STAT 302 - Chapter 1 : Simple Linear Regression - Part 3

Harsha Perera

Key Topics

- Transformations/Reexpressions
- Outliers and Influential Points
- Extrapolation

Transformations/Reexpressions

If one or more of the conditions for a simple linear regression model are not satisfied, we can consider **transformations** on one or both of the variables to address the problems.

Examples :

- Lack of normality in residuals
- Patterns in residuals
- Heteroscedasticity (nonconstant variance)

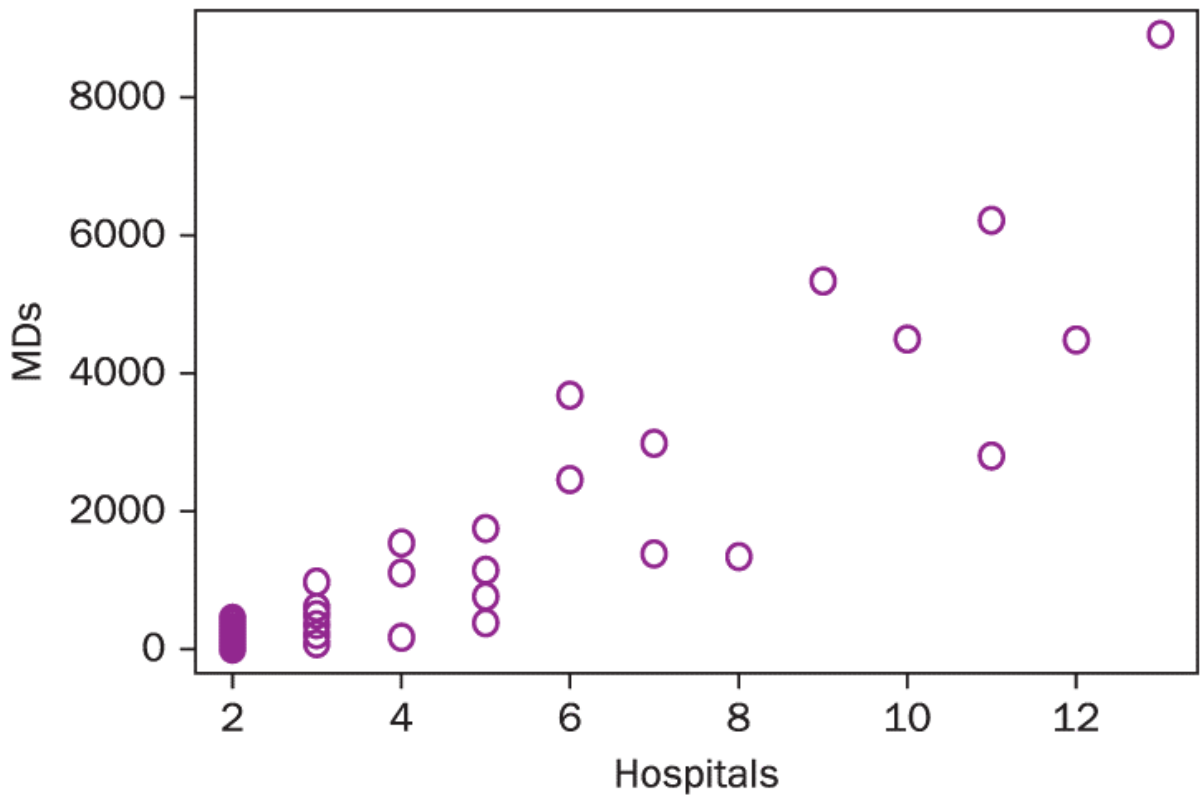
We can use Data Transformations/Reexpressions to :

- Address nonlinear patterns
- Stabilize variance
- Remove skewness from residuals

Transformations/Reexpressions - Example

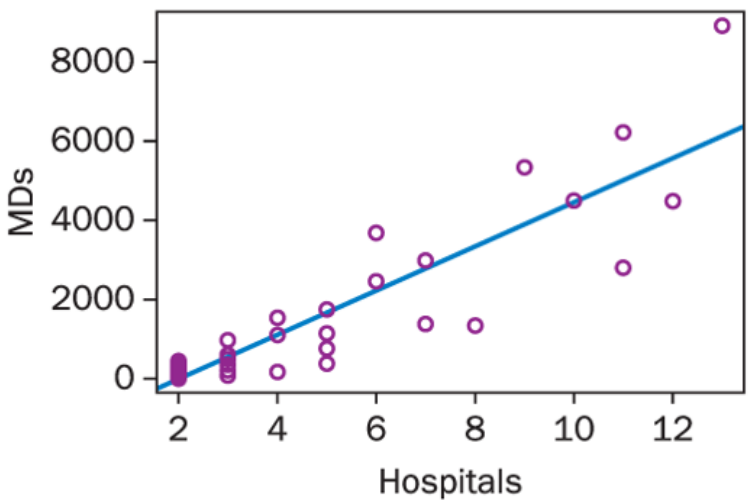
- Example 1.7 - Doctors and hospitals in counties on page 35

- Start the process of finding a model to predict the number of doctors (MDs) from the number of hospitals (Hospitals) by examining the scatterplot of the two variables.

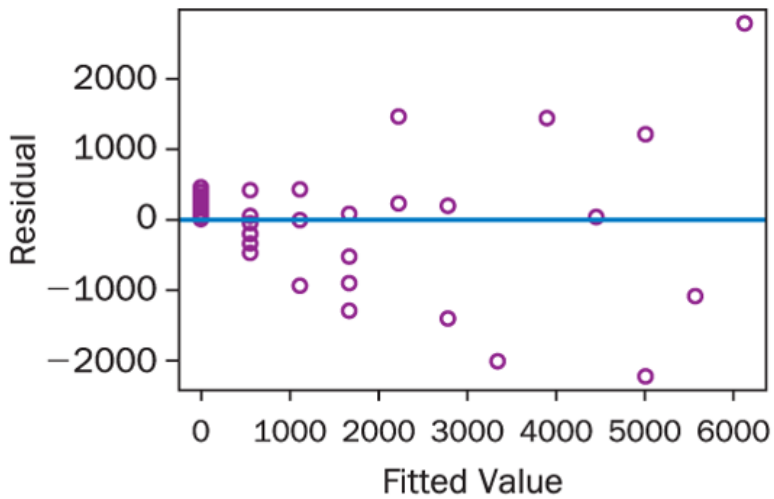


Example -1.7

- The least square regression line does a fairly good job showing the increasing trend between MDs and Hospitals.
- However the residual plots show some considerable departures from our standard regression conditions. Figure (b) shows a fan shape which is the variability in the residuals tending to increase as the fitted values get larger.



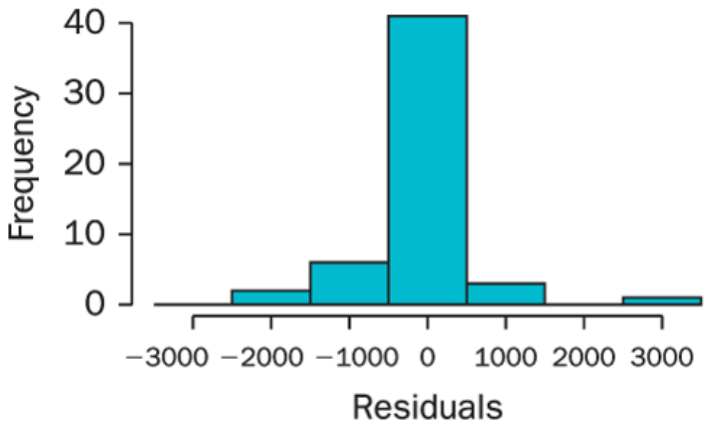
(a) Least squares line



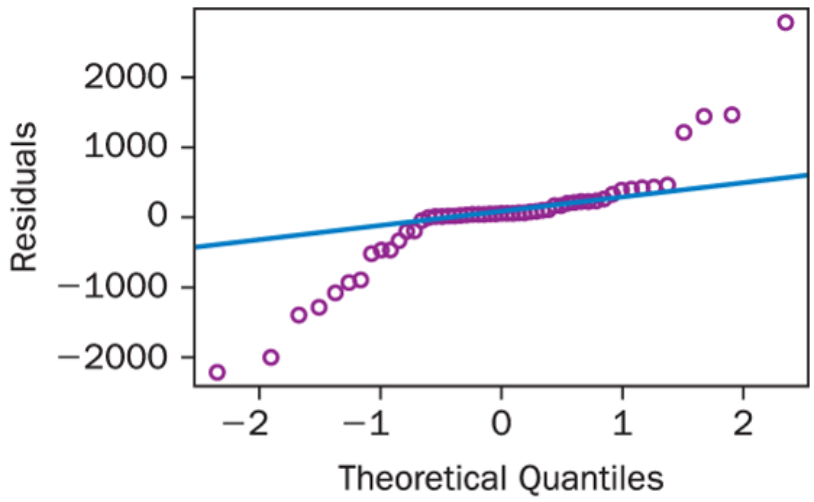
(b) Residuals versus fits

Example -1.7

- Histogram of the residuals and normal quantile plot show the violation of the assumption of normality.



(a) Histogram of residuals



(b) Normal quantile plot

- Since the assumptions are vioalted we have to do a transformation/reexpression.

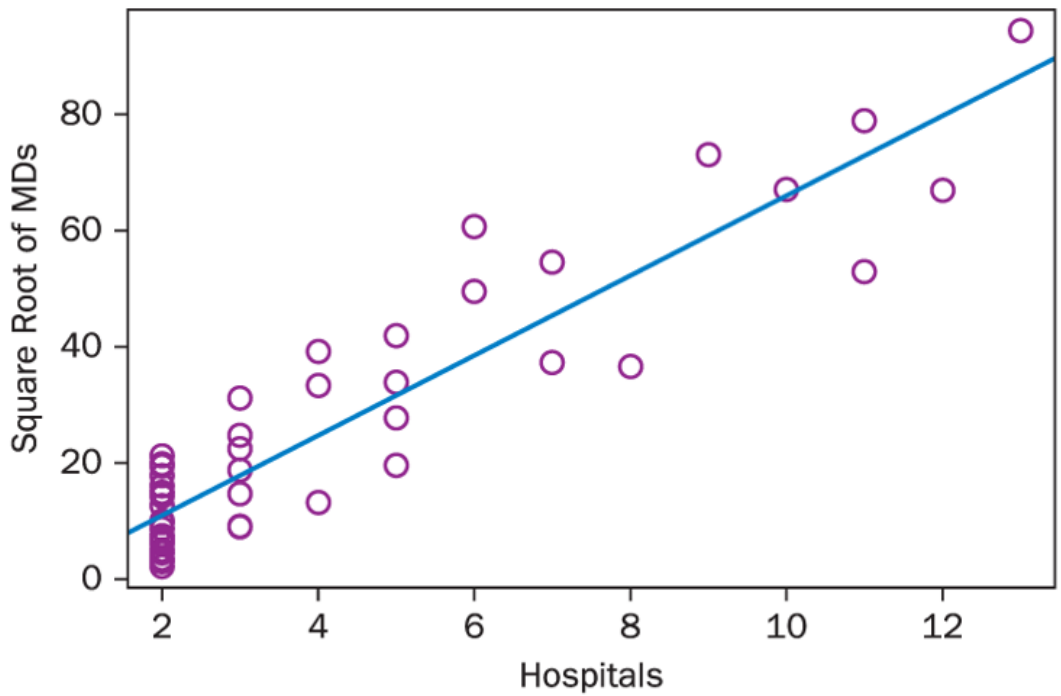
Common Transformations

For either the response (Y) or predictor (X) or both...

- Square root $Y \rightarrow \sqrt{Y}$
- logarithm $Y \rightarrow log(Y)$
- Power Function $Y \rightarrow Y^2$, Y^3 etc
- Reciprocal $Y \rightarrow 1/Y$

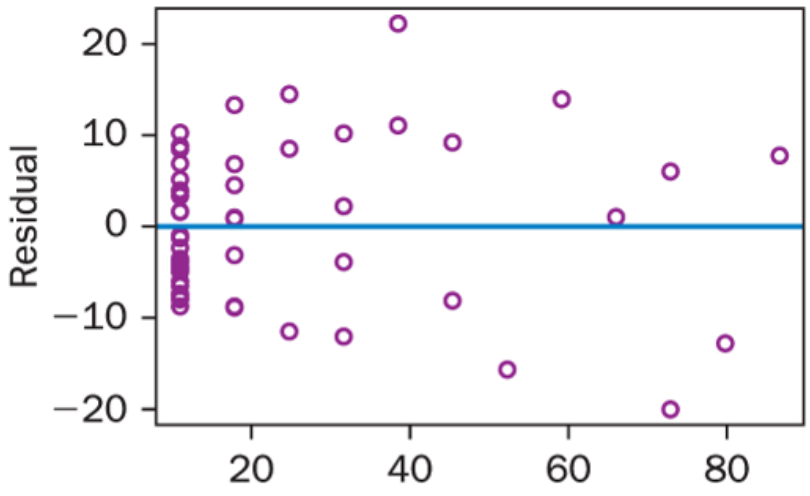
Example 1.7

- Since we have count data in this data set, such as the number of doctors and hospitals, a **square root** transformation is useful.
- The scatter plot shows the least squares regression line fit to the square root transformed data.

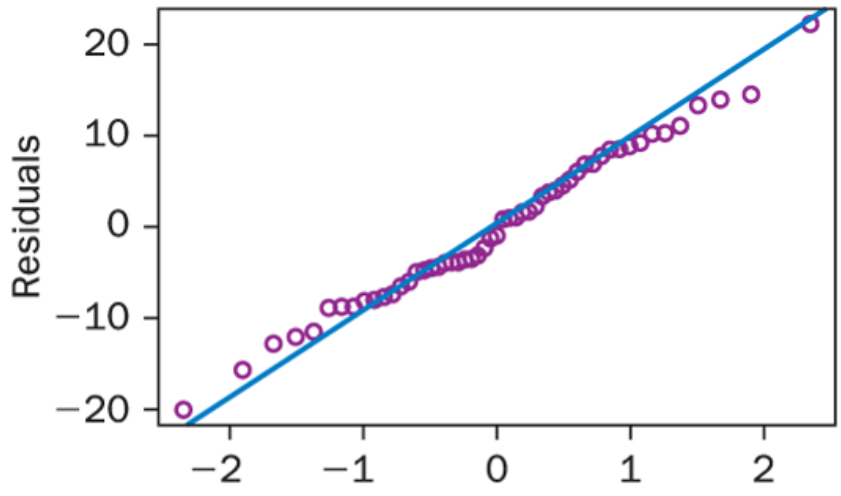


Example 1.7

- The residuals versus fitted values plot for the transformed data and the normal quantile plot of the residuals show a considerable improvement at meeting the constant variance and normality conditions.



(a) Residuals versus fits



(b) Normal plot of residuals

- Remember that our transformed linear model is predicting $\sqrt{MDs}$, so we must **square** the predicted values to obtain the actual number of doctors.

Choosing a Transformation

- There is no guarantee that transformations will eliminate or reduce the problems with departures from the conditions for a simple linear regression model.
- Often a logarithm transformation is very helpful, but not always.
- Finding an appropriate transformation can be as much an art as a science, yet there are some guidelines.
- Please refer the text book section about "Choosing a Transformation" on page 40 and 41 and refer examples 1.8(Species by area) and 1.9(Areas of circles).

Outliers and Influential Points

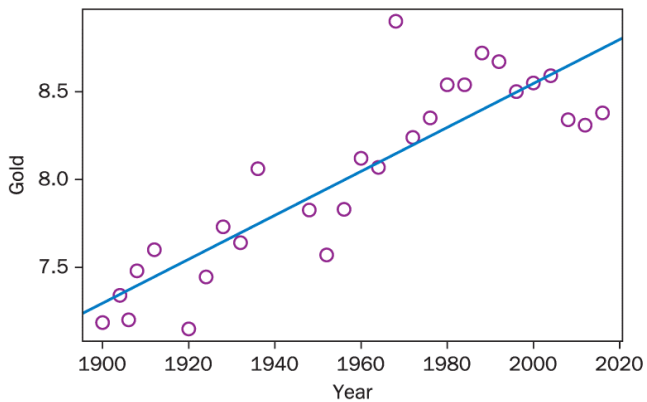
- **Outlier** - An unusual data point that doesn't fit with the other points in a scatterplot vertically.
- **Influential point** - A data point which differs from the others horizontally and vertically.
- Will examine some quick methods for detecting outliers and influential data points using **graphs** and **summary statistics**.

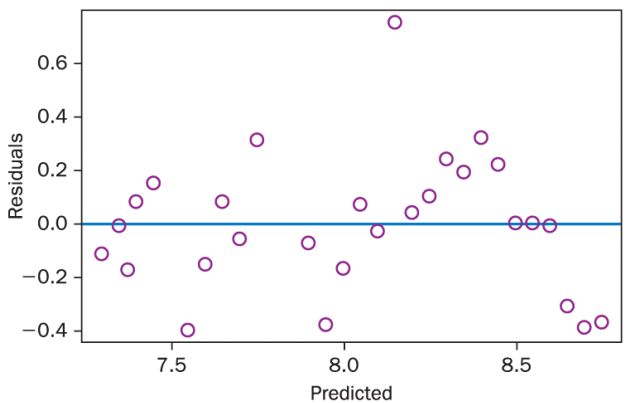
Outliers

- In the setting of simple linear regression, we call a data point an "outlier" if the magnitude of the residual is unusually large.
- How large must a residual be for a data point to be called an outlier ?
- That depends on the variability of all the residuals.

Example 1.10 - Olympic Long Jump

- The 1968 data point clearly stands above the others and is far removed from the regression line. Because this point doesn't fit the general pattern in the scatterplot, it is an **outlier**. The unusual nature of this point is perhaps even more evident in the residual plot for the fitted least squares model. The 1968 residual is 8.90 - 8.15 = 0.75



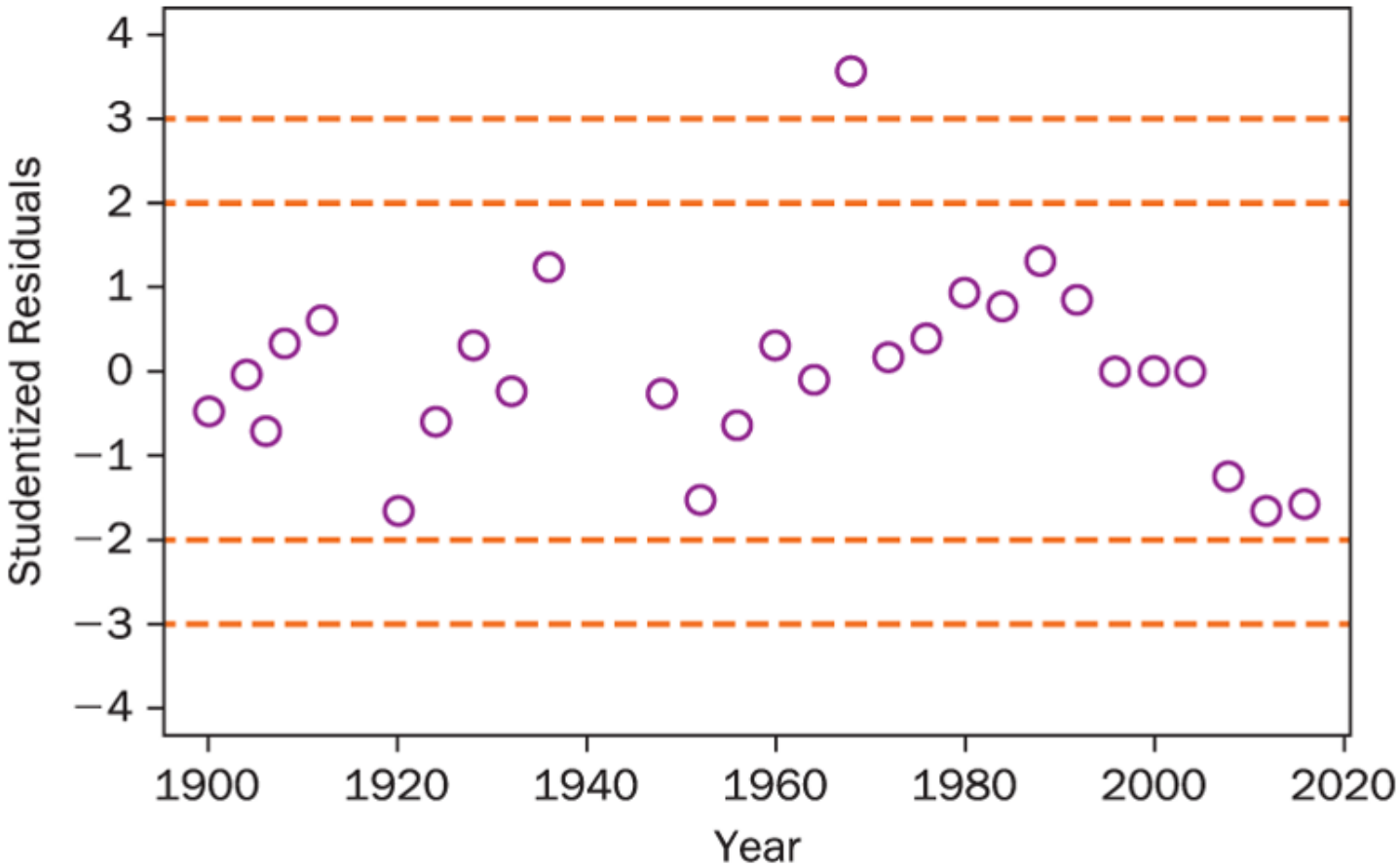


Different Types of Residuals

- One method to help decide when a residual is extreme is to put the residuals on a standard scale.
- Since the estimated standard deviation of the regression error, $\hat{\sigma}_\epsilon$ reflects the size of a typical error, we could standardize each residual using $\frac{y-\hat{y}}{\hat{\sigma}_\epsilon}$
- Standardized residual
- Studentized residual (or deleted-t residual)

Studentized Residual Plot

If the conditions of a simple linear model hold, approximately 95% of the residuals should be within 2 standard deviations of the residual mean of zero. So we would expect most standardized or studentized residuals to be less than 2 in absolute value



How should we deal with points like this?

- The answer will really depend. For example, it could be the case that after identifying an unusual point like this, you can examine the original data and realise that someone simply wrote the number down wrong. But in general, unless there is some good reason to change or remove the point, it should probably be retained. In analysing data, it would be a good idea to highlight such unusual points, even if we cannot say any more.

- In general, outliers should not be removed unless there is sufficient justification that data point should not be in the same sample as the rest of the data.

Influential Points

- If we see an unusual point in the horizontal direction , then we need to think about how that point influences the overall fit of the regression model.

- The amount of **influence** that a single point has on a regression fit depends on how well it aligns with the pattern of the rest of the points and its value for the predictor variable.

- Generally, points farther from the mean value of the predictor ($\bar{x}$) have greater potential to influence the slope of a fitted regression line. This concept is known as the **leverage** of a point.

- One potential solution would be to perform the analysis with and without this point included, which will highlight the difference it makes to the model. Then, someone with extra knowledge of the data may be able to help you decide what to do.

Extrapolation

- Trying to make a prediction for an unusually large or small value of the explanatory (x) variable.
- Results may be very unreliable when trying to predict for a value that is far from the data that is used to build the model.
- If we try to use the model in example 1.3 (Honda Accord Prices) to predict the price of a Honda Accord with 200 thousand miles (which is outside the X data range), we would see the predicted price as -3.15. Clearly we shouldn't expect to get money and the car.