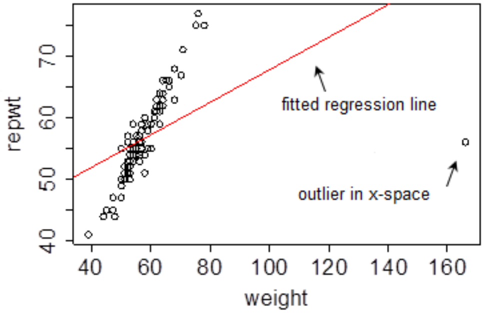
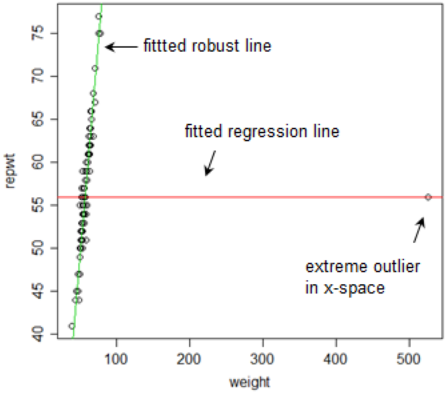
GLM-Part 1

Robust and Resistant Regression Techniques



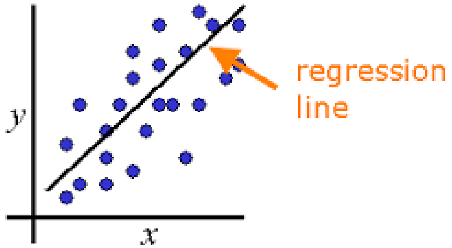


Dr. Rita Chakravarti
Institute of Systems Science
National University of Singapore
Email: rita@nus.edu.sg

Topics

- Revisit Linear Regression
- Outliers and Influential Observations
- Robust and Resistant Regression techniques
- How to Fit Robust Regression Models

Revisit Linear Regression



- We have data linking a response variable y with an independent set of variables .
- We wish to fit a model of the form

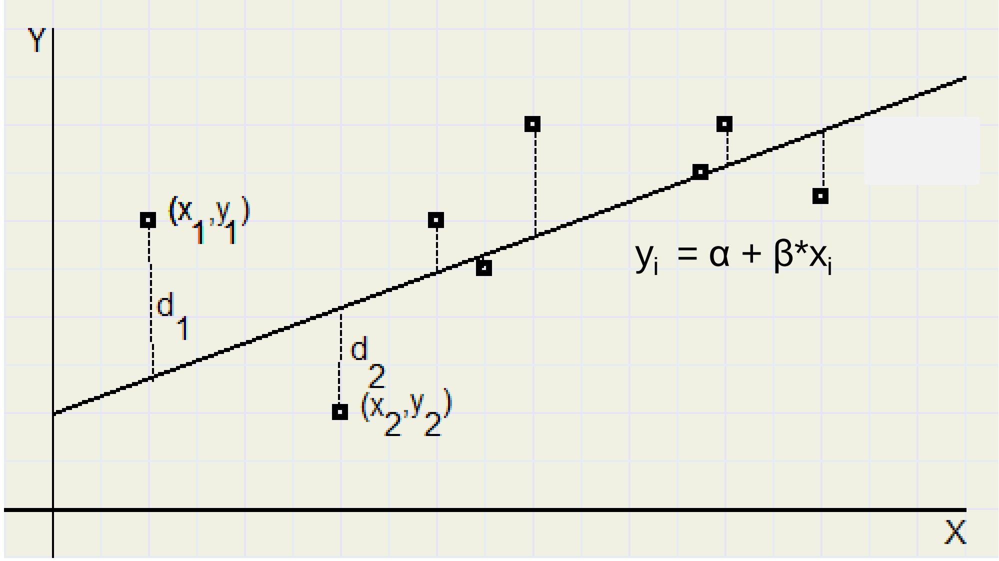
$y_i = \alpha + \beta * x_i + \varepsilon_i$, i = 1, ……..n , n observations

Assumptions of regression analysis (LINE)

- 1. The relation is linear so that the errors all have expected value zero; $E(\varepsilon_{i}) = 0$ for all i **(L)**
- 2. The errors are independent of each other **(I)**
- 3. The errors are all normally distributed: ε_{i} is normally distributed for all *i* **(N)**
- 4. The errors all have same equal variance: $Var(\varepsilon_{i}) = \sigma^2$ for all *i* **(E)**

Method of Least Squares

- Suppose we have collected two sets of observations (y_i ,x_i), i=1,…n for n individuals
  - y_i is the Weight
  - x_i is the explanatory variable Height



- Suppose further we assume a model of the form

$$y_i = \alpha + \beta{*}x_i$$

- Then if we know α , β and x_i, then we would predict y_i by $\hat{y}_i$, where

$$\hat{y}_i = \alpha + \beta{*}x_i$$

- If we represent this by a straight line through the points then the deviation between the ith point ($\hat{y}_i$, x_i) and the predicted ($\hat{y}_{i,}$, x_i) is given by

$$d_i = \hat{y}_i - y_i$$

- If we let the sum of deviations be

$$SSE = \Sigma_{i=1,n} \; (y_i - \hat{y}_i)^2 = \Sigma_{i=1,n} \, \varepsilon_i^2$$

then it would be reasonable to attempt to minimise SSE by choosing the correct line, i.e. by choosing α & β to minimise SSE

- This method is called the method of least squares and our estimates of β and α are
  - $\hat{\beta} = \dfrac{\Sigma\,(y_i{*}(x_i - \bar{x}))}{\Sigma\,(x_i - \bar{x}\,)^2}$ $\quad\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x}\quad$ where $\bar{x} = \dfrac{\Sigma\, x_i}{n}$ etc..

Limitations of the Least Square Linear Models

- Usually relies on the errors being normally distributed
- Very sensitive to outliers and influential observations
- Has issues with small data sets and poor quality data

Data with outliers and/or fails to meet the Regression Analysis assumptions

The data set : Belgian Phone represents the number of phone calls in millions in Belgium between 1950 and 1973.

There is a strong linear relationship with time, except that the values for 1963 through 1970 are very much above the apparent regression line for the other points (this is especially true for 1964 through 1969).

However, between 1964 and 1969 the total length of calls (in minutes) were recorded rather than the number, and both recording systems were used during parts of 1963 and 1970.

These outliers have a severe effect on the fitted least squares regression line, as the following plot shows:

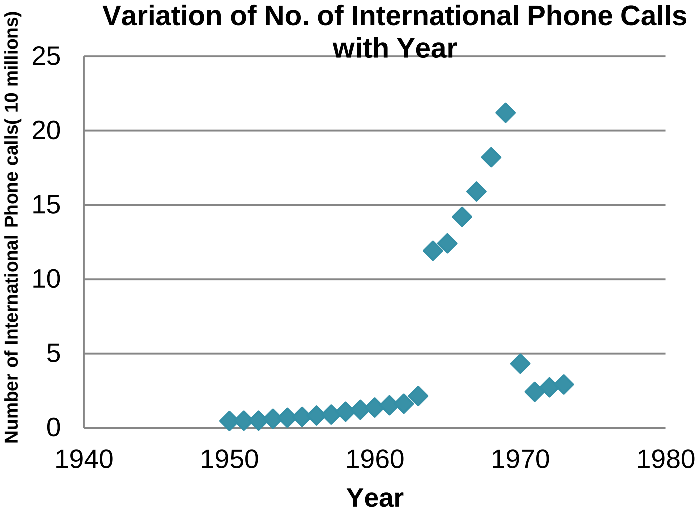
Year	No. of tens of millions of International phone calls
1950	0.44
1951	0.47
1952	0.47
1953	0.59
1954	0.66
1955	0.73
1956	0.81
1957	0.88
1958	1.06
1959	1.2
1960	1.35
1961	1.49
1962	1.61
1963	2.12
1964	11.9
1965	12.4
1966	14.2
1967	15.9
1968	18.2
1969	21.2
1970	4.3
1971	2.4
1972	2.7
1973	2.9

Data with outliers and/or fails to meet the Regression Analysis assumptions(cont.)

- Clearly the Regression line is inadequate
- The regression equation is

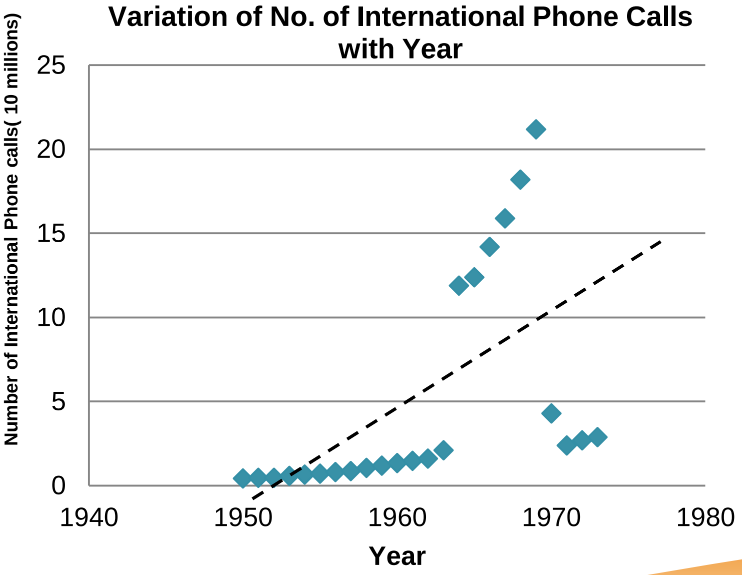
No. of phone calls = - 984 + 0.504 YEAR

Predictor	Coef	StDev	t-value	p-value
Constant	-983.9	325.2	-3.03	0.006
YEAR	0.5041	0.1658	3.04	0.006



- But R-Sq = 29.6%, R-Sq(adj) = 26.4% which is quite bad

- So how would we deal with this ?
  - If these are outliers how would we define them?

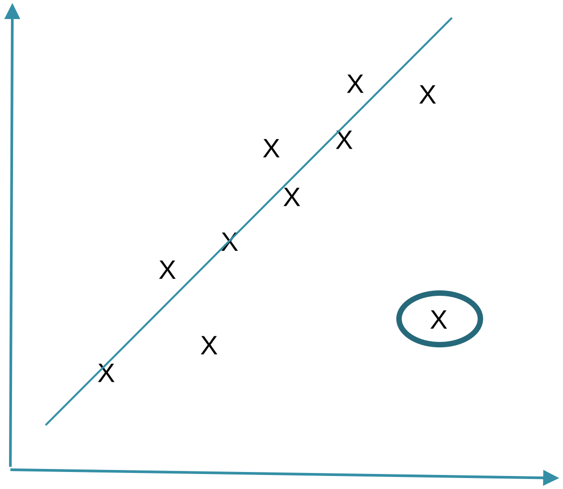


Revisit Outliers (abnormalities, discordants, deviants, anomalies)

- In **statistics**, an **outlier** is an observation point that is distant from other observations. An **outlier** may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.
- There may be reasons for the outlier :
  - There was an error in recording the value.
  - The point does not belong in the population we are supposed to be sampling .
  - However, the observation may still be valid.
- Identify outliers either from
  - The scatter diagram
  - Observation **3 standard deviations** from the mean in case of **<u>Normality</u>**
    - Examining the standardized residuals

      $s_i = (y_i-\hat{y}_i)/MSE$; where $MSE = \Sigma\,(y_i - y)^2 / (n-2)$
      It is usual to believe an observation is an outlier

      ⇨ if it's standardized residual > 3



Why can't we just remove the outliers?

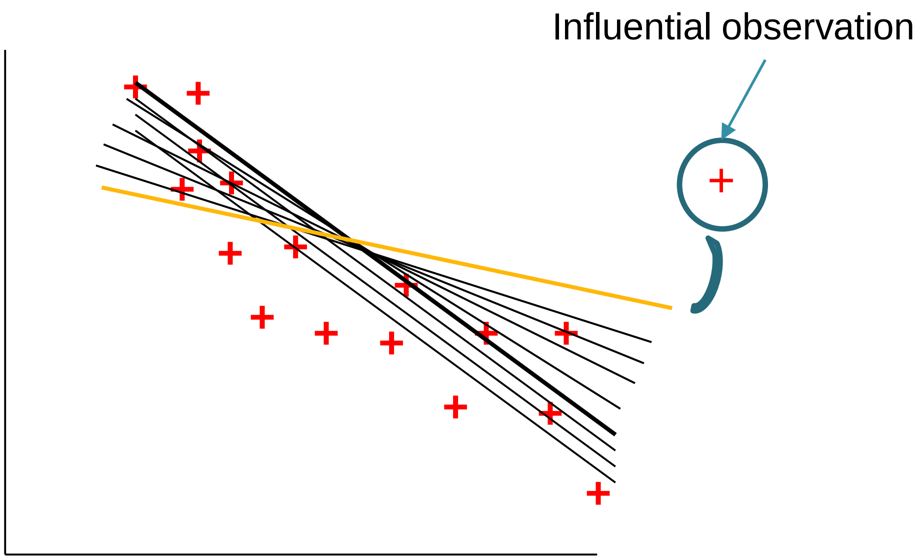
- Rejecting the outliers affects the assumed underlying distribution which ought then to be adjusted.
  - Variances will be underestimated from the "cleaned" data,
    - ⇨ Leading to incorrect significance test, confidence intervals , etc.
- Outliers may contain useful information on the limits of a sample or population
- The decision to keep or reject an outlier may result from subjective judgement
- A better approach is to apply small weights to outliers rather than rejecting them,
  - But we can still reject completely wrong observations.
- So a better alternative is to
  - A more considered approach to outliers
  - Or try robust or resistant regression

A Structured Approach to Deal with Outliers

- Check the measurement process and environment for defects
- Check the outlier point for unusual features, e.g. has it unusually high or low values
  - Example; in a sample of people you may find one individual who displays unusual features (e.g. high blood pressure). On examination you may find they are very old or very young
- If, after analysis, you still find that this a data point that belongs in the population of interest, then you may
  - Discard it if it has high influence in terms of determining the various parameter estimates $(\hat{\alpha}, \hat{\beta})$ where $y = \alpha + \beta * x$
  - Otherwise leave it in the data set, but consider using a robust or resistant regression technique to minimize it’s impact

We must also consider Influential Observations

- **An influential observation** is an observation in linear regression analysis whose deletion from the dataset would noticeably change the results of the regression
- Can be detected visually
- Also by calculation of Cook's distance D_i which measures the effect of deleting the ith observation.
- Large values of D indicate high influence
- When we discover highly influential observations then we must be assured that these are genuine
- Influential observations are often (but not always) outliers
- Once again we may consider using a robust or resistant regression technique



<u>Cooks Distance</u>

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\ \mathrm{MSE}},$$

$\hat{Y}_j$ is the prediction from the full regression model for observation *j*;

$\hat{Y}_{j(i)}$ is the prediction for observation *j* from a refitted regression model in which observation *i* has been omitted;

p is the number of fitted parameters in the model;

MSE is the mean square error of the regression model.

Robust & Resistant Regression Methods

- Robust Regression
  - Robust regression is an alternative to least squares regression when the assumptions of linear regression are violated
    - They are not affected by small deviations from the assumptions of the model
  - Most resistant estimators are also robust in relation to the assumption about normally distributed residuals
- Resistant Regression
  - These are regression Methods where outliers have minimal impact of the fitted model
    - They are not affected by small errors or changes in the sample data
    - Most resistant estimators are also robust in relation to the assumption about normally distributed residuals

Overview of Robust Regression

- These techniques were introduced in the 1960s
- Robust regression techniques are used as an alternative to Least Square method when errors are non-normal
- They can also be used to confirm the appropriateness of the ordinary least square model.
- In OLS Outliers affect the estimates of
  - Parameters
  - Standard errors (standard deviation of parameters)
  - Coefficient of determination
  - Test statistics
  - And many other statistics
- Robust regression tries to protect against this by giving less weight to such cases, not by excluding them
- Primarily helpful in finding cases that are outlying with respect to their y values (many outliers in long tails)
- Also helps in dealing with unusual X-values that have great leverage and hence cause problems
  - But robust methods can't overcome problems caused by complex variance structure.
- And it is more complex to evaluate the precision (confidence in) of the regression coefficients, compared to the least squares model.

Robust Regression Fitting Techniques

- Fitting is done by iterated re-weighted least squares (IWLS)
- IWLS (IRLS) uses weights based on how far outlying a data point is, as measured by the residual for that data point.
  - Weights vary inversely with size of the residual
  - We begin by drawing a best fit line, and then calculating the residuals from each data point to the line
  - We then redraw the line using the weights described above in an IWLS model
  - Continue iteration until process converges(i.e. you converge to a stable solution)
  - NOTE : sometimes this will cause you to have different end estimates , depending on which line you started from

A Popular Robust Technique: M-estimators

- Remember the classical regression model is obtained by selecting the least squares of deviations, i.e.

  Given the equation $y = \alpha + \beta * x + \varepsilon$, and observations (y_i, x_i)

  Where ε_i are the residuals from each pair of observations y_i, x_i

  We choose estimates of α and β such that the mean square sum of errors $\sum \varepsilon^2_i$ is minimized

- The basis of M-estimation (Maximum Likelihood Type) is a generalisation of least squares
  - We have a function of the errors $\rho(\varepsilon_i) = \rho(y_i - \hat{y}_i)$
  - The general M-estimator minimizes the objective function $\sum \rho(\varepsilon_i)$
  - If $\rho(\varepsilon_i) = \varepsilon^2_i$ then we have the least squares approach
- The function ρ can be chosen in such a way to provide the estimator desirable properties (in terms of bias and efficiency) when the data are truly from the assumed distribution, and 'not bad' behaviour when the data are generated from a model that is, in some sense, *close* to the assumed distribution.
- The Huber M-estimator is obtained by setting $\rho(\varepsilon_i) = w(\varepsilon_i) * \varepsilon_i$

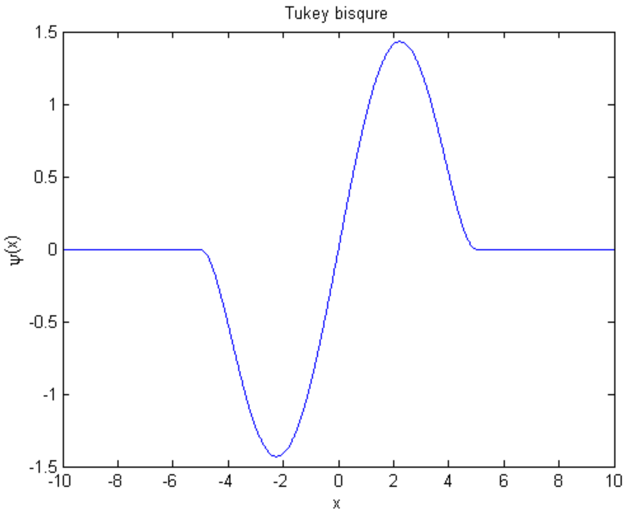
  where $w(\varepsilon_i)$ is a weight and a function of ε_i
  - In general the weights w_i are smaller when the residual ε_i is large *(Not dissimilar to the IWLS approach)*

The Bi-square Estimator

- Another M-Estimator is the Bi-square Estimator
- Tukey proposed an M-estimator that has the following ρ(x) weighting function

$$\rho(\varepsilon)=\varepsilon(1-(\varepsilon/k)^2)^2 \quad \text{when} |\varepsilon| \leq k$$

- This function is plotted in the below figure for k=5.



"Tukey" by Osnatgp - Own work. Licensed under Public Domain via Common https://commons.wikimedia.org/wiki/File:Tukey.png#/media/File:Tukey.png

Overview of Resistant Regression Methods

- Resistant Regression techniques are similar to robust techniques , but are model-based. , i.e. their answer is always the same.
- They reject all possible outliers, and so can be useful to detect outlier
- But the are inefficient, as they are only taking into account a portion of the data (they have taken away the data points that are outliers)
- They are more resistant to outliers than robust methods
- Two common resistant regression types :
  - Least Trimmed Squares (LTS)
  - Least Median of Squares (LMS)

Resistant Regression Techniques: Least Trimmed Squares (LTS)

- The LTS Method is similar in some ways to Hubers M-estimating method
- Instead of minimising the mean square sum of errors $\sum\varepsilon^2_i$ for all n points in your data set
  - The LTS method attempts to minimise the sum of squared residuals over a subset S_k of k of those points.
    - ⇨ The n-k points which are not used do not influence the fit.
- Whether or not a point belongs to S_k depends on the size of the residual ε_k
- We decide what the value k should be (e.g for 100 observations it may be 90)
- We then fit a least regression model and calculate the absolute value of each ε_k
- We then order these and work out what the kth highest value of ε_k is
- Then all those points which have residuals less than this value are in our set
  - Those which have residuals higher than this value are removed
- We then re-fit a least regression model to these data points
  - This is our least trimmed square regression model

Resistant Regression Techniques: Least Median of Squares (LMS)

- The least-median-of-squares (LMedS) method estimates the parameters by minimizing the median over all i of $|\varepsilon_i|$ or ε_i^2
  - i.e the values of the estimates (α, β) , must give the smallest value for the median of squared residuals computed for the entire data set.
- This replaces minimizing the sum in the Least Square Model method with median.
- Does not work very well for small samples, but is resistant to outliers

Other Approaches: Method of least absolute deviation (L1 norm)

- Least Squares seeks to find the parameters b_0 , b_1….. that minimise $\sum(y_i - \hat{y}_i)^2$

  where $\hat{y}_i = b_0 + b_1 * x_1 + \ldots\ldots.$

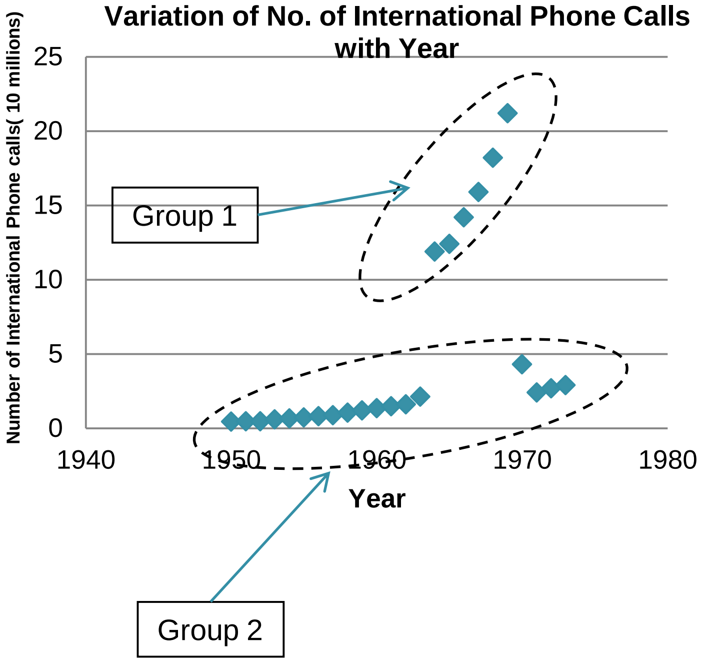
- Instead we seek estimated values of the unknown parameters b_0 , b_1….. that minimize the sum of the absolute values of the residuals:

  $\sum |(\hat{y}_i - y_i)|$

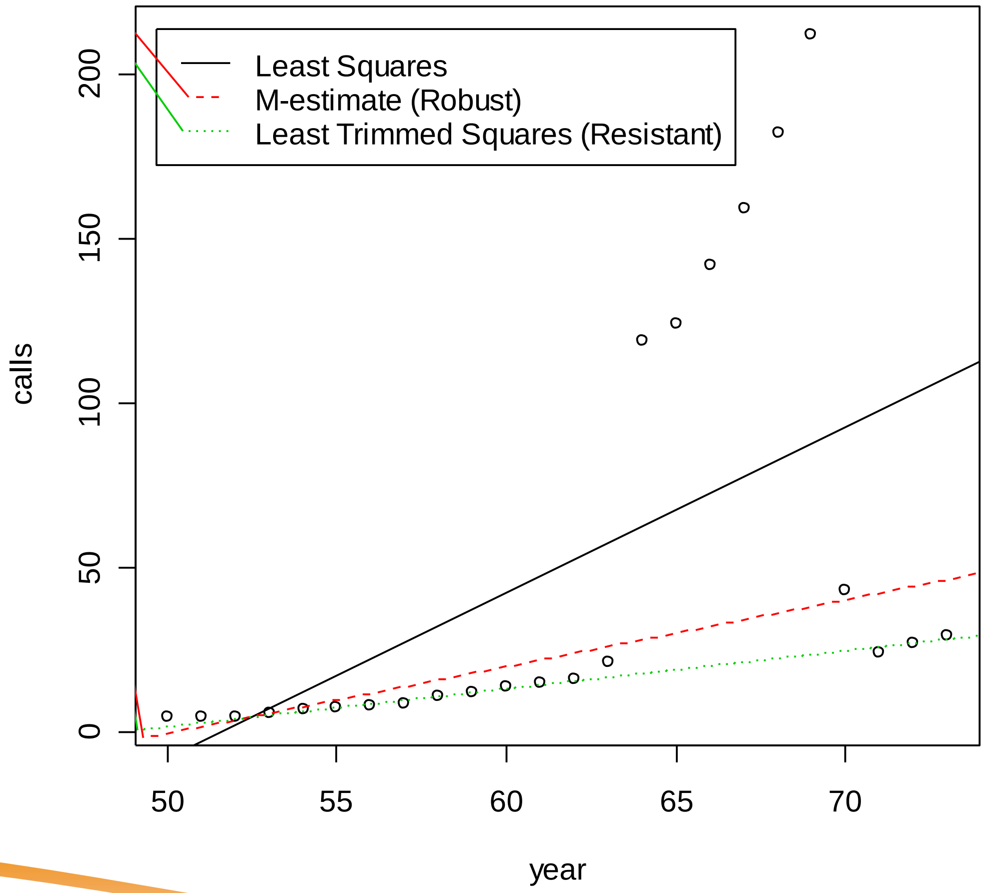
- Not so easy to calculate as least squares , but can be solved using linear programming
  - May get multiple possible solutions
- The model created is less sensitive to outliers than least squares

Alternative Approach: Group the data based on Doman knowledge and create more than one model

- Given a set of heterogeneous data, we could also
  - Split the data into groups
  - Fit different models for groups
- So in the case of the telephone data
  - We could build models for two groups (although the model for length of call may not be interesting to us)
- In general, we can break the data into groups using
  - Clustering
  - Classification Trees
  - Prior knowledge
  - Visual analysis
- However usually this approach must be driven and supported by expert opinion



Phones data with Least Squares, Robust, and Resistant regression lines



Robust and Resistant Methods in Practice

- Robust and resistant methods are often confused and mixed together
- Both robust and resistant methods requires much more computing resources than least squares
- Robust and resistant methods tend to underestimate the errors and uncertainties associated with estimates

Fitting Robust/Resistant models

1. Begin by conducting OLS regression
2. Then identify outliers; calculate residual vales and Cooks distance
   - Also look to see if distributional assumptions of OLS are violated
3. Decide whether to simply eliminate them and redo OLS regression
   - Seek domain knowledge
4. If you don’t think elimination is a good strategy

   ⇨ You should use several models and plot each model results on a graph and see which seems “most reasonable”

Reference

- https://stats.idre.ucla.edu/stata/dae/robust-regression/
- https://stats.idre.ucla.edu/r/dae/robust-regression/
- http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch13.pdf
- Outlier Analysis by Charu C Agarwal, Springer, 2013

Conclusions

- There are many, many robust and resistant regression techniques
- They were not so popular as they were not terribly well supported by analytical software packages. But situation has improved now
- You should consider using them where
  - There are outliers in the data that well greatly influence the application of the OLS regression model
  - There is reasonable evidence that the assumptions of the OLS regression model are not true
- You may use several robust/resistant regression techniques and see which is most "reasonable"
- Other situations may call for the use of alternative techniques

The End