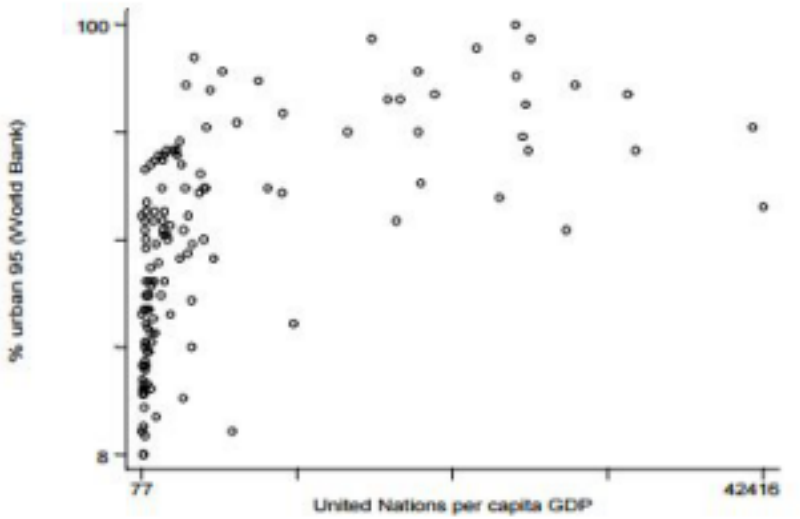
**CS 403/725**
**Tutorial - 1**

(1) What will be the effect on the solution of Least square analysis if we apply the following transformations on the training set:
   (a) Add a real number k to the output value of each datapoint.
   (b) Multiply by k the output value of each datapoint.
   (c) Rotate all data points by a fixed angle.

(2) Consider the regression of % urban population (1995) on per capita GNP:




   a) Can you fit a line through this data?
   b) What is the transformation you would do to apply the concepts of linear regression on such data points.

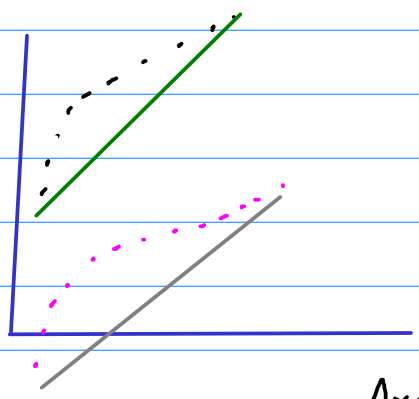
(3) Problems with least square regression

Least squares regression can perform very badly when some points in the training data have excessively large or small values for the dependent variable compared to the rest of the training data. The reason for this is that since the least squares method is concerned with minimizing the sum of the squared error, any training point that has a dependent value that differs a lot from the rest of the data will have a disproportionately large effect on the resulting constants that are being solved for.

Problem ① ⓐ $y_j^{new} = y_j + k$.

Argument: $w^{new}$ s.t $(w^{new})^T \phi(x_j) = w^T \phi(x) + k$

Why?



(Assume b is part of w)

$$w = \underset{\hat{w}}{\operatorname{argmin}} \| \phi \hat{w} - y \|^2 \quad (1)$$

$$w^{new} = \underset{\tilde{w}}{\operatorname{argmin}} \| \phi \tilde{w} - y - k \|^2 \quad (2)$$

Argument: Let $w^{new}$ be optimal soln to (2)

Then $w$ s.t $\phi w^{new} - y - k$ minimizing (2) ⇒

$(\phi w^{new} - k) - y$ minimizes (1)

(1) (a) Suppose $\tilde{w}^*$ was optimal soln after the shift & $\tilde{E}^*$ the optimal value of error, our claim is

$\tilde{w}_0^* = w_0^* + k$ & $\tilde{w}_j^* = w_j^*$ should be the

$\phi_0(x_i) = 1$ case

Claim: If not, I could present a better solution to the orginal problem (or the new problem)

New error at old optimum shifted $= \sum_j \left[ \left( \sum_i (\tilde{w}_i^* \phi_i(x_j) - y_j) \right) + (k-k) \right]^2 = E^*$

Consider an Example:

Suppose we would like to predict height of a person based on his weight and age. Fig 2 shows the linear regression line (hyperplane) through the data.

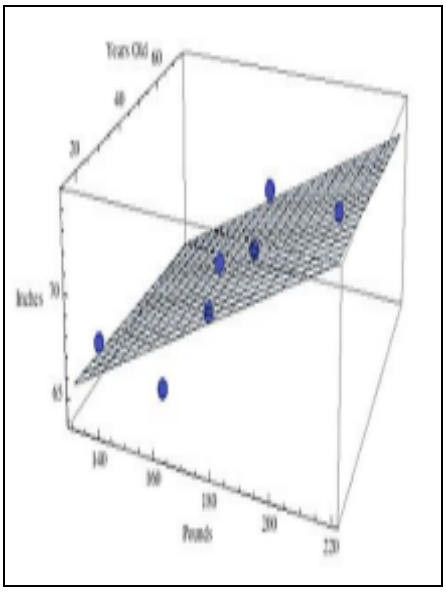



**Fig. 2**

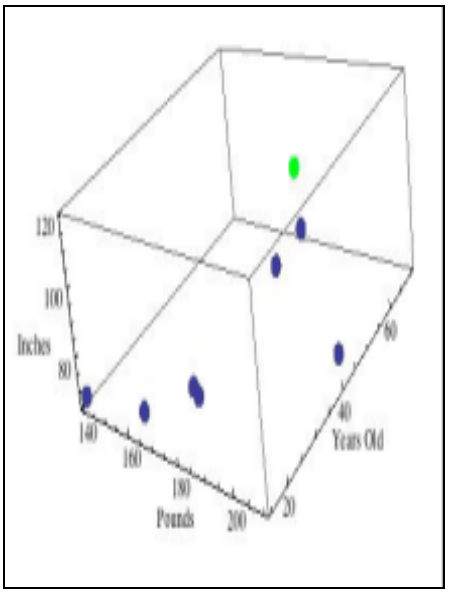
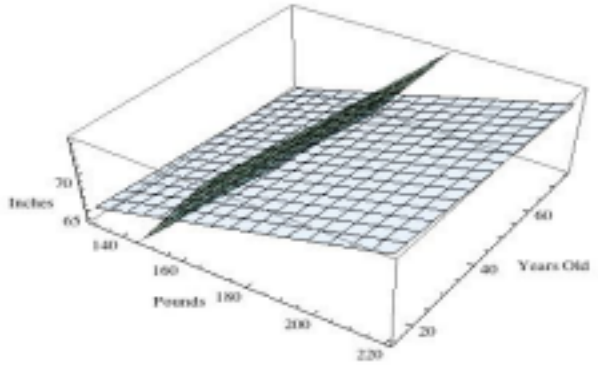



**Fig 3**

Now if we have an outlier i.e. a 10 foot tall 40 year old who weighs 200 pounds man (shown as green) in our original data the figure would look like fig 3.

Below we have a plot of the old least squares solution (in blue) prior to adding the outlier point to our training set, and the new least squares solution (in green) which is attained after the outlier is added:



General idea: $\min_w E(f(w), y) + \lambda \Omega(w)$ → Regularization penalty model complexity such as $\frac{1}{2}\|w\|_2^2$ or $\|w\|_1$

As you can see in the image above, the outlier we added dramatically distorts the least squares solution and hence will lead to much less accurate predictions.

Suggest some methods to improve the optimization function in the linear regression to work around this problem.

(Hint: It should be noted that bad outliers can sometimes lead to excessively large regression constants we would like to fix this problem)
[src: clockbackward.com]

(Spirit behind boosting algos)

eg: Iteratively wted regression: Diagonal $M = \begin{bmatrix} r_1 & 0 \\ 0 & \ddots \; r_m \end{bmatrix}$ $r_i = 1$ initially

Iteratively solve $w^* = \arg\min_w \|M\phi w - y\|^2$

(4) We propose a modification to the least square regression formulation: Instead of taking the squared sum of the error values, we will instead raise their absolute value to the power of p. p is a parameter which we can tune.

$r_j^{new} = |(f(x_j) - y_j)|^2$

a. For least squares, p = 2. If we change p what problems do you think we can run into? Higher $p \Rightarrow$ more sensitive to outliers
b. How will you implement the method for a general p? Can you find a closed formula? Gradient descent etc (to come)
c. Take-home question: It is observed that if p < 2 then the method tends to be more robust to outliers. Can you think of an experiment to test this?

[Objective is differentiable]

$1 < p < 2$ gives sparsity yet preserving convexity (and is differentiable). $p=1$ makes problem hard

**Problem 5.** In the class, we illustrated the problem of overfitting by supplying training data to the Applet `http://mste.illinois.edu/users/exner/java.f/leastsquares/#simulation` and plotting polynomial and printing its coefficients for increasing polynomial degree. As an exercise, try this out yourselves either with the same or another applet or writing code in R/scilab. Share your observations. What, in your opinion, leads to overfitting?

**Problem 6.** In the class, we presented the solution to the least squares linear regression problem

$$\mathbf{w} = (\phi^{\mathbf{T}}\phi)^{-1}\phi^{T}\mathbf{y}$$

and expressed that if $\phi$ is full column rank, $\phi^T\phi$ will be invertible.

filtering → pairwise correlation

- When is $\phi$ not full column rank? What are associated problems and fixes? (Redundant features) $\phi^T\phi$ may not be invertible
- How can we find a solution if $\phi$ is not full column rank?
- When will $\phi^T\phi$ be positive (semi) definite? What is the relationship between positive definiteness and invertibility?

$(\phi^T\phi)w^* = \phi^T y$ ←

Solve the linear system using Gauss elimination ==> Multiple solutions might exist!

$(\phi^T\phi)w = (\phi^T y)$ $\quad [Ax = b]$

Given $\phi^T\phi$ is not full rank, when with $\phi^T y$ NOT belong to column space of $\phi^T\phi$?