 (Maldives) and 84 (Tajikistan) as being problematic and influential.

```
dfbeta(lm.1)[c(53, 84), ]

    (Intercept) educ_female  high_gni educ_female:high_gni
53 1.775101e-15   0.0000000 -9.299315            0.9242394
84 2.603053e+00  -0.8039487 -2.603053            0.8039487
```

Based on the DFBETA values, removing these observations would change the magnitude of the unstandardized coefficients in the following way:

- Intercept: –2.60
- Main effect of female education level: +0.80
- Main effect of GNI: +11.90
- Interaction effect: –1.72

$$d_{ij} = B_j - B_{j(-i)}$$



Measuring Influence on SEs: COVRATIO

Most of the influence measures we have examined look at the effect on the coefficients. If it is of importance, you may also want to look at the effect on the SEs for the coefficients.

One way to measure an observation's impact on the SEs is to measure the change in the **joint confidence region** for the coefficients (analogous to the length of the coefficient CIs, which is proportional to the SE). Belsley et al. (1980) proposed an influence measure to do just this:

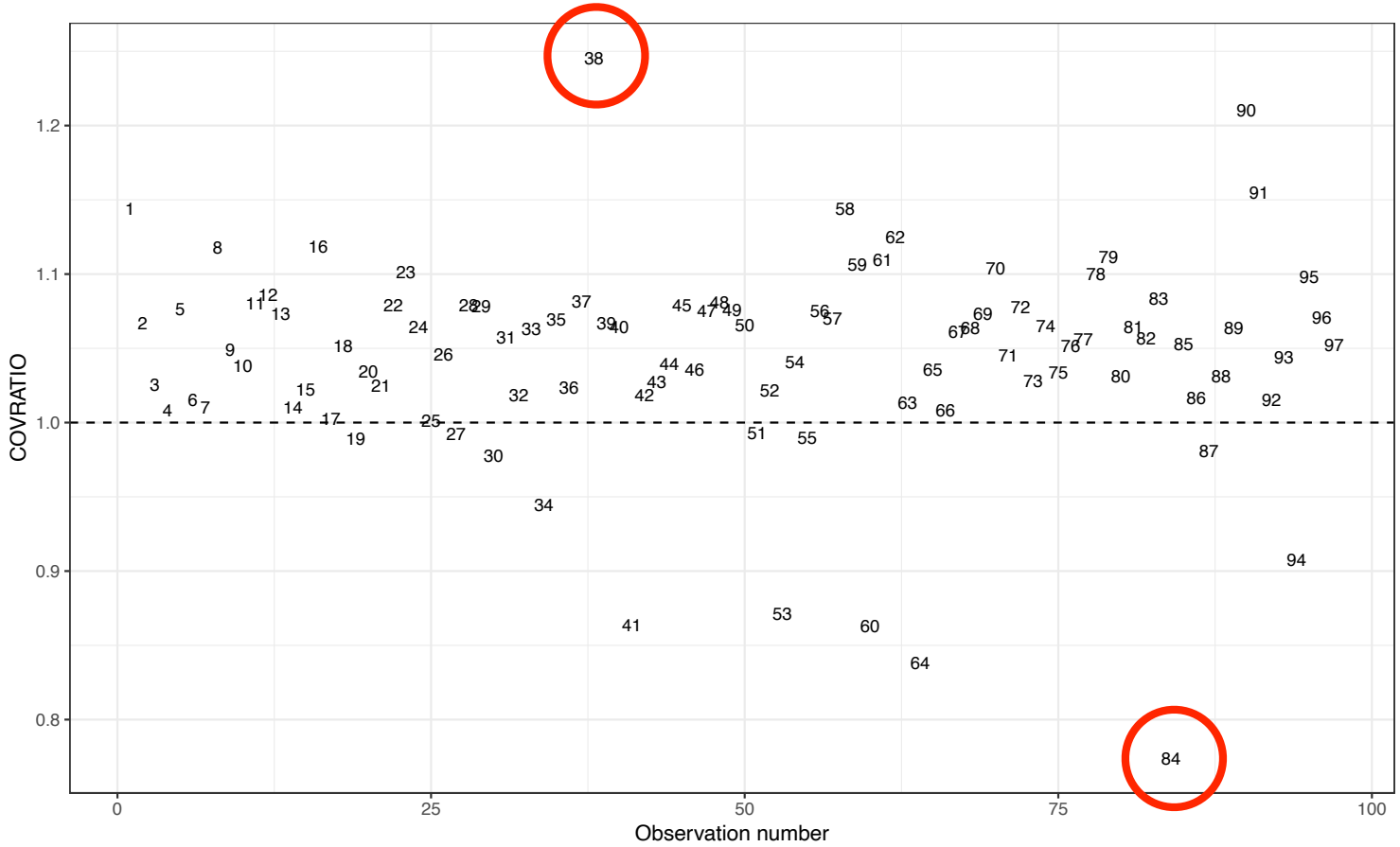
$$\text{COVRATIO}_i = \frac{1}{(1-h_{ii})\left(\frac{n-k-2+t_i^2}{n-k-1}\right)^{k+1}}$$

where h_{ii} is the leverage value, n is the sample size, k is the number of predictors, and t_i is the studentized residual. COVRATIO is a measure of the ratio for the squared volume of the confidence region between the full data and the observation deleted data. Thus, COVRATIO values that differ from 1 are suspicious.

We can use the `covratio()` function to compute each observation's COVRATIO value. Here we show this value for Observation 1.

```
covratio(lm.1)

       1
1.143885
```

We also examine an index plot of the COVRATIO values.



Numerical Cutoffs for Determining Influence

Fox points out that one needs to examine the distribution of influence values and look for unusual values rather than adhere to numerical cutoffs for saying an observation is influential. He also goes on to say that cutoff values are most effective when they are added to graphical presentations to call attention to suspicious observations.

- Hat-values (leverage): $h_{ii} > 2\bar{h}$
- Studentized residuals: $|t_i| > 2$
- Standardized DFBETA: $|\mathrm{DFBETA}| > \frac{2}{\sqrt{n}}$
- Cook's D: $D_i > \frac{4}{n-k-1}$
- DFFITS: $|\mathrm{DFFITS}| > 2\sqrt{\frac{(k+1)}{n-k-1}}$
- COVRATIO: $|\mathrm{COVRATIO} - 1| > 3\frac{(k+1)}{n}$

Joint Influence

Added-Variable Plots

Identification and removal of individual influential observations using the methods so far is sequential:

- Identify the most influential observation
- Remove the observation
- Re-fit the model and re-calculate the influence statistics
- Repeat

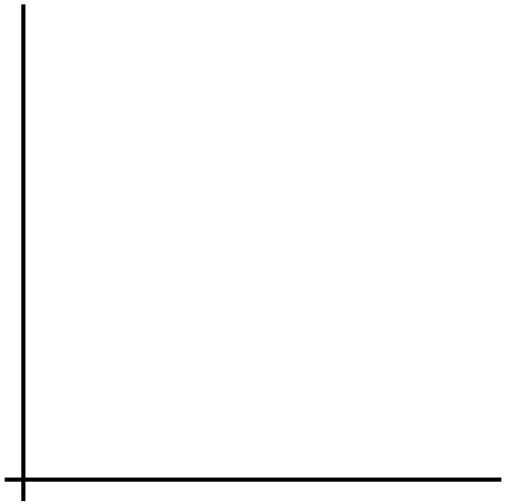
It is also difficult to use this sequential methodology to identify and remove multiple observations simultaneously that **jointly influence** the regression coefficients.

Added-Variable Plots (a.k.a., *partial-regression leverage plots*) are a graphical tool to identify subsets of points that **jointly influence** the regression coefficients for removal.

To create an added-variable plot to determine the joint-influence of observations on a particular predictor, say **X1** in a two predictor model:

(1) Regress **Y** on $\mathbf{X}_{(-X1)}$ (omit **X1** from the model) — Y | X2
(2) Regress **X1** on $\mathbf{X}_{(-X1)}$ — X1 | X2
(3) Plot the residuals from (1) versus the residuals from (2)

These residuals indicate how unusual the outcome value is given the other predictor values

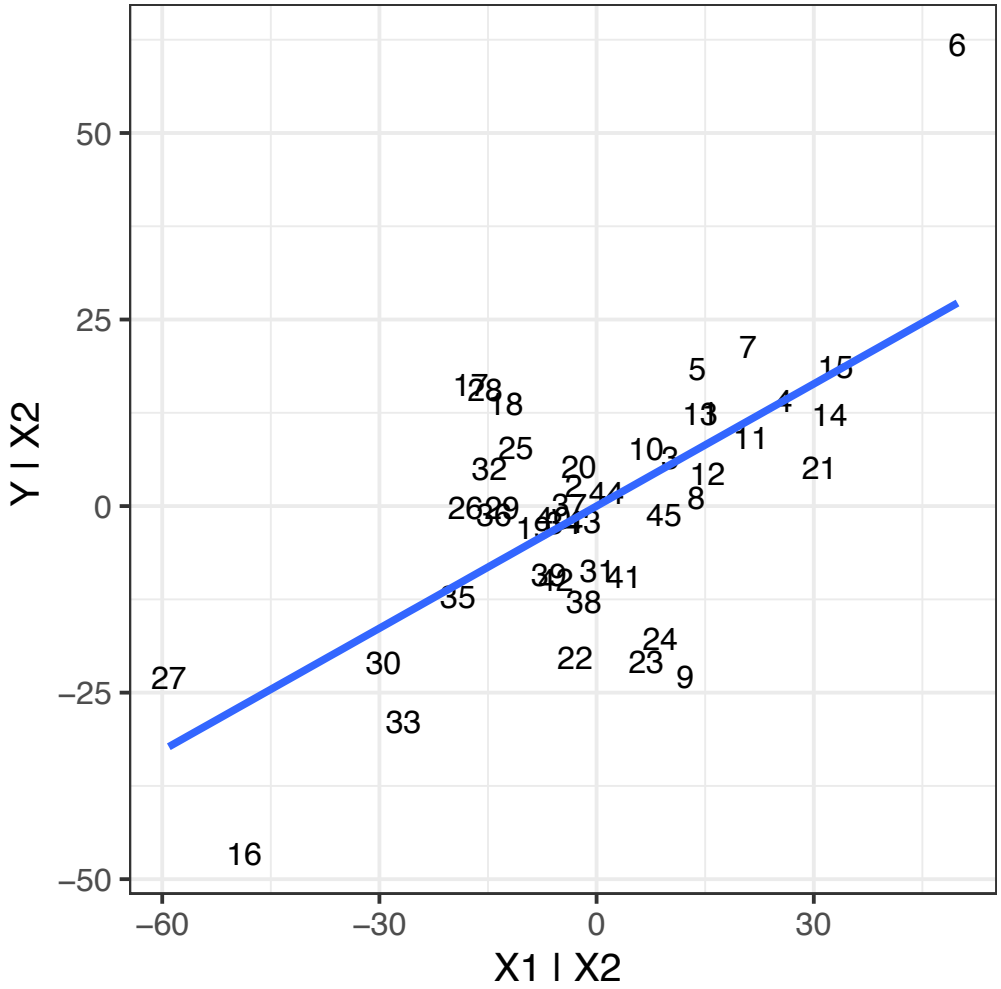


These residuals indicate how unusual the predictor value for X1 is given the other predictor values

Added-Variable Plot: Fake Two-Predictor Example

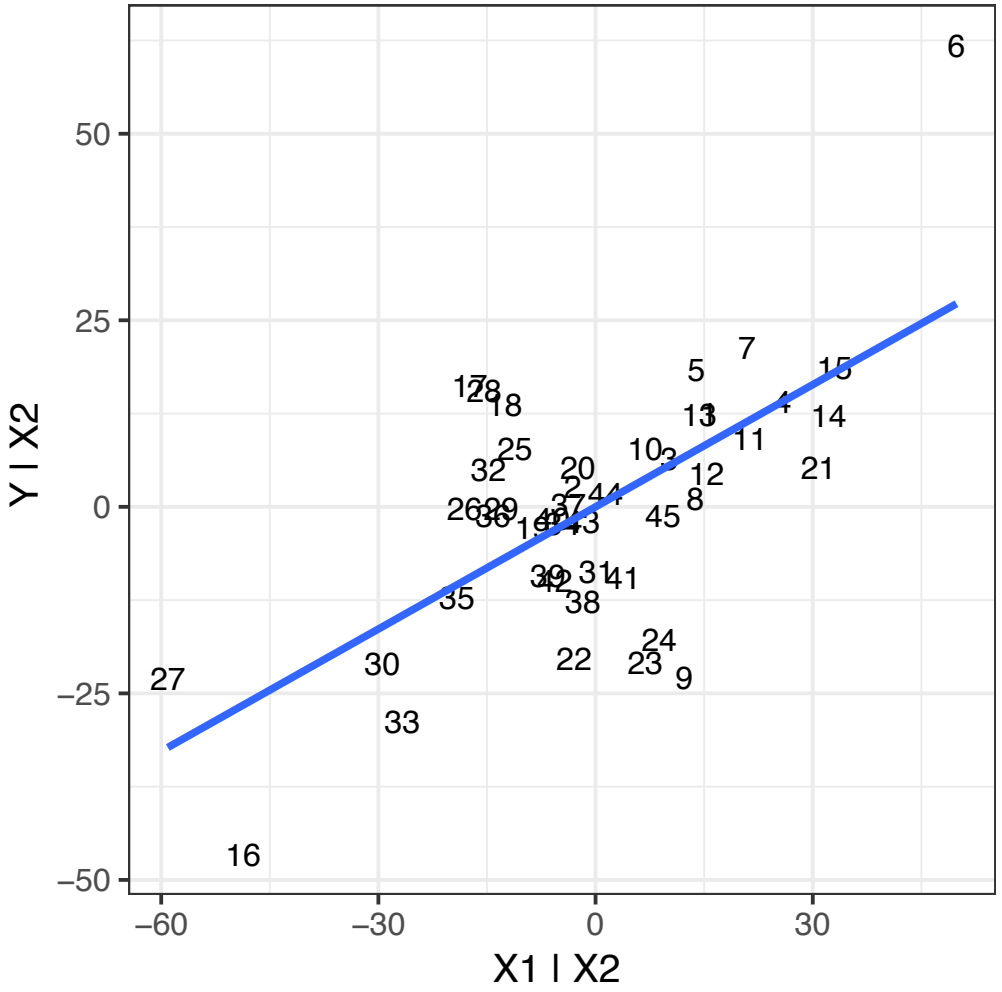
The added-variable plot shows the scatterplot of the residuals. These residuals have several interesting properties:

- The slope of regressing the residuals of `Y|X2` (**e**[1]) on those from `X1 |X2` (**e**[2]) is the slope associated with X1 from the full multiple regression (`Y | X1 + X2`)
- The SE associated with the slope estimate when regressing **e**[1] on **e**[2] is the same as the SE associated with the partial slope for X1 in the full multiple regression.

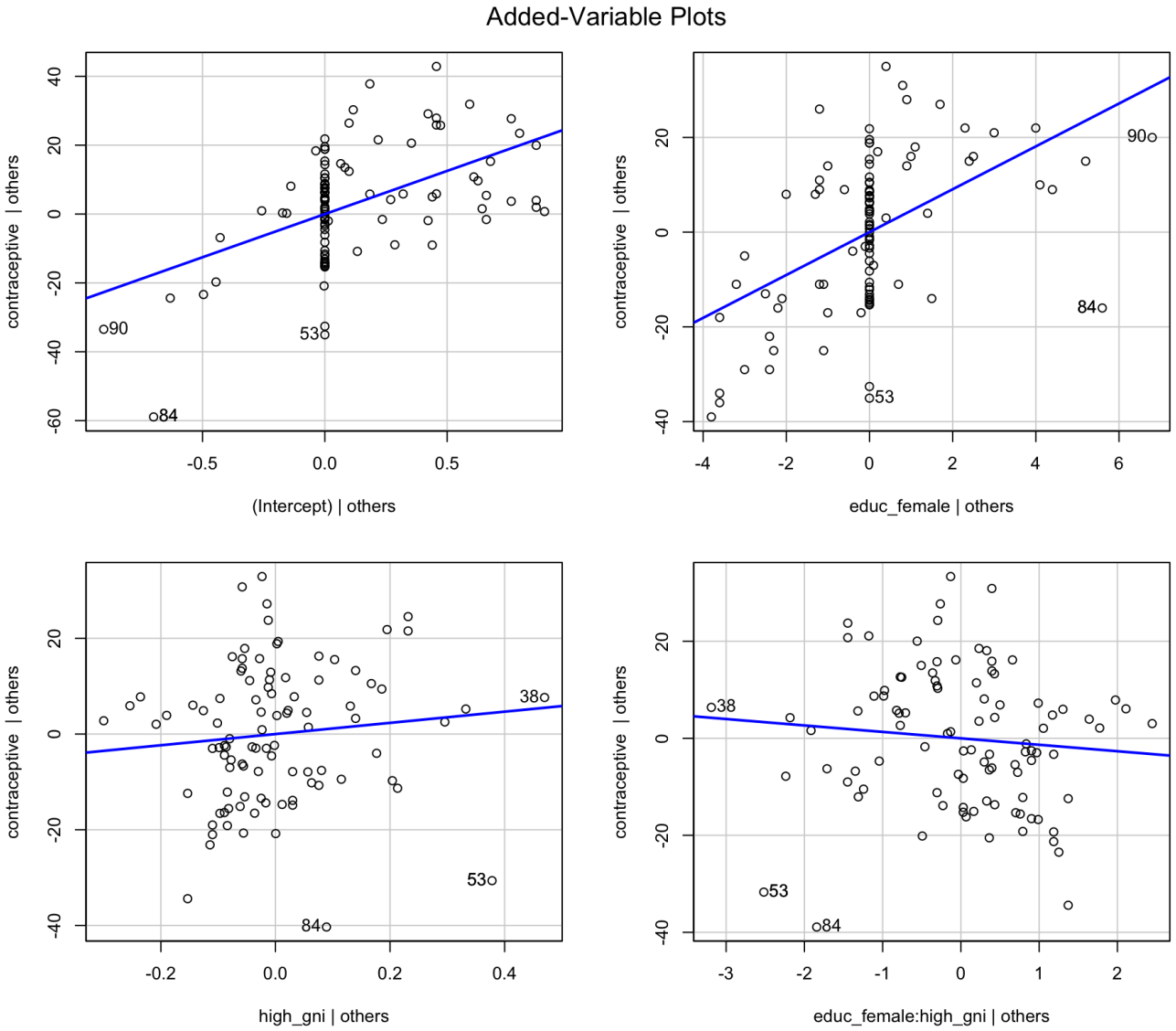


How do these properties help us interpret and use the added-variable plot to identify influential observations?

- Removing observations above the line will decrease the magnitude of B1, and vice-versa
- The degree of influence depends how far above or below the line the observation is *and* also how far to the left/right (leverage) the observation is from the center mass.
  - Removing *Observation 6* will have a larger effect than removing *Observation 7*; it is more influential.
  - Observations 27 and 16 have high leverage (along with observation 6) but they are not as influential.
- By evaluating how the regression line in the added-variable plot will change as we remove observations, we directly predict how their removal influences B1 in the full multiple regression



Here are the added-variable plots for the interaction model to predict contraception rate. We compute these using the `avPlots()` function from the **car** library.



Based on these plots, Observations 53 and 84 are clearly influential on each of the coefficients. Observation 53 is perhaps less problematic as it is not a high leverage point in those plots.

Observations 38 (Iraq) and 90 (Ukraine) are also identified as suspicious (high leverage). However given these plots they do not appear to be as influential (they are close to the line).

Should we remove unusual observations?

- WHY are the observations are unusual.
- Do the observations help re-specify the model (including omitted predictors)?
- Will a data transformation help?
- Should you use a more robust method of fitting a model (e.g., nonparametric regression)?

Be careful not to over-fit to a small number of unusual cases. Large samples help immeasurably. In large samples, unusual cases rarely have much influence on the estimated coefficients or SEs.

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*, 17–22.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics, 19,* 15–18.

Davis, C. (1990). Body image and weight preoccupation: A comparison between exercising and non-exercising women. *Appetite, 15,* 13–21.

Ericksen, E. P., Kadane, J. B., & Tukey, J. W. (1989). Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association, 84,* 927–944.

Fox, J. (1991). *Regression diagnostics.* Thousand Oaks, CA: Sage.

Fox, J. (2016). *Applied regression analysis & generalized linear models* (3rd ed.). Thousand Oaks, CA: Sage.