Influential points and outliers
Variable selection

GR 5205 / GU 4205
Section 3

Columbia University
Xiaofei Shi

Outliers and influential points

- An outlier is a point with large residuals.

  An influential point is a point that has a large impact on the regression.

- They do not mean the same thing

- Four famous datasets due to Frank Anscombe.

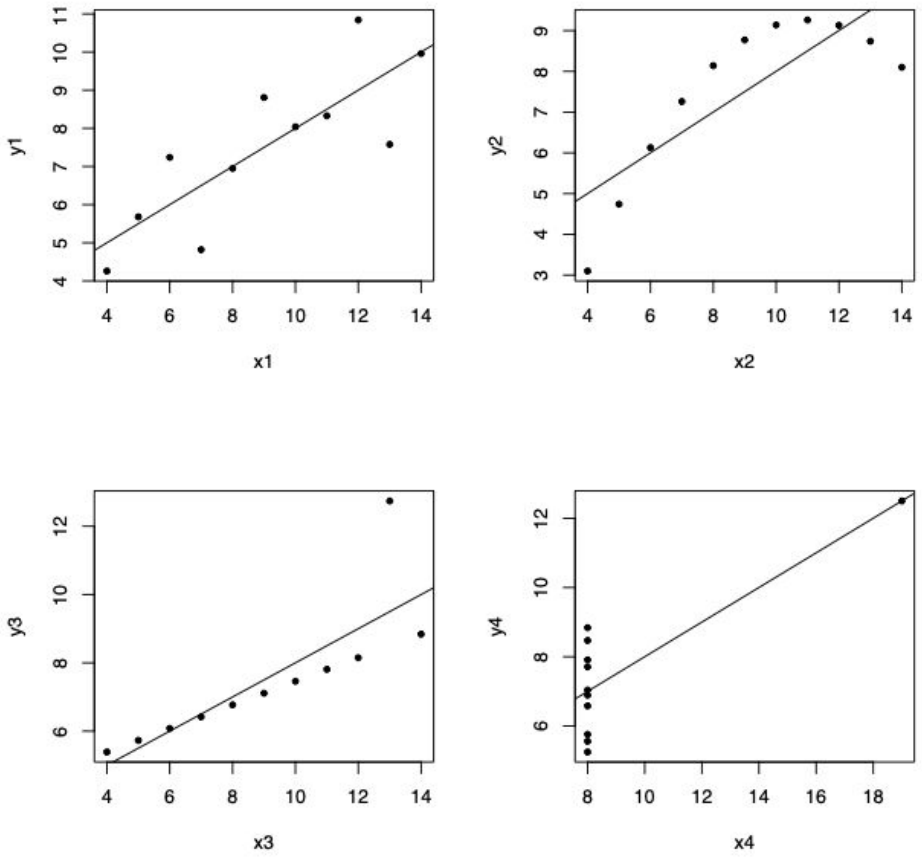


FIGURE 1: *For data sets that have the same fitted line. Top left: no problems. Top right: a non-linear pattern. Bottom left: An outlier. Bottom right: an influential point.*

Modified residuals

- Let e be the n-dim vector of residuals. Recall that

$$e = (I - x(x^\top x)^{-1}x^\top)Y, \quad \mathbb{E}[e] = 0, \quad \mathrm{Var}[e] = \sigma^2 \left(I - x(x^\top x)^{-1}x^\top\right)$$

- Define $H := x(x^\top x)^{-1}x^\top$

  Then the standard error of the i-th residual is $\widehat{\sigma}\sqrt{1-h_{ii}}$

- Standardized residuals

$$r_i = \frac{e_i}{\widehat{\sigma}\sqrt{1-h_{ii}}}$$

Studentized residuals

$$t_i = r_i\sqrt{\frac{n-p-2}{n-p-1-r_i^2}} = \frac{e_i}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}$$

Influence $\hat{Y} = HY$

- A linear combination of elements of H
- In particular, h_{ii} is the contribution of the i-th data point to the i-th estimation;

  We call h_{ii} the leverage.
- We want to know how influential the i-th data point could be.

Leave-one-out

- The Cook's distance $D_i = \frac{\left\| \widehat{Y} - \widehat{Y}^{(-i)} \right\|^2}{p\widehat{\sigma}^2}$
- Leave-one-out estimation $\widehat{\beta}^{(-i)} = \widehat{\beta} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i^T e_i}{1 - h_{ii}}.$

Diagnostics

- We can look at their leverage, which depends only on the value of the predictors.
- We can look at their studentized residuals, either ordinary or cross-validated, which depend on how far they are from the regression line.
- We can look at their Cook's statistics, which say how much removing each point shifts all the fitted values; it depends on the product of leverage and residuals.

Leverage
0.005
0.010
0.015
0
200
400
600
Index

Standardized residuals
−4
−2
0
2
4
6
8
0
200
400
600
Index

Cross−validated studentized residuals
−4
−2
0
2
4
6
8
0
200
400
600
Index

Cook's statistic
0.00
0.04
0.08
0.12
0
200
400
600
Index

Dealing with outliers

- Deletion

  - the data point is wrong and there is no way to fix it.

  - the data point isn’t wrong, exactly, but belongs to a different phenomenon or population from the one you’re studying.

  - the data point looks really weird compared to all the others.

Dealing with outliers

- Deletion

  - the data point is wrong and there is no way to fix it.

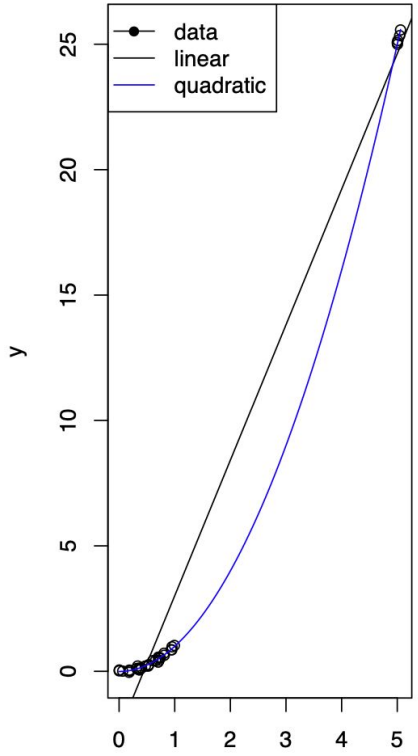
  - the data point isn’t wrong, exactly, but belongs to a different phenomenon or population from the one you’re studying.

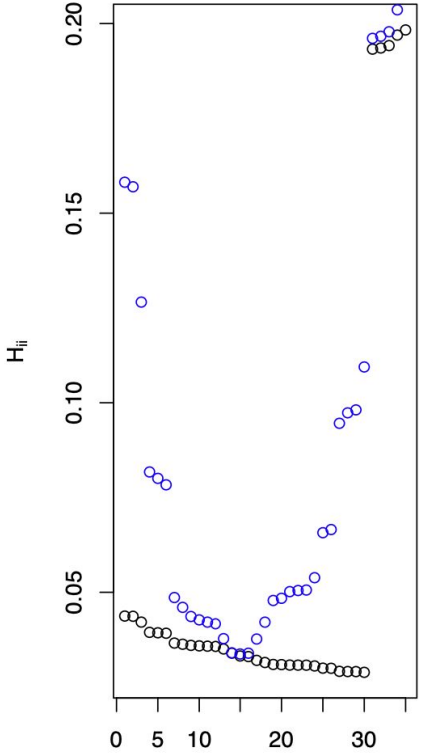
  - the data point looks really weird compared to all the others.

  A counter example: the Mpemba’s effect

Dealing with outliers

- Changing the model





Dealing with outliers

- Robust linear regression $\tilde{\beta} = \underset{\mathbf{b}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i \mathbf{b}).$

  - Different choice of the loss function yields different estimators

  - Huber loss: in between linear and quadratic

Variable Selection

- Which variables and functions of variables should we include in our model.

- Why do we need variable selection:
  - have simpler model
  - avoid singular design matrix
  - bias-variance trade-off
- Generally speaking, small models (with few predictors) have low variance and high bias. Large models (with many predictors) have high variance and low bias. The challenge in variable selection is to choose a model with small prediction error and this requires that we balance the bias and variance.

Using p-value is not a good idea

- It is very tempting, and common, to use the p-values which come from this test to select variables: significant variables get included, insignificant ones do not, ones with smaller p-values (hence larger test statistics) are higher priorities to include than ones with smaller test statistics.
- But this might be a bad idea:
  - Larger coefficients will have larger test statistics and be more significant
  - Reducing the noise around the regression line will increase all the test statistics, and make every variable more significant
  - Increase the sample size will increase all the test statistics, and make every variable more significant
  - More correlation between the current and the other predictors will, all else being equal, decrease the test statistic and make the variable less significant

Using p-value is not a good idea

- Can we reliably detect that this coefficient isn't exactly zero?

  - the test statistic, and thus the p-value, runs together an estimate of the actual size of the coefficient with how well we can measure that particular coefficient.

  - Every variable whose coefficient isn't exactly zero will eventually (as $n \to \infty$) have an arbitrarily large test statistic, and an arbitrarily small p-value.

- None of this is even much help in answering the question "Which variables help us predict the response?" F-tests on groups of coefficients don't help either. t-tests on individual coefficients.

Which variables help predict the response

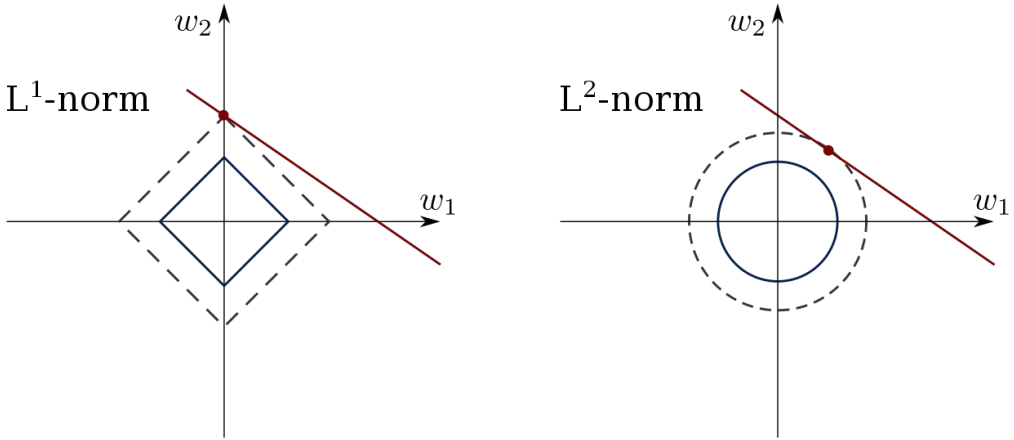
- Usually first standardize the predictors and the response variables, so we are comparing variables on a similar scale
- When the number of variables is not too large:

  - Cross-validation, AIC, Cp
- When the number of variables is large:

  - Stepwise method

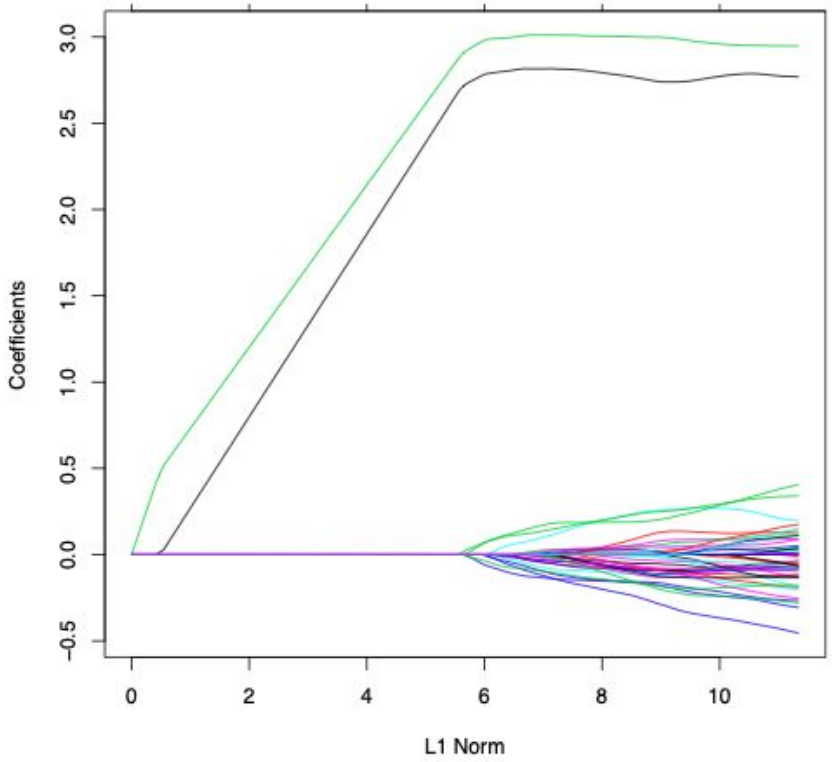
  - Lasso

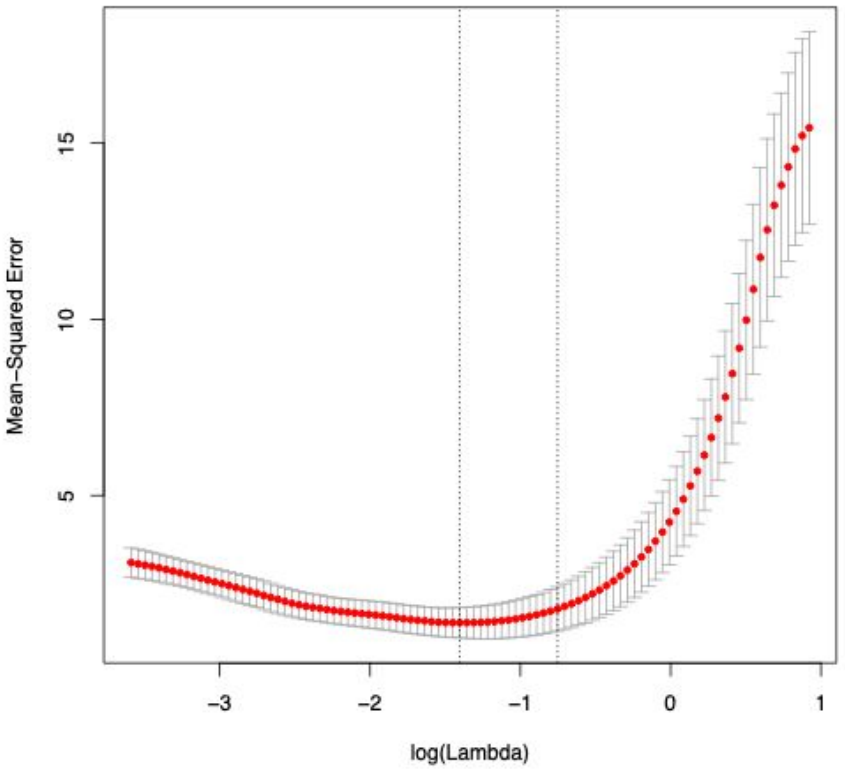
Lasso



- http://statweb.stanford.edu/~tibs/lasso.html
- Implement:
  - R:
  - Python:

Tuning of the hyperparameters





Inference after selection

- If all you care about is prediction, you may not need to do hypothesis testing or confidence intervals.
- The standard inferential statistics (like the p-values on individual coefficients) are not valid if you do variable selection. The easy cure is to split the data in half at random, and use one part to do model selection and the other half to do inference for your selected model.