Set R02 - Regression diagnostics

STAT 401 (Engineering) - Iowa State University

March 20, 2017

All models are wrong!

George Box (Empirical Model-Building and Response Surfaces, 1987):

> *All models are wrong, but some are useful.*

http:
//stats.stackexchange.com/questions/57407/what-is-the-meaning-of-all-models-are-wrong-but-some-are-useful

> *"All models are wrong" that is, every model is wrong because it is a simplification of reality. Some models, especially in the "hard" sciences, are only a little wrong. They ignore things like friction or the gravitational effect of tiny bodies. Other models are a lot wrong - they ignore bigger things.*
>
> *"But some are useful" - simplifications of reality can be quite useful. They can help us explain, predict and understand the universe and all its various components.*
>
> *This isn't just true in statistics! Maps are a type of model; they are wrong. But good maps are very useful.*

Regression

The simpler linear regression model is

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

this can be rewritten as

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad e_i \overset{iid}{\sim} N(0, \sigma^2)$$

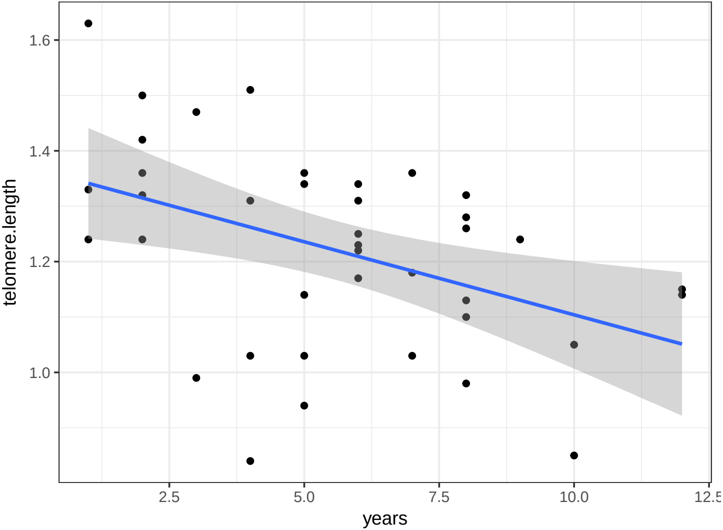
where we estimate the errors via the residuals

$$r_i = \hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

Key assumptions are:

- The errors are
  - normally distributed,
  - have constant variance, and
  - are independent of each other.
- There is a linear relationship between the expected response and the explanatory variables.

Telomere data

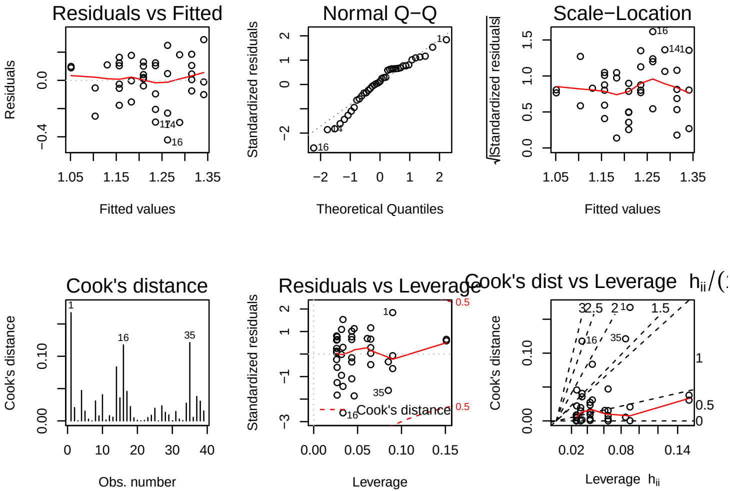


Case statistics

To evaluate these assumptions, we will calculate a variety of case statistics:

- Fitted values
- Leverage
- Residuals
  - Standardized residuals
  - Studentized residuals
- Cook's distance

Default diagnostic plots in R



Residuals and Fitted values

Another rewrite of a regression model is

$$\begin{aligned} Y_i &\stackrel{ind}{\sim} N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 X_i \end{aligned}$$

A fitted value $\hat{Y}_i$ for an observation i is

$$\hat{Y}_i = \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

and thus the residual is

$$\begin{aligned} r_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\mu}_i \\ &= Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \\ &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \end{aligned}$$

Leverage

Definition

The leverage $(0 \leq h_i \leq 1)$ of an observation i is a measure of how far away the observations explanatory variable value is away from the other observations. Larger leverage indicates a larger *potential* influence of a single observation on the regression model.

In simple linear regression,

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{SXX}$$

which is involved in the standard error for the line for a location x_i.

The variability in the residuals is a function of the leverage, i.e.

$$Var(r_i) = \sigma^2(1 - h_i)$$

Telomere data

```
m <- lm(telomere.length~years, Telomeres)

cbind(Telomeres, leverage = hatvalues(m)) %>%
  select(years, leverage) %>%
  unique() %>%
  arrange(-years)
```

```
   years   leverage
1     12 0.15113547
2     10 0.08504307
3      9 0.06115897
4      8 0.04338293
5      7 0.03171496
6      6 0.02615505
7      5 0.02670321
8      4 0.03335944
9      3 0.04612373
10     2 0.06499608
11     1 0.08997651
12     1 0.08997651
```

Standardized residuals

Typically, we don't visualize residuals themselves, but instead we standardize them, i.e.

$$\frac{r_i}{\sqrt{\widehat{Var}(r_i)}} = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}$$

If $|r_i|$ is large, it will have a large impact on $\hat{\sigma}^2 = \sum_{i=1}^n r_i^2/(n-2)$. Thus, we can calculate a externally studentized residual

$$\frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

where $\hat{\sigma}_{(i)} = \sum_{j \neq i} r_j^2/(n-3)$.

These residuals should generally be compared to a standard normal distribution.

Telomere data

```
m <- lm(telomere.length˜years, Telomeres)

cbind(Telomeres,
      leverage     = hatvalues(m),
      residual     = residuals(m),
      standardized = rstandard(m),
      studentized  = rstudent(m))


   years telomere.length   leverage     residual standardized studentized
1      1            1.63 0.08997651  0.288692247   1.84050794  1.90475158
2      1            1.24 0.08997651 -0.101307753  -0.64587021 -0.64070443
3      1            1.33 0.08997651 -0.011307753  -0.07209064 -0.07111476
4      2            1.50 0.06499608  0.185066562   1.16399233  1.16977226
5      2            1.42 0.06499608  0.105066562   0.66082533  0.65571510
6      2            1.36 0.06499608  0.045066562   0.28345009  0.27989750
7      2            1.32 0.06499608  0.005066562   0.03186659  0.03143344
8      3            1.47 0.04612373  0.181440877   1.12984272  1.13420749
9      2            1.24 0.06499608 -0.074933438  -0.47130041 -0.46628962
10     4            1.51 0.03335944  0.247815192   1.53293696  1.56251168
11     4            1.31 0.03335944  0.047815192   0.29577555  0.29209673
12     5            1.36 0.02670321  0.124189507   0.76558098  0.76121769
13     5            1.34 0.02670321  0.104189507   0.64228860  0.63711129
14     3            0.99 0.04612373 -0.298559123  -1.85914473 -1.92601533
15     4            1.03 0.03335944 -0.232184808  -1.43625042 -1.45793267
16     4            0.84 0.03335944 -0.422184808  -2.61155376 -2.85227987
17     5            0.94 0.02670321 -0.295810493  -1.82355895 -1.88546999
18     5            1.03 0.02670321 -0.205810493  -1.26874325 -1.27962563
19     5            1.14 0.02670321 -0.095810493  -0.59063518 -0.58536500
20     6            1.17 0.02615505 -0.039436179  -0.24304058 -0.23992534
21     6            1.23 0.02615505  0.020563821   0.12673244  0.12503525
```

Cook's distance

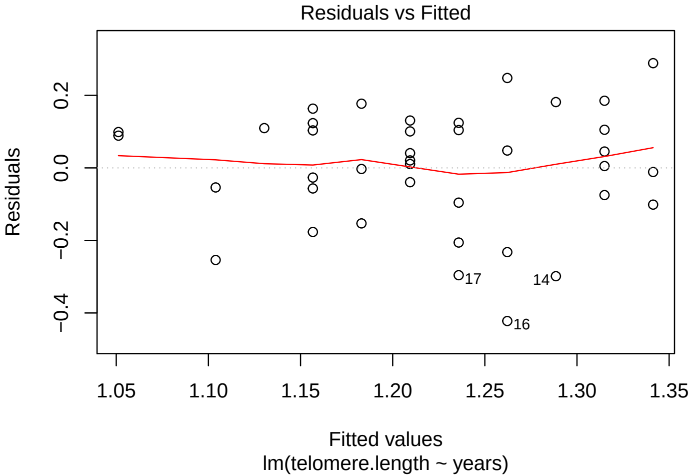
If a particular observation is highly influential in estimating the parameters of the regression model, we can assess how influential it is by fitting the regression model with and without that observation and evaluating how the model parameters change.

Definition

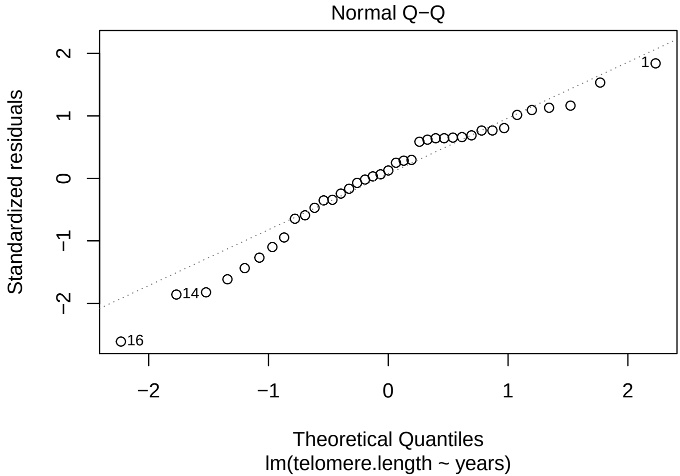
The Cook's distance for an observation i $(d_i > 0)$ is a measure of how much the estimates of the regression parameters change when that observation is included versus when it is excluded.

Operationally, we might be concerned when d_i is

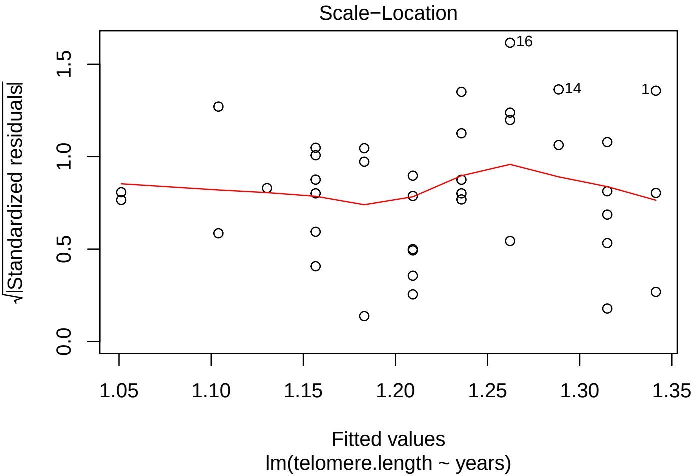
- larger than 1 or
- larger then 4/n.



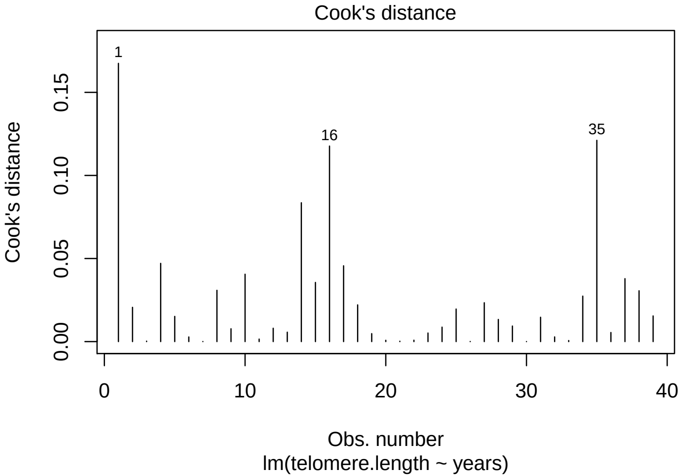
Assumption	Violation
Linearity	Quadratic curve
Constant variance	Funnel shape



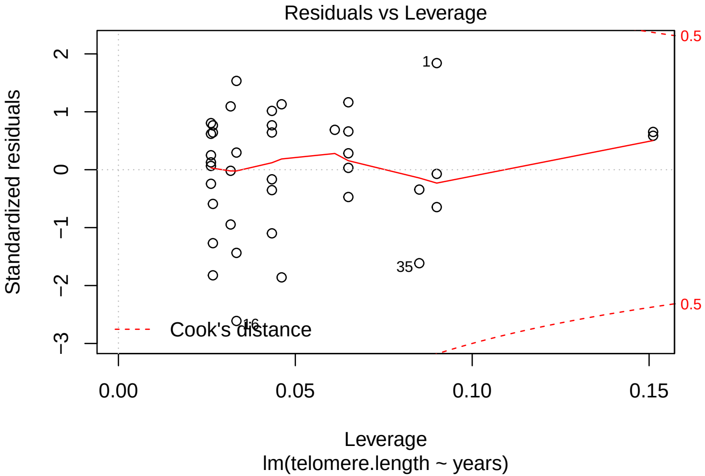
Assumption	Violation
Normality	Points don't generally fall along the line



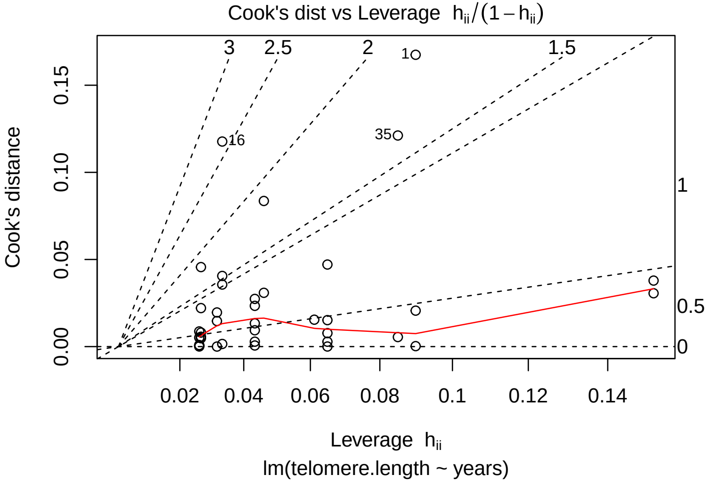
Assumption	Violation
Normality	Too many large values
Constant variance	Increasing or decreasing trend



Outlier	Violation
Influential observation	Large Cook's distance



Outlier	Violation
Influential observation	Points outside red dashed lines



Outlier	Violation
Influential observation	Points outside red dashed lines

Summary

Case statistics:

- Fitted values
- Leverage
- Residuals
  - Standardized residuals
  - Studentized residuals
- Cook's distance

Model assumptions:

- Normality
- Constant variance
- Independence
- Linearity

Default plots in R do not assess all model assumptions. So we'll need to make more.