DATA SCIENCE

Dr. Karthik Chandrasekhar

Department of Science and Humanities

karthikchandrasekhar@pes.edu

DATA SCIENCE

Checking Assumptions and Transforming Data

Karthik Chandrasekhar
Department of Science and Humanities

DATA SCIENCE

Checking Assumptions to form a Linear Model

The methods discussed so far are valid under the assumption that the relationship between the variables x and y satisfies the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where the errors ε_i satisfy assumptions 1 through 4. These assumptions are:

1. The errors ε_1 , . . . , ε_n are random and independent. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1}.
2. The errors ε_1 , . . . , ε_n all have mean 0.
3. The errors ε_1 , . . . , ε_n all have the same variance, which we denote by σ^2.
4. The errors ε_1 , . . . , ε_n are normally distributed.

DATA SCIENCE

Checking Assumptions to form a Linear Model

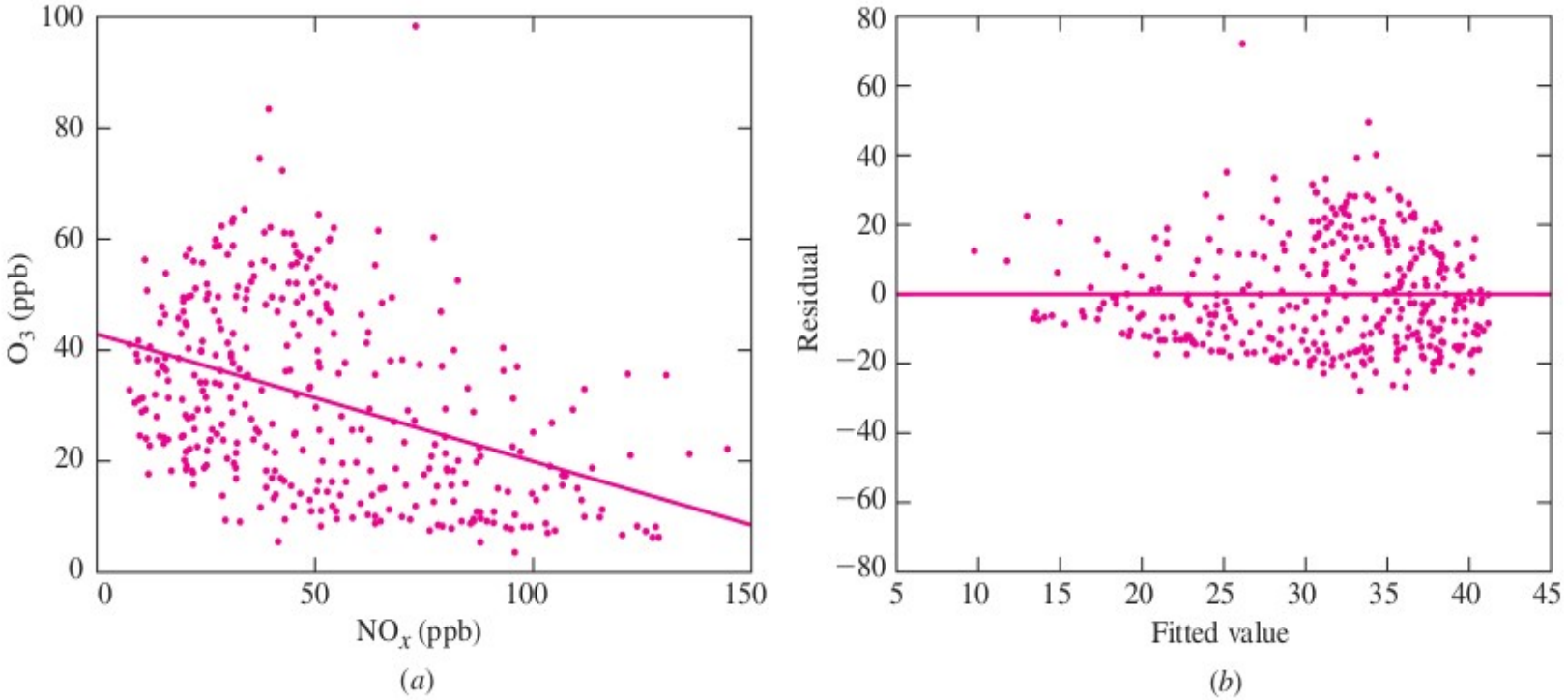
The single best diagnostic for least-squares regression is a plot of residuals e_i versus fitted values y_i, sometimes called a **residual plot.**

Recall that the residuals e_i are the difference $y_i - y_i$ between the true value y_i and the fitted value y_i, versus the fitted value y_i.

DATA SCIENCE

Checking Assumptions to form a Linear Model

- **Example of a residual plot:** On the left is the plot of x versus the fitted values of y, on the right the residual with the fitted values of y



(Source: “Statistics for Engineers and Scientists” - William Navidi)

DATA SCIENCE

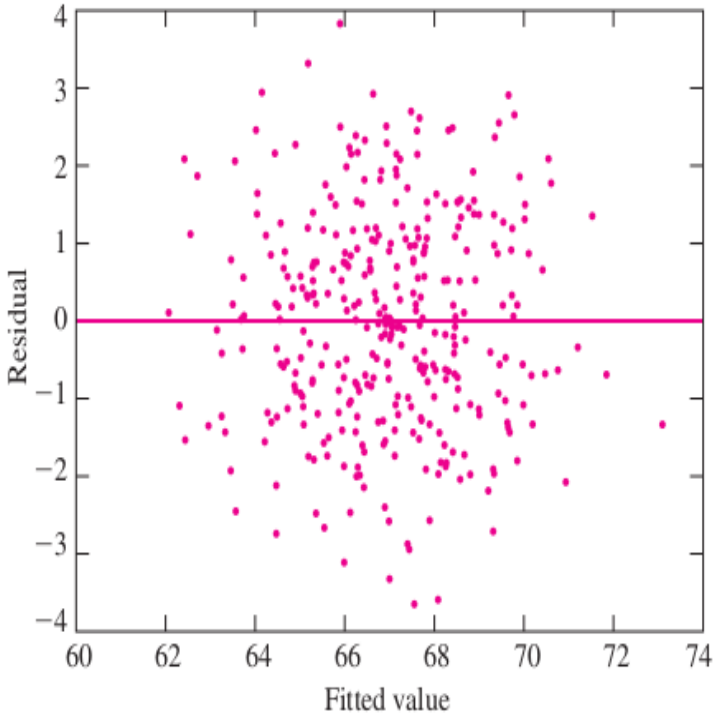
Checking Assumptions to form a Linear Model

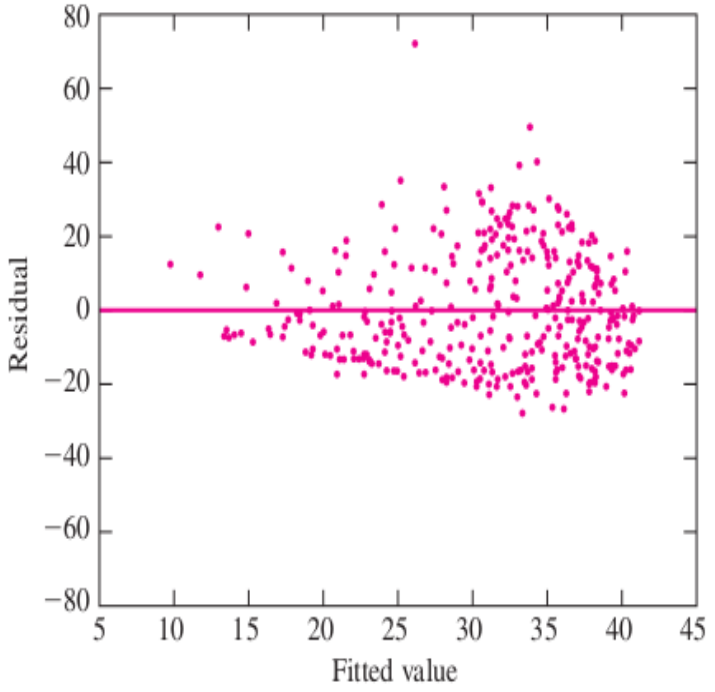
- *A bit of terminology:*
- If the vertical spread shows does not vary with the fitted value, we call the residual plot **homoscedastic**. Else we call the plot **heteroscedastic.**

DATA SCIENCE

Checking Assumptions to form a Linear Model

- Below on the left the plot is homoscedastic, while on the right the spread increases with the fitted value and is thus heteroscedastic.





(Source: “Statistics for Engineers and Scientists” - William Navidi)

DATA SCIENCE

Homoscedasticity or Heteroscedasticity? The way forward...

- If the residual plot is **homoscedastic**, and shows no substantial trend or curve, then a linear model can be found for the data plotted.
- If the residual plot is **heteroscedastic**, or shows a substantial trend or curve, then the assumptions for a linear model certainly do NOT hold! In such cases we need to transform the data or pursue other methods.

DATA SCIENCE

Which transformation to apply?

It is possible with experience to look at a scatterplot, or a residual plot, and make an educated guess as to how to transform the variables.

Mathematical methods are also available to determine a good transformation.

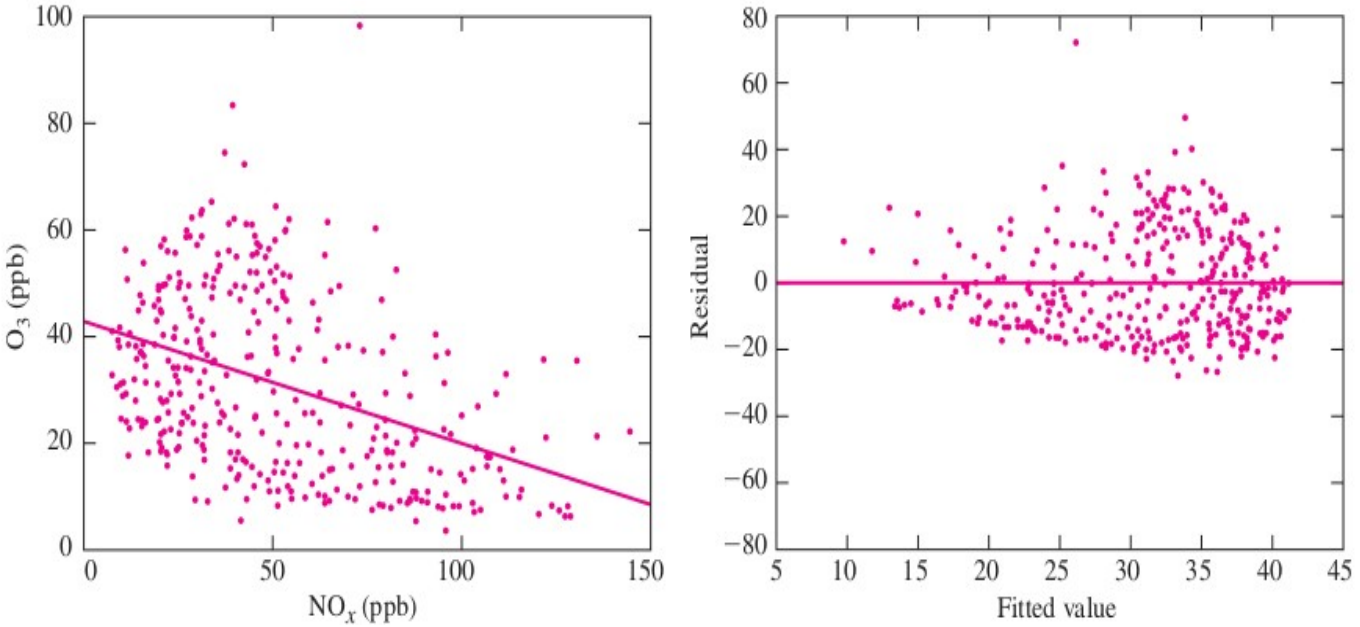
Trial and Error is fine – Try various powers on both x and y (including ln x and ln y), look at the residual plots, and hope to find a homoscedastic one with no discernible pattern.

More advanced discussion in Draper and Smith (1998).

DATA SCIENCE

Which transformation to apply?

Recall the earlier example of a scatter plot (O_3 concentration vs NO_X concentration) whose residual plot on the right is heteroscedastic as shown below. Linear model NOT GOOD! Uh oh! Also notice the outlier with ozone concentration nearly 100.

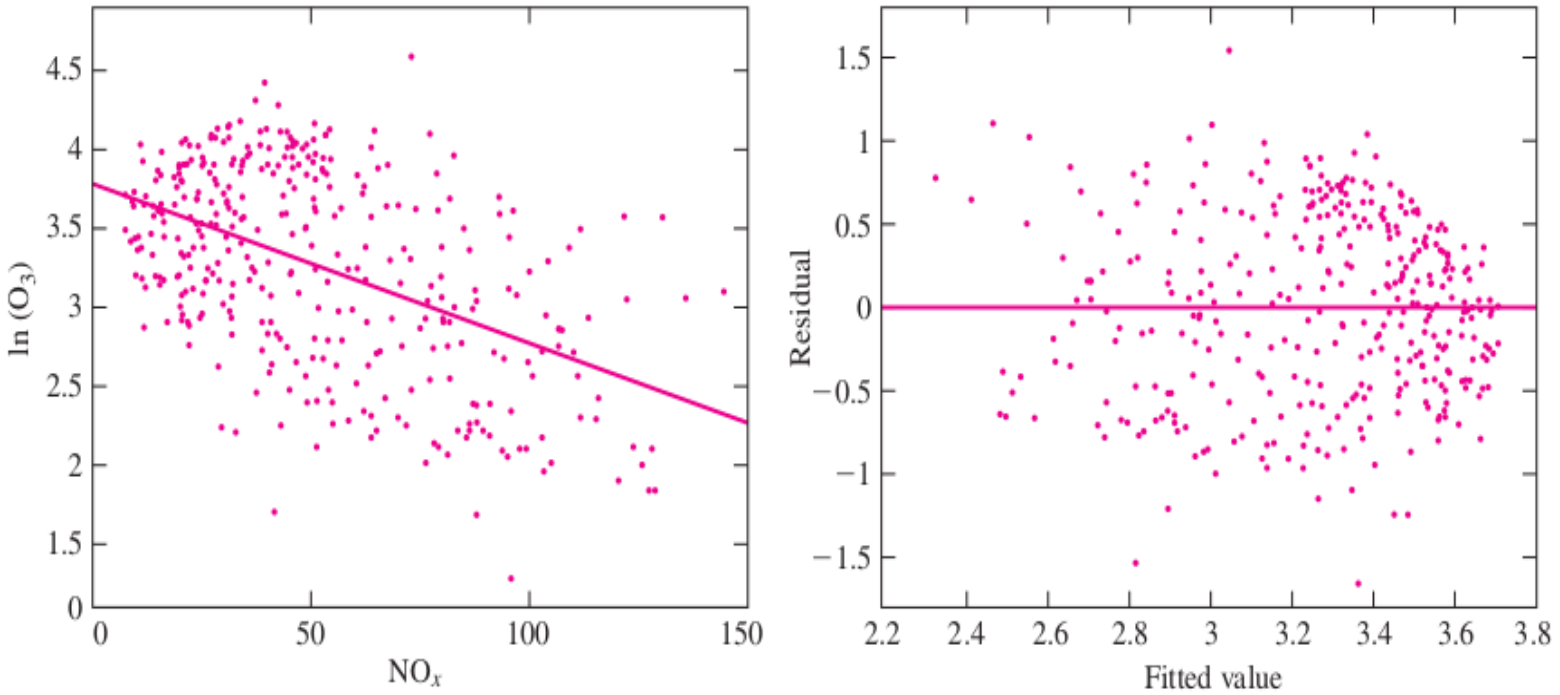


(Source: “Statistics for Engineers and Scientists” - William Navidi)

DATA SCIENCE

Logarithm Transformation on One Axis

Applying the logarithm on y-axis (O_3 concentration) and obtain the following scatter plot and its residual on the right. Linear model looks GOOD! YAY! The outlier is less prominent too!

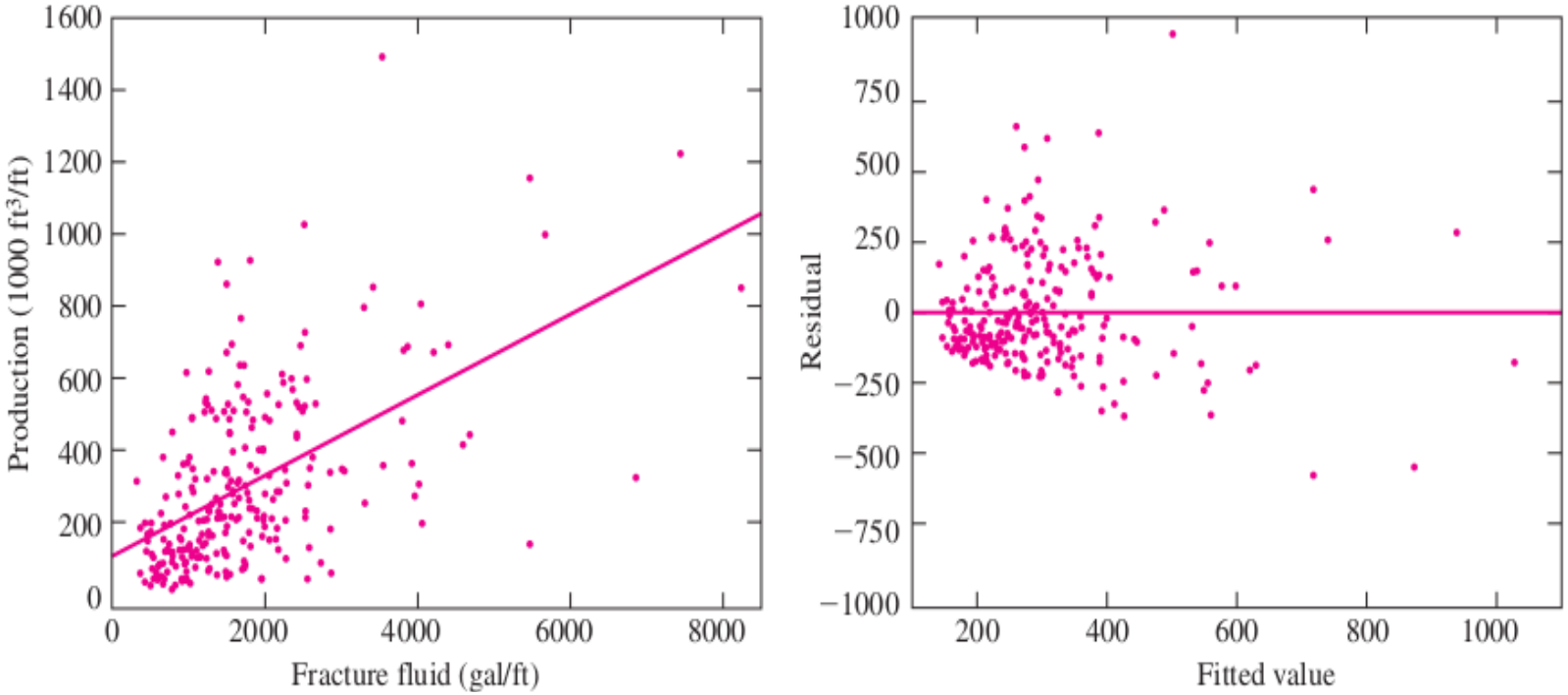


(Source: “Statistics for Engineers and Scientists” - William Navidi)

DATA SCIENCE

Logarithm Transformation on Both Axes

Now consider an example below where The plot on the left is Production (ft^3/ft) vs Fracture fluid (gal/ft) and the residual plot is largely heteroscedastic! Not good for a linear model.

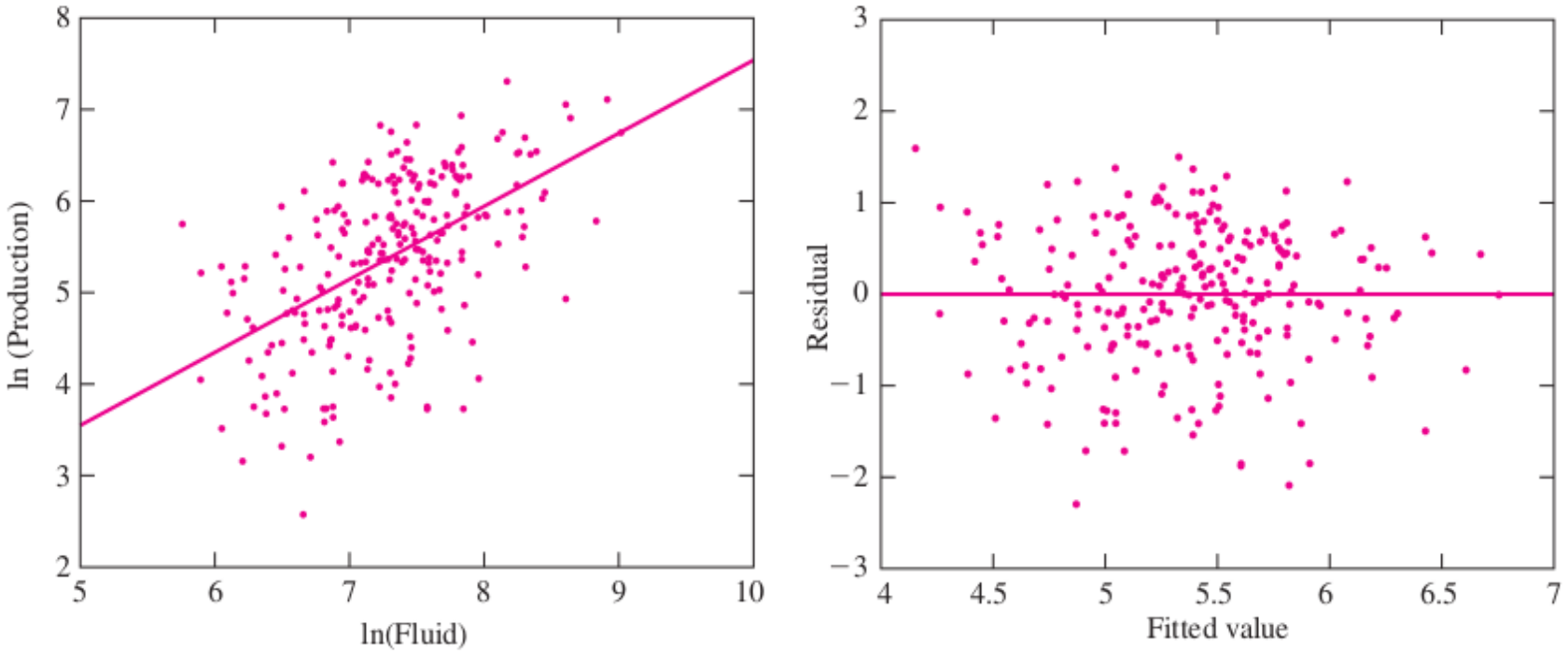


(Source: “Statistics for Engineers and Scientists” - William Navidi)

DATA SCIENCE

Logarithm Transformation on Both Axes

Below is a plot of ln (production) vs ln (fracture fluid) for the same data. This time the residual plot is homoscedastic, good for linear model!

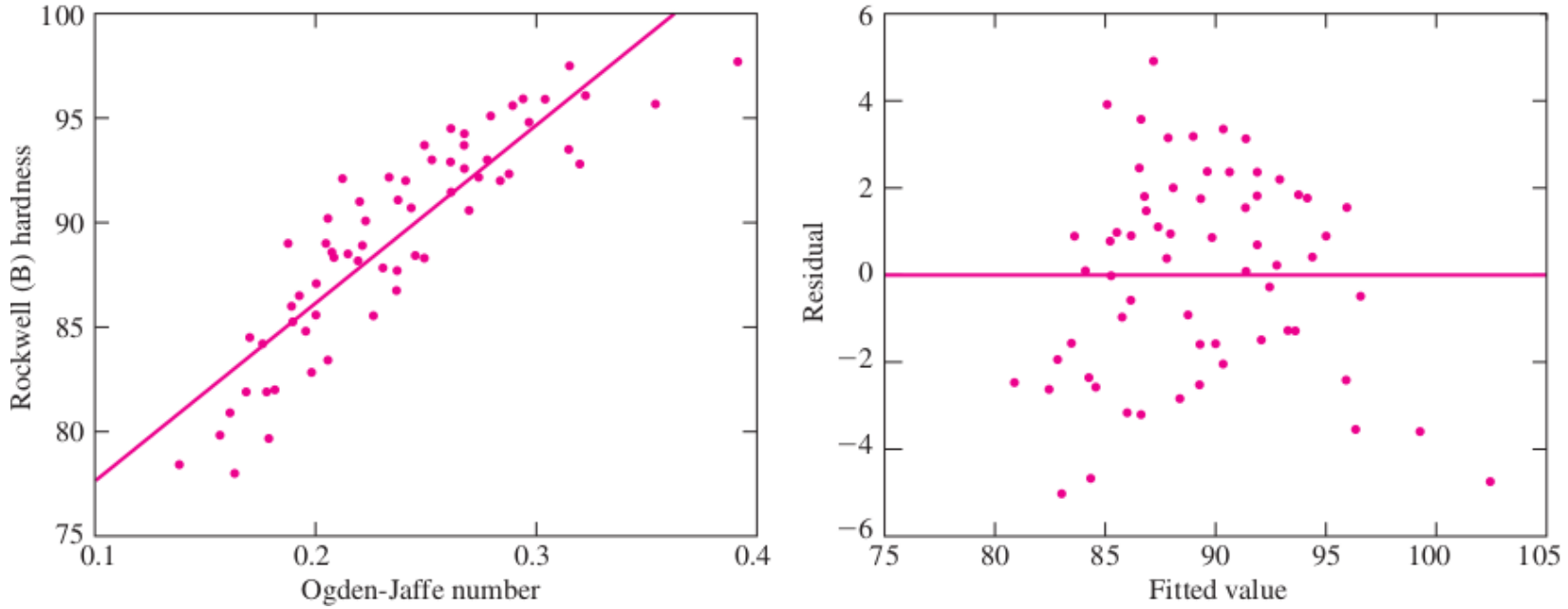


(Source: “Statistics for Engineers and Scientists” - William Navidi)

DATA SCIENCE

Power Transformations – The reciprocal

Below (left side) is a plot of Rockwell (B scale) hardness of welds versus their Ogden-Jaffe number. The residual plot (right side) shows a pattern where negative residual is observed for the extreme fitted values and positive residual for the ones in the middle. Linear model NOT OK.



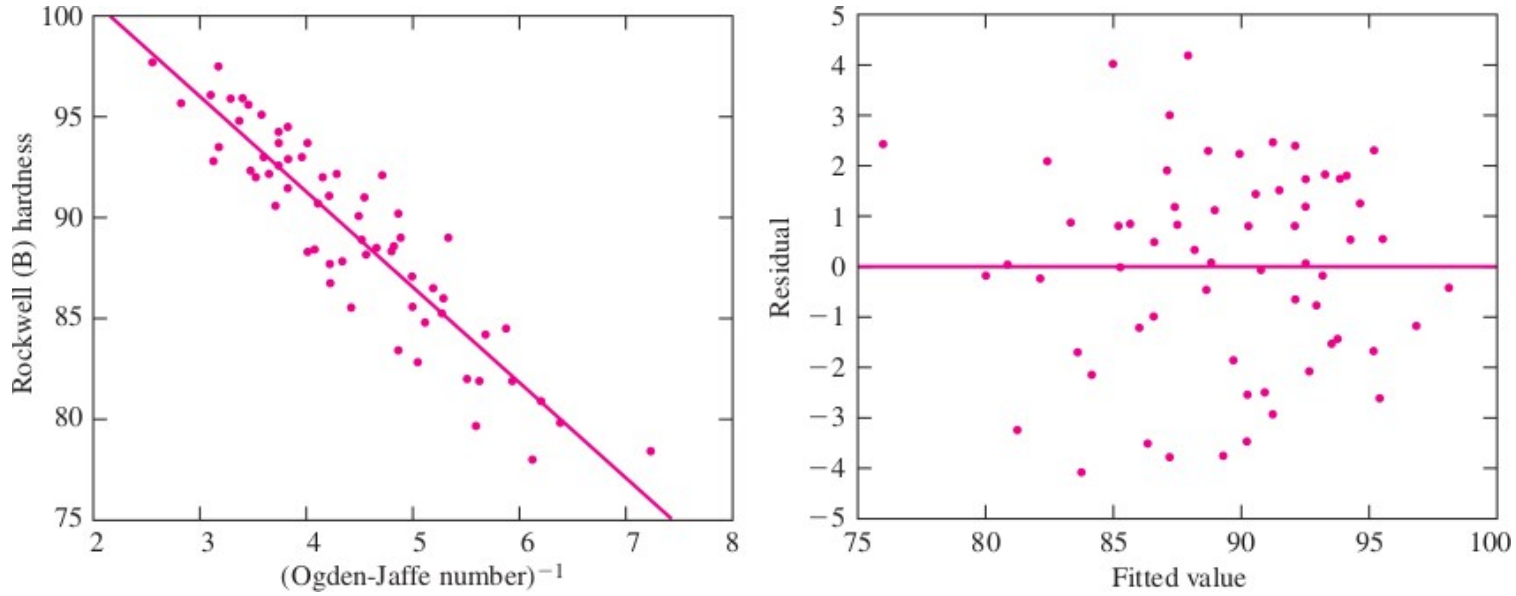
(Source: “Statistics for Engineers and Scientists” - William Navidi)

DATA SCIENCE

Power Transformations – The reciprocal

We plot the graph of Rockwell Hardness vs (Ogden-Jaffe)$^{-1}$ for the same data (below, left side) and find that the residual plot (below, right side) is homoscedastic, having no discernible pattern.

Linear model is OK.



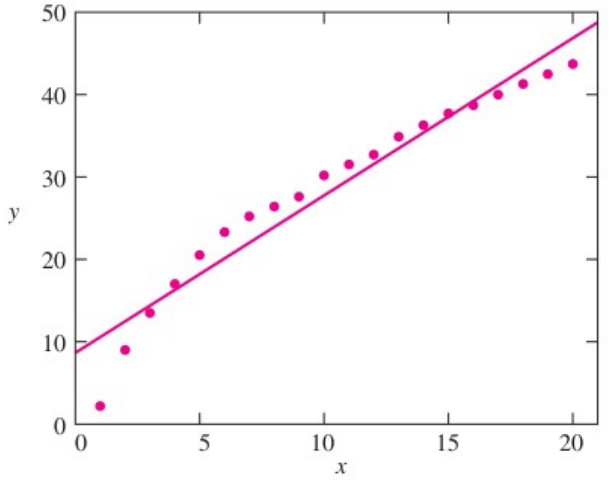
(Source: “Statistics for Engineers and Scientists” - William Navidi)

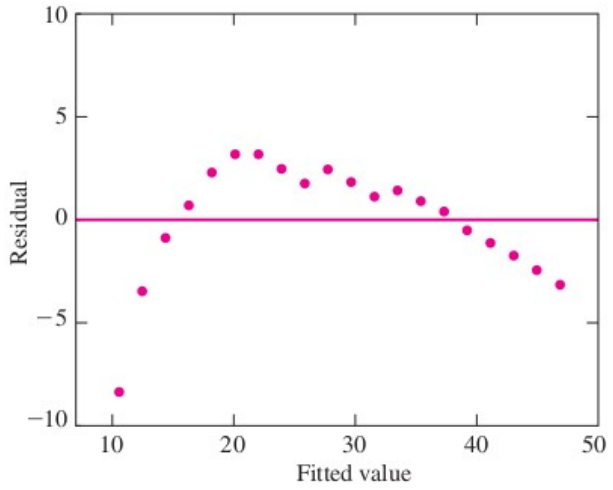
DATA SCIENCE

Power Transformations – Positive Powers

Plot y vs x and its residual plot which exhibits a discrenible pattern.
Linear model is NOT OK.

x	y	x	y
1	2.2	11	31.5
2	9	12	32.7
3	13.5	13	34.9
4	17	14	36.3
5	20.5	15	37.7
6	23.3	16	38.7
7	25.2	17	40
8	26.4	18	41.3
9	27.6	19	42.5
10	30.2	20	43.7





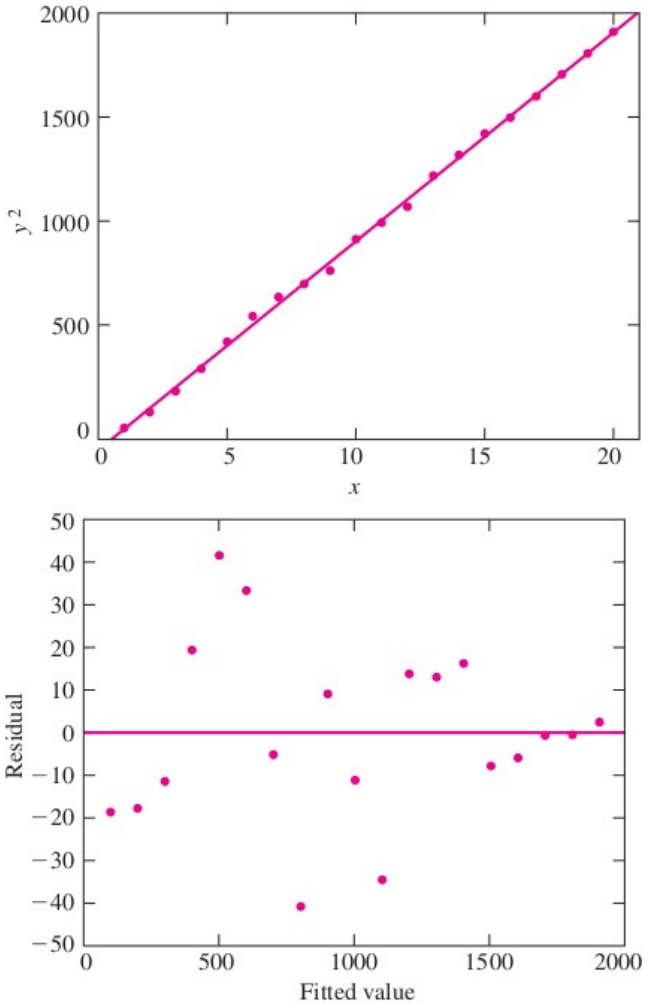
(Source: “Statistics for Engineers and Scientists” - William Navidi)

DATA SCIENCE

Power Transformations – Positive Powers

Plot y^2 vs x and its homoscedastic residual plot which exhibits no discernible pattern.
Linear model is OK.

x	y^2		x	y^2
1	4.84		11	992.25
2	81		12	1069.29
3	182.25		13	1218.01
4	289		14	1317.69
5	420.25		15	1421.29
6	542.89		16	1497.69
7	635.04		17	1600
8	696.96		18	1705.69
9	761.76		19	1806.25
10	912.04		20	1909.69



(Source: “Statistics for Engineers and Scientists” - William Navidi)

DATA SCIENCE

Transformations – Do they always work?

It is important to remember that power transformations don't always work.

Sometimes, none of the residual plots looks good, no matter what transformations are tried. In these cases, other methods should be used. One of these is multiple regression which is not covered here.

Some other methods are briefly mentioned in the next slide.

DATA SCIENCE

Alternatives to Transformations

The popular methods other than transformation are:

- Weighted Least Squares
  - We assign greater weights to points in regions where the vertical spread is smaller and vice versa.

- Multiple Regression
  - We add more independent variables in order to explain the variation in the dependent variable. The reader is directed to Draper and Smith (1998) for more about this topic.

DATA SCIENCE

How Many Points Make a Reliable Residual Plot?

When there are too few points on the residual plot, then…

- … it may appear to have a pattern or be heteroscedastic in spite of that being just a visual effect created by one or two points.

- … detecting outliers may become difficult

What to do if you can’t interpret a residual plot reliably?

You can start by fitting a linear model but declare your result tentative; wait for more data and then a reliable decision can be made.

DATA SCIENCE

How Many Points Make a Reliable Residual Plot?

NOT all residual plots with few points turn out to be hard to interpret.

Some of these show a pattern which cannot be changed by relocating just one or two points.

In such a case a linear model should NOT be used!

THANK YOU

Karthik Chandrasekhar
Department of Science and Humanities
karthikchandrasekhar@pes.edu