R02 - Regression diagnostics

STAT 587 (Engineering)
Iowa State University

March 30, 2021

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.*

Simple Linear Regression

The simple linear regression model is

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

this can be rewritten as

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

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.

Multiple Regression

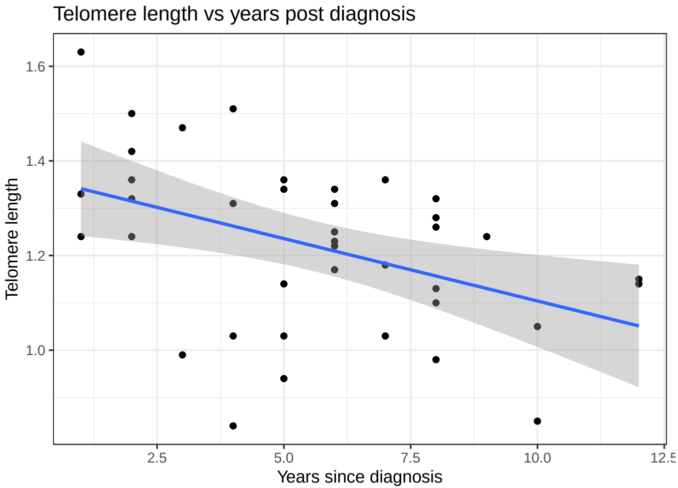
The multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + e_i \quad e_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

Key assumptions are:

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

Telomere data

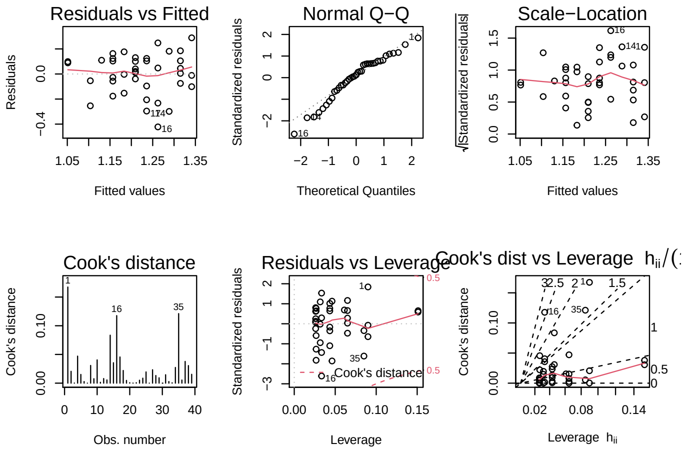


Case statistics

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

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

Default diagnostic plots in R



Leverage

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

In simple linear regression,

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

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

```
   years   leverage
37    12 0.15113547
35    10 0.08504307
39     9 0.06115897
27     8 0.04338293
25     7 0.03171496
20     6 0.02615505
12     5 0.02670321
10     4 0.03335944
8      3 0.04612373
4      2 0.06499608
1      1 0.08997651
2      1 0.08997651
```

Residuals and Fitted values

A regression model can be expressed as

$$Y_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2) \qquad \text{and} \qquad \mu_i = \beta_0 + \beta_1 X_i$$

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 the residual is

$$r_i \quad = Y_i - \hat{Y}_i$$

Standardized residuals

Often we will standardize residuals, 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 an externally studentized residual

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

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

Both of these residuals can be compared to a standard normal distribution.

Telomere data: residuals

	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
22	6	1.25	0.02615505	0.040563821	0.24999011	0.24679724
23	6	1.31	0.02615505	0.100563821	0.61976313	0.61452870
24	6	1.34	0.02615505	0.130563821	0.80464964	0.80073848
25	7	1.36	0.03171496	0.176938136	1.09357535	1.09656310
26	6	1.22	0.02615505	0.010563821	0.06510360	0.06422148

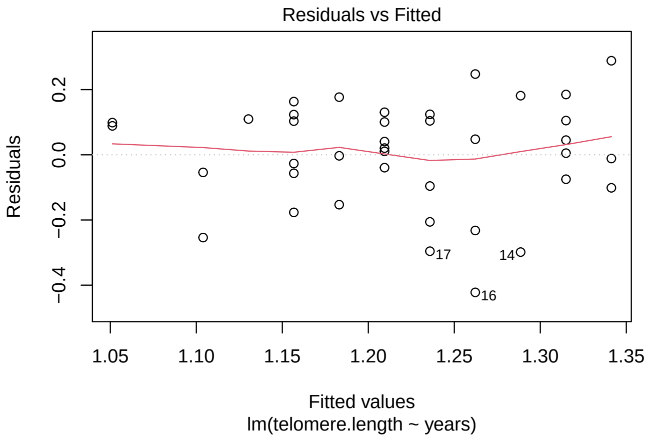
Cook's distance

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

Operationally, we might be concerned when d_i is

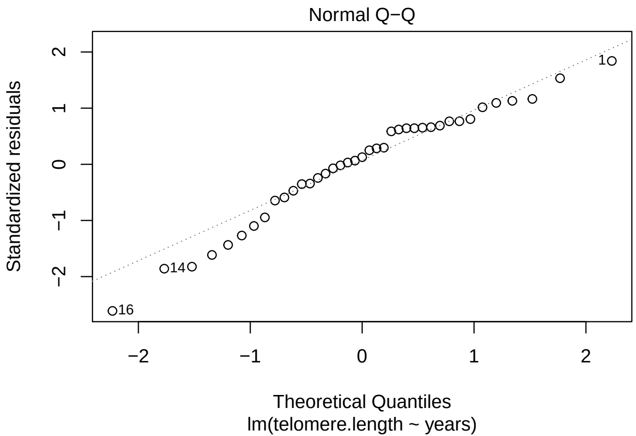
- larger than 1 or
- larger then 4/n.

Residuals vs fitted values



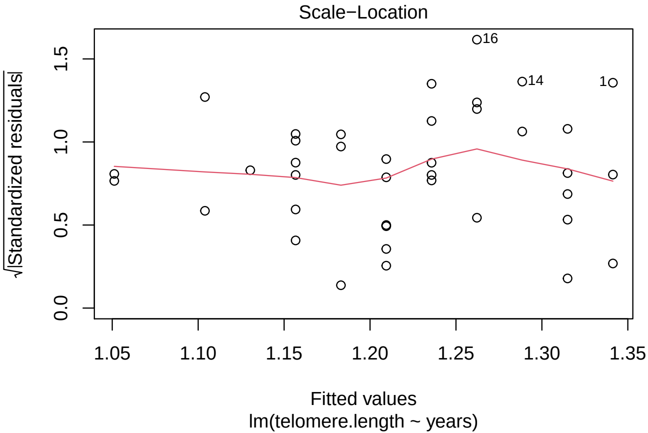
Assumption	Violation
Linearity	Curvature
Constant variance	Funnel shape

QQ-plot



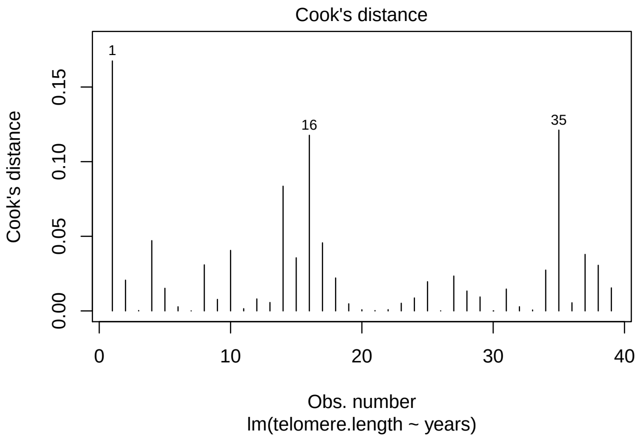
Assumption	Violation
Normality	Points don't generally fall along the line

Absolute standardized residuals vs fitted values



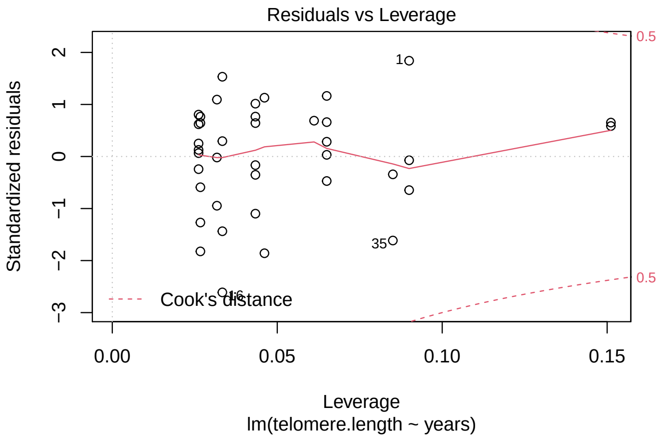
Assumption	Violation
Constant variance	Increasing (or decreasing) trend

Cook's distance



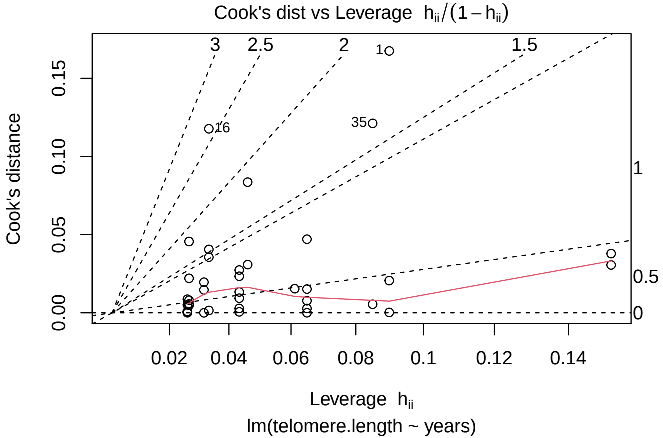
Outlier	Violation
Influential observation	Cook's distance larger than (1 or 4/n)

Residuals vs leverage



Outlier	Violation
Influential observation	Points outside red dashed lines

Cooks' distance vs leverage



This plot is pretty confusing.

Additional plots

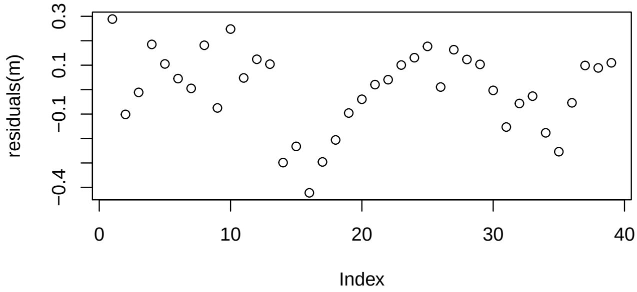
Default plots do not assess all model assumptions.

Two additional suggested plots:

- Residuals vs row number
- Residuals vs (each) explanatory variable

Plot residuals vs row number (index)

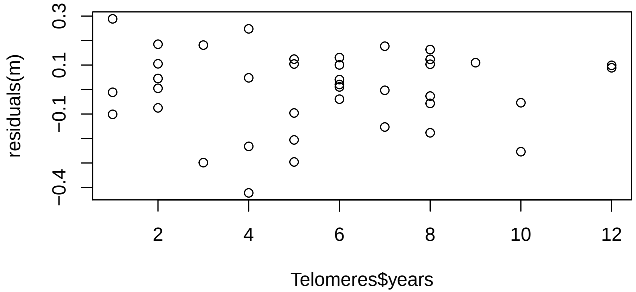
```
plot(residuals(m))
```



Assumption	Violation
Independence	A pattern suggests temporal correlation

Residual vs explanatory variable

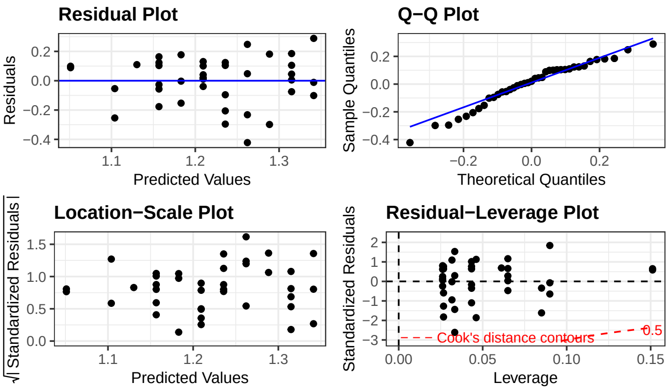
```
plot(Telomeres$years, residuals(m))
```



Assumption	Violation
Linearity	A pattern suggests non-linearity

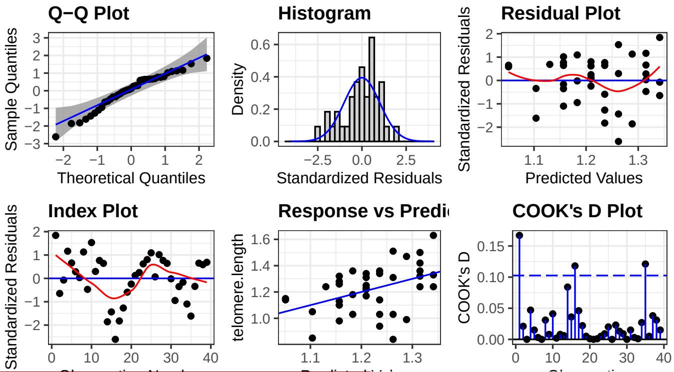
ggResidpanel: R default

```r
resid_panel(m, plots = "R")
```



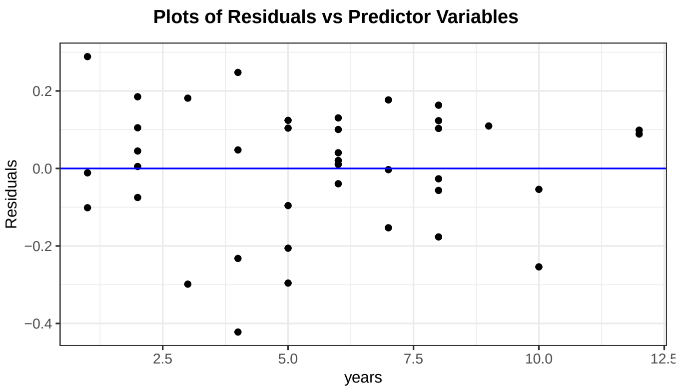
ggResidpanel: R all plots

```
resid_panel(m, plots = c("qq", "hist", "resid", "index", "yvp", "cookd"),
            bins = 30, smoother = TRUE, qqbands = TRUE,
            type = "standardized") # what I was calling studentized
```



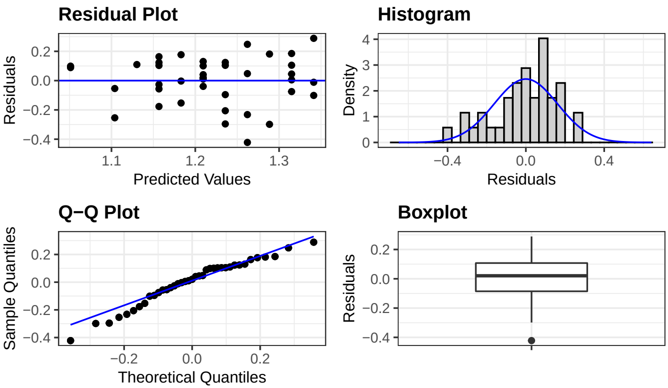
ggResidpanel: R explanatory

```
resid_xpanel(m)
```



ggResidpanel: SAS

```
resid_panel(m, plots = "SAS")
```



Summary

Case statistics:

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

Model assumptions:

- Normality
- Constant variance
- Independence
- Linearity