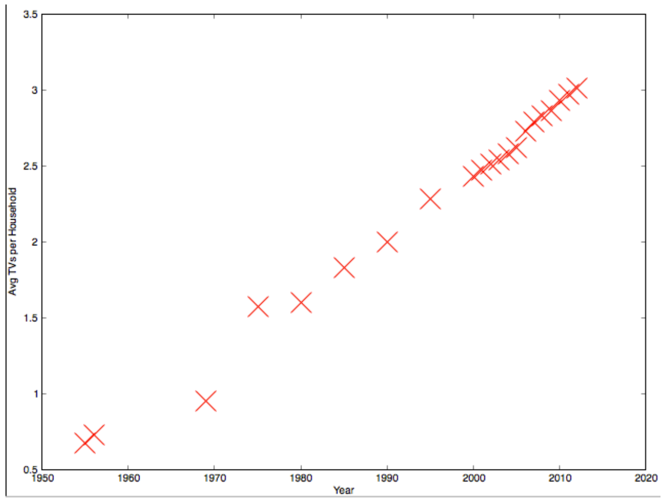
Linear Regression Machine Learning – Television Sets Over Time

At what rate have American households been buying television sets? Using raw data I found for television audiences since 1955, I will train an unsupervised learning algorithm to answer the question.

The null hypothesis is that there is no statistically significant trend in the rate of increase in television sets per household. First, plotting the data:



We now train our linear regression model to find a line of best fit, which we'll refer to as our hypothesis. Formally, the hypothesis is $h_\theta(x) = \theta_0 + \theta_1 x$, which is the equation of a line. It can be reformulated into just 1 term: $h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 = \theta^T x$ such that $\begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix} \mid \theta \in \mathbb{R}^2$

We will have the algorithm create many lines of best fit and then choose the most correct one. That is, choose the line that minimizes MSE (Mean Squared Error) - the distance between the points and the line. Where m is the number of training examples or, put simply, (x,y) coordinates in our data, the function calculating the MSE can be formalized to:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Its vectorized form for MATLAB use is:

$$J(\theta) = \frac{1}{2m}(X\theta - y)^T(X\theta - y)$$

Such that

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} \quad X = \begin{bmatrix} —(x^{(1)})^T— \\ —(x^{(2)})^T— \\ \vdots \\ —(x^{(m)})^T— \end{bmatrix}$$

We need to fill in θ_0 and θ_1 in the hypothesis, which are the intercept and slope. To find the error-minimizing line of best fit, we minimize for theta.

We want our algorithm to adjust θ values but it won't know when to stop so we'll specify, say numIterations=1500 steps. You will notice that the update θ := θ - $\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2\, x^{(i)}$ will reduce or increase theta to a magnitude equivalent to the slope of J(θ) at point x_θ. However, this may overshoot past the minimum and our algorithm will be useless… We must, then, introduce a constant α like 0.01 that is our learning rate. As a rule of thumb, 0.001 <= α <= 10. We'll have to keep our eye on the learning rate to see if 0.01 is appropriate but more on this later. Anyway, our θ will gradually descend to the minimum of cost function. Simultaneously updating for our 2 variables, our pseudo-code is:

```
temp0 := $\theta_0 - \frac{\alpha}{m}\sum_{i=1}^{m}(h_{\theta 0}(x^{(i)}) - y^{(i)})^2$
temp1 := $\theta_1 - \frac{\alpha}{m}\sum_{i=1}^{m}(h_{\theta 1}(x^{(i)}) - y^{(i)})^2\, x^{(i)}$
θ := temp0
θ := temp1
```

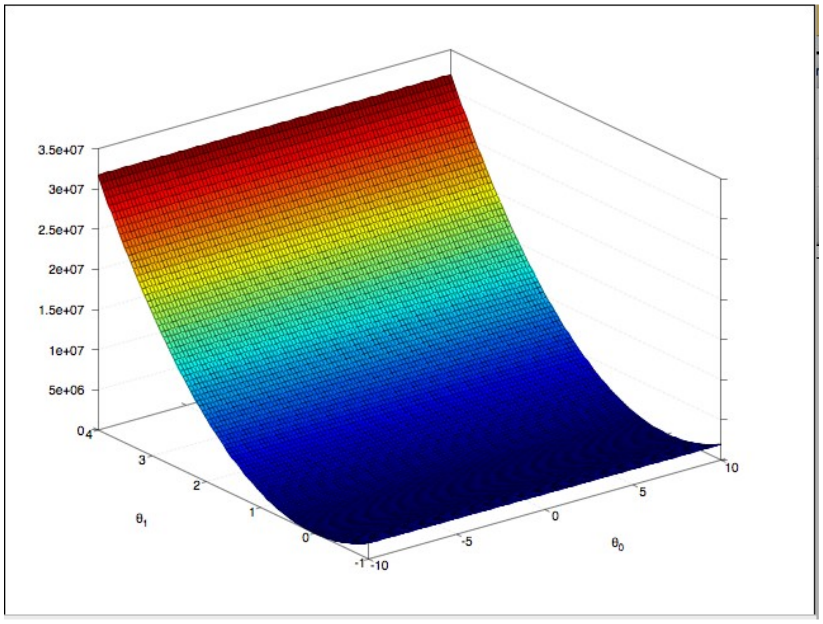
For MATLAB,

```
X = [ones(m,1), TVdata(:,2)];               % ℝm x 2  where first column is 1
m=length(TVdata (:,2));                     %m=# of training examples
prediction=X*theta;
J = (1/(2*m))*sum((prediction- y).^2)       %compute the cost to fit data in X and y
                                              for plotting
for 1:numIterations                         %update theta
        theta = theta – (alpha/m)(prediction – y)*(X')
                                            %X' is X transpose, so we can multiply it
                                              with the matrix X*theta – y
                                            %Recall: this is the derivative of the cost
                                              function
```

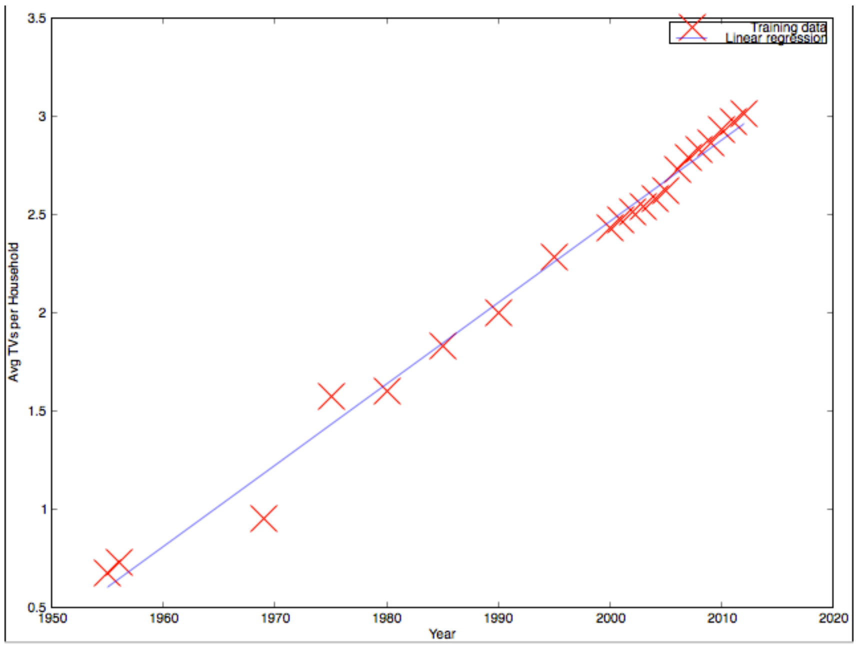
Interestingly, the theta values in fact diverge to infinity. Let us look at our feature scaling… The range of the TV values is 0 to ~3. So, this shouldn't be an issue. Therefore, our α has overshot past the minimum. Instead of playing with α, we will opt for the analytic method of finding θ:

$$\theta = \text{pinv}(X^T X * X^T y)$$

This gives us the J value 2.7530 and the θ values -80.262851, 0.041363. We can generate a x-y-z plot where the z-axis is J(θ), x is θ_0 and y is θ_1.



Using these values, it seems American households buy 0.04 TV sets per year on average.


Training data
Linear regression
Avg TVs per Household
Year