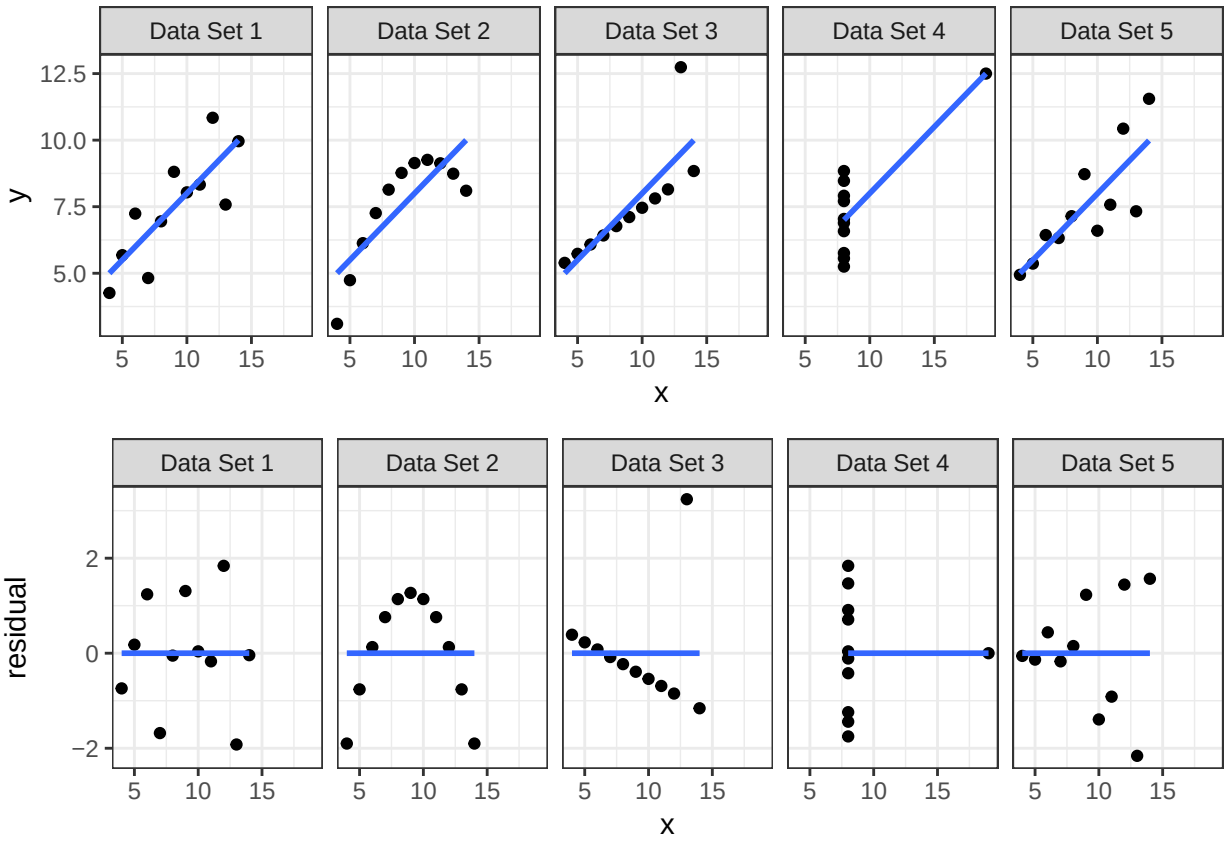
Chapter 11: Outliers and Influential Observations

Recall Anscombe's Data



- For today, let's focus on Data Sets 3 and 4
- Definitions (note: there are not universally agreed on definitions for these terms):
  - An **outlier** is an observation that "doesn't fit" with the patterns in the rest of the data
    - Both data set 3 and data set 4 have outliers
  - An **influential observation** is an observation whose removal from the data set would substantially change the model fit (coefficient estimates)
    - Both data set 3 and data set 4 have influential observations
    - The point in data set 4 is *more influential*
  - A **high leverage observation** is one whose explanatory variable values are far from the explanatory variable values of other observations
    - Data set 4 has a high leverage observation
    - Data set 3 does not
- Note that residual plots **exactly fail** to identify very influential/high leverage observations!!!

Leverage

- If the model has 1 X variable, the leverage of observation i is defined to be

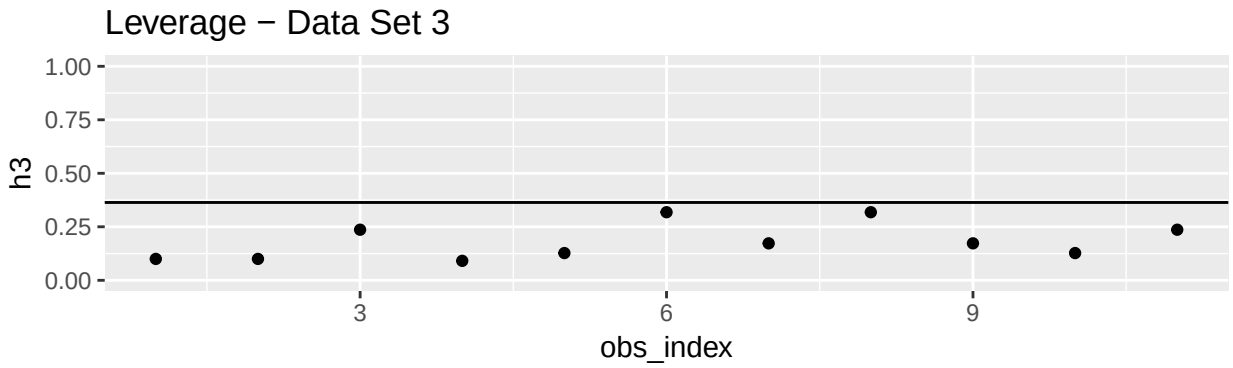
$h_i = \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2} + \frac{1}{n}$

- Basically, how far is X_i from $\bar{X}$, standardized in an obsure way.
  - $\frac{1}{n} \leq h_i \leq 1$
  - Average of leverages is p/n (where p is the number of parameters for the mean in the model).
- As a very rough guide, $h_i > 2p/n$ indicates an observation is worth looking into more

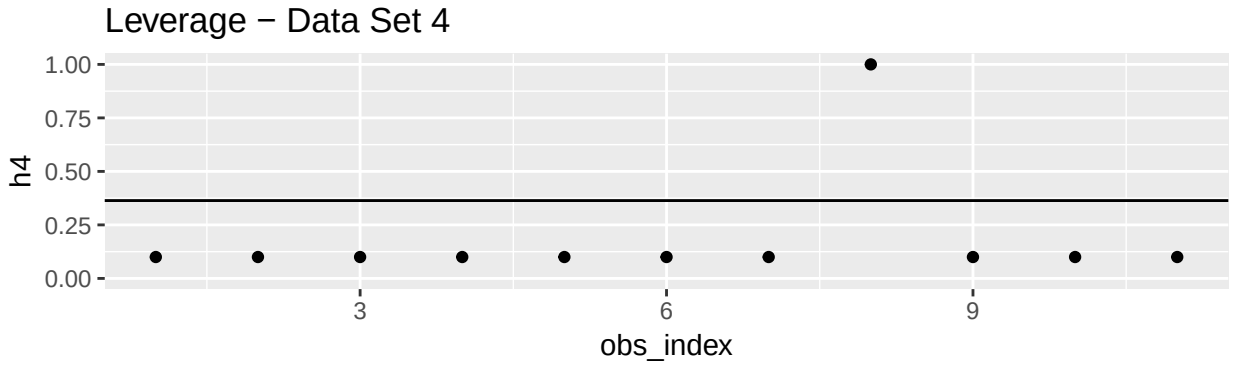
Plots of leverage vs. observation index (code will be shown later)

```
# 2p/n; p = 2 since we have beta_0 and beta_1 in our simple linear regression model
2 * 2 / nrow(anscombe)
```

```
## [1] 0.3636364
```



Looks OK




```
# confirm observation 8 is the one with a big X!
anscombe$x4[8]
```

```
## [1] 19
```

Studentized Residuals

- Observations with high leverage tend to have small residuals!
  - $SD(res_i) = \sigma\sqrt{1-h_i}$
- Looking at just the residuals can be misleading
- The **studentized residuals** adjust by dividing residual by its estimated standard deviation

$studres_i = \frac{res_i}{\hat{\sigma}\sqrt{1-h_i}}$

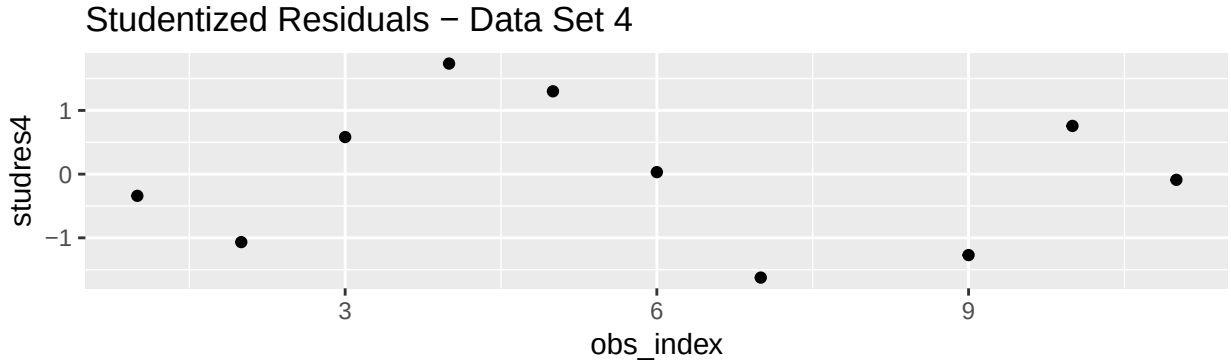
- A studentized residual less than -2 or greater than 2 could indicate problems **if other diagnostics also indicate issues**
- We expect about 5% of *studentized residuals* to be less than -2 or greater than 2.

Plots of studentized residuals vs. observation index (code will be shown later)


```
# confirm observation 3 is the one with a big Y!
anscombe$y3[3]
```

```
## [1] 12.74
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Looks OK… what's up with that warning?

```
anscombe$studres4
```

```
##           1           2           3           4           5           6
## -0.34104165 -1.06669299  0.58216636  1.73514504  1.30031318  0.03136768
##           7           8           9          10          11
## -1.62381807         NaN -1.27046922  0.75677904 -0.08931624
```

```
anscombe$h4
```

```
##   1   2   3   4   5   6   7   8   9  10  11
## 0.1 0.1 0.1 0.1 0.1 0.1 0.1 1.0 0.1 0.1 0.1
```

Cook's Distance

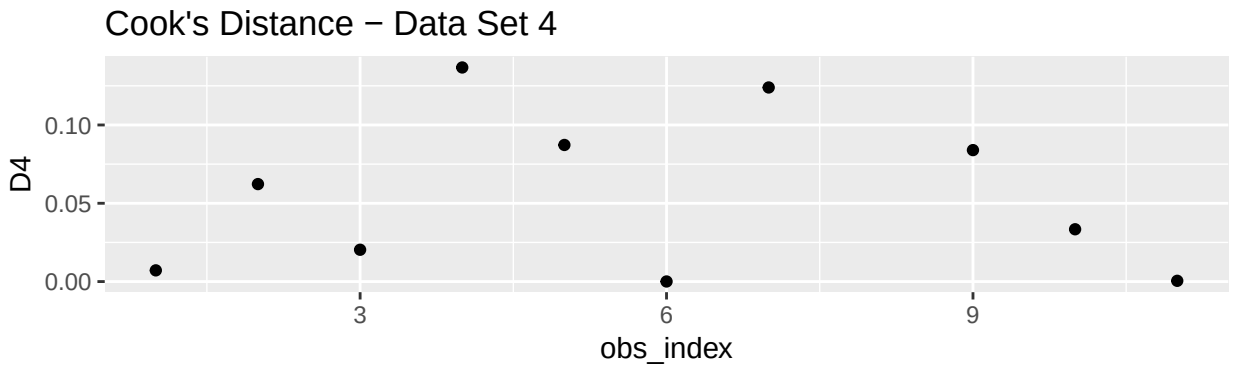
- Measures how different predicted values for all observations are when observation i is or is not used for model estimation
- Fit model using all observations; get predicted values $\hat{Y}_j$ for each $j = 1, \ldots, n$
- Fit model using all observations **other than** i; get predicted values $\hat{Y}_{j(i)}$ for each $j = 1, \ldots, n$
- $D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_{j(i)} - \hat{Y}_j)^2}{p\hat{\sigma}^2}$
- As a very rough guide, $D_i > 1$ indicates an observation is worth looking into more

Plots of Cook's distance vs. observation index (code will be shown later)


```
# confirm observation 3 is the one with a big Y!
anscombe$y3[3]
```

```
## [1] 12.74
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Looks OK… what's up with that warning?

```
anscombe$D4
```

```
##            1            2            3            4            5
## 7.165166e-03 6.225950e-02 2.032144e-02 1.367179e-01 8.723799e-02
##            6            7            8            9           10
## 6.148813e-05 1.239465e-01          NaN 8.394407e-02 3.340334e-02
##           11
## 4.980902e-04
```

```
anscombe$h4
```

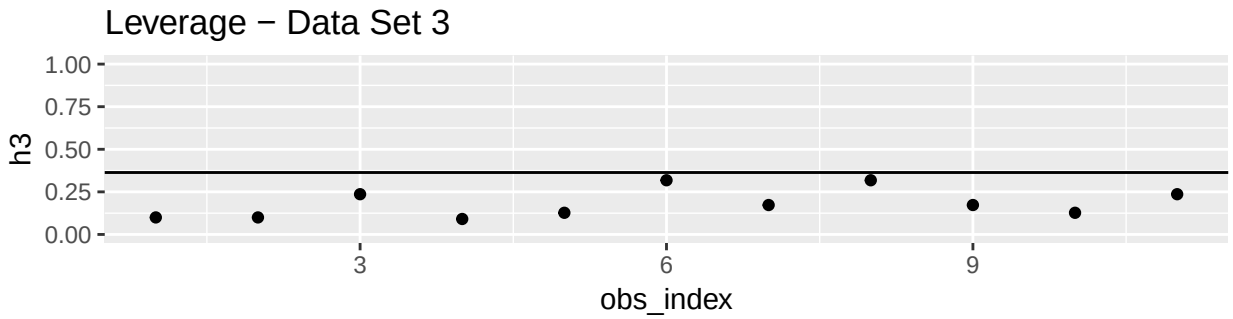
```
##   1   2   3   4   5   6   7   8   9  10  11
## 0.1 0.1 0.1 0.1 0.1 0.1 0.1 1.0 0.1 0.1 0.1
```

R Code: Manual Plots

- Every statistical software package will give you different plots by default
- Our book suggests the plots we've looked at so far, which are not the defaults for R/require more code to create:

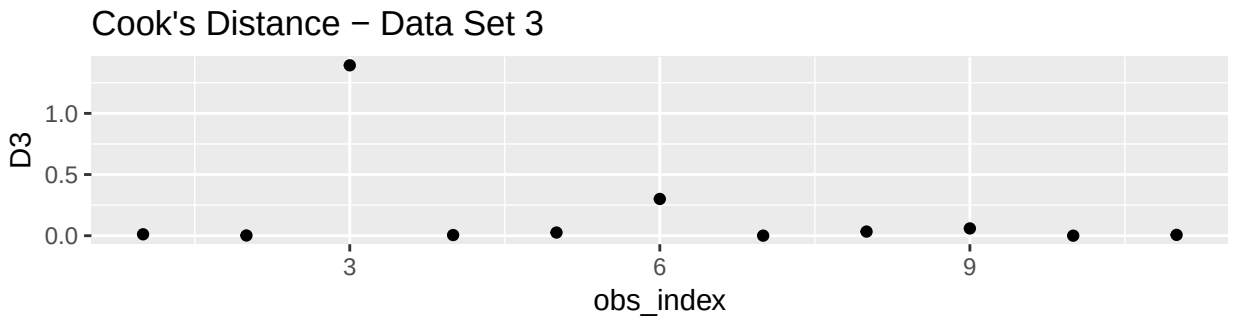
```
anscombe <- anscombe %>%
  mutate(
    obs_index = row_number(),
    h3 = hatvalues(fit3),
    studres3 = rstudent(fit3),
    D3 = cooks.distance(fit3)
  )

ggplot(data = anscombe, mapping = aes(x = obs_index, y = h3)) +
  geom_point() +
  geom_hline(yintercept = 2 * 2 / nrow(anscombe)) +
  ylim(0, 1) +
  ggtitle("Leverage - Data Set 3")
```




```
ggplot(data = anscombe, mapping = aes(x = obs_index, y = studres3)) +
  geom_point() +
  ggtitle("Studentized Residuals - Data Set 3")
```

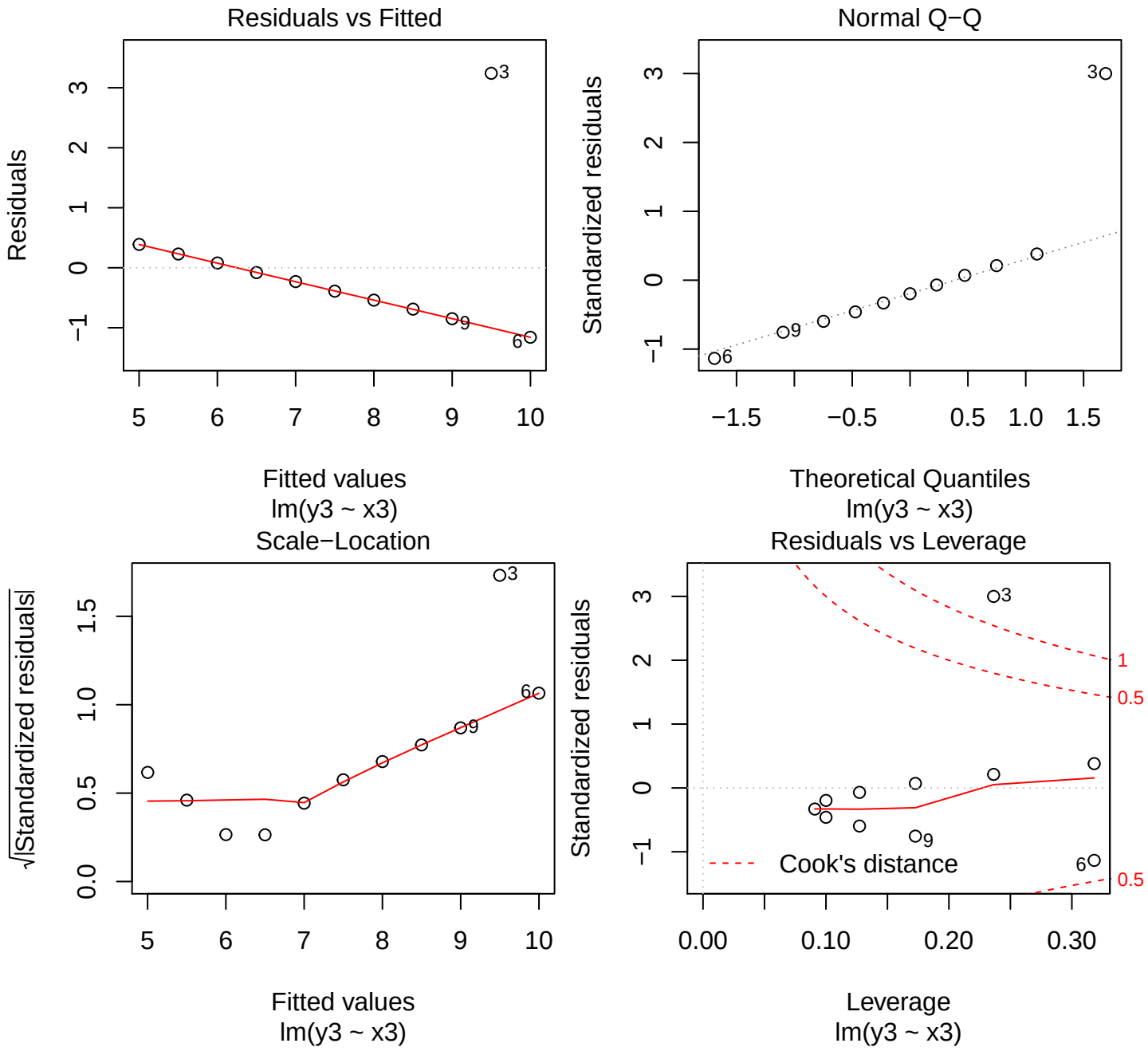

```
ggplot(data = anscombe, mapping = aes(x = obs_index, y = D3)) +
  geom_point() +
  ggtitle("Cook's Distance - Data Set 3")
```



R Code: Default Plots

You can get a set of different diagnostic plots more easily, but I find the plot involving Cook's distance and Leverage less intuitive:

```
plot(fit3)
```



Note: to get the plots to all show up in the knitted pdf, I had to set figure height and width in the code chunk declaration:

```
```{r, fig.height = 4, fig.width = 4}
```
```