Multiple Linear Regression

Authors: Yingjie Chen, Dawson Cook, Priyanka Choudhary. (PDF)

MLR is one of the multiple ways to model the data at hand. The expected answer for the dependent variable is expressed as y(hat) and the parameters (betas) are found out by the model that explain the fit of independent variables with the y. In a perfect model, the true Y is assumed to be generated using the MLR model along with an added zero mean gaussian noise. This concept also was reflected in one of our quiz questions that we later discussed in class. If we have less data points, we would want them to be distributed along along a wide range to get a good MLR fit.

We assume the follwing in each MLR model: 1. Y|X is linear in X: All the terms are linear in X, there is no higher order term. However, we can fit an MLR model to an equation which has higher order terms by assuming them to be in the higher order variable axis.

2. The error term is normally distributed and they are all independent of each other. The error term is a reflection of the unexplained portion of the true Y which the model cannot explain. The variance and mean of all error terms is constant (0 and sigma squared)

Given a set of Xs, the Y is a gaussian with the mean as the $\beta^T X$ and the variance is sigma squared.

The cost funtion for the MLR/OLS is the mean squared error which we try to minimise to find the optimal fit.

It's a property of gaussian distributions that an MLR with gaussian noise gives the same solution as the maximum likelihood estimate, which is why we use gaussian distributions to model the noise in the dataset.

MLR results

Model: $Sales_i = \beta_0 + \beta_1 P1_i + \beta_2 P2_i + \epsilon_i,\ \epsilon \sim N(0, \sigma^2)$

Regression Statistics	
Multiple R	0.99
R Square	0.99
Adjusted R Square	0.99
Standard Error	28.42
Observations	100.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.00	6004047.24	3002023.62	3717.29	0.00
Residual	97.00	78335.60	807.58		
Total	99.00	6082382.84			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	115.72	8.55	13.54	0.00	98.75	132.68
p1	-97.66	2.67	-36.60	0.00	-102.95	-92.36
p2	108.80	1.41	77.20	0.00	106.00	111.60

Why use a R Square?

Unfortunately, the data one works with typically will be biased. This will result in the R Square being optimistic. The true accuracy actually could be lower, but not any higher. This is where the Adjusted R Square comes into play. Adjusted R Square better reflects what you would expect in the future when you apply a model. The adjusted R-squared compensates for the addition of variables and will decrease when a predictor improves the model less than what is predicted by chance. So as you add more variables, the R Square might improve, but it is also making your model more complex. Adjusted R Square solves this problem by taking into account complexity. Resulting in a more accurate value. Since this problem is very simple with few parameters, the Adjusted R Square is too small to see. It typically is smaller than the R Square.

Standard Error:

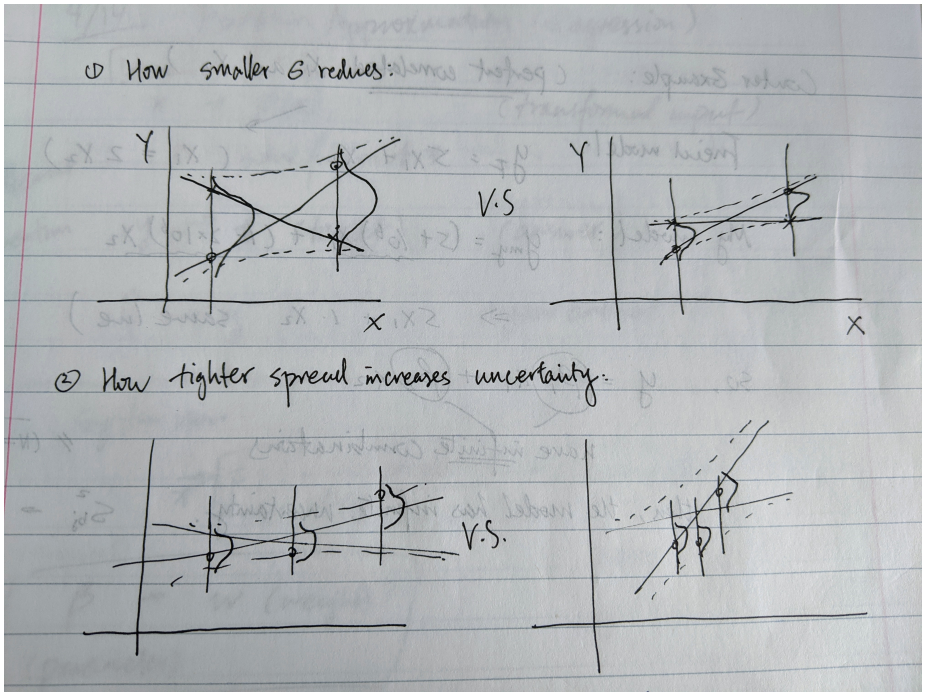
Horizontal error is indicating how much noise you estimate is in the output, and can be classified as σ or the RMSE. The vertical error is the uncertainties associated with each of your betas or the coefficients. So what this means is that there is a 95% chance that the coefficients will be two times plus or minus their respected vertical error. Each vertical standard error is the standard deviation associated with that estimate. In a perfect world you would want the standard error to be zero, but in reality you just want it as low as possible. The higher it is, the more uncertainty you have. This all relates to the collinearity problem. Collinearity is a condition in which some of the independent variables are highly correlated. If this is the case, your vertical standard error will be very high. To account for this problem, you should try and access more data or lower your Horizontal error(less noisy).

Ways to Reduce Uncertainty in the Model:

$s^2_{b_j} = \frac{s^2}{(N-1)*(\text{Variation in } X_j, \text{not associated with other X's})}$

Note: s_{b_j} is the standard error for b_j, so our goal is to lower the value on the LHS: $s^2_{b_j}$

(1) Lower σ, less noise: Note the s^2 term in the numerator of the equation above is the σ^2 in target (noise). So lowering this particular term will reduce the error (as shown below).



(2) More data available: The N-1 term in the denominator of the equation represents the number of data available to the model. Increasing N will in terms reduce $s^2_{b_j}$

(3) More spread: If X's are really close to each other, the predictions for points outside that tight spread will have a increasing uncertainty and error. (as shown in scratch above)

(4) Reduce multicollinearity: Collinearilty will in fact reduce (inflate) this variation term Variation in X_j not associated with other X's in the equation, resulting in higher $s^2_{b_j}$.

For example, if given "height", then one can guess "weight" to a certain extend. Therefore, the variation in weight not associated with other X's will decrease because it's closely associated with "height". It increase the $s^2_{b_j}$ term.

Counter Example (with perfectly correlated X1 and X2):

Consider the case with perfectly correlated X1 and X2, such that $X_1 = 2 * X_2$

My friend has the model: $y_{Friend} = 5 * X_1 + X_2$ and I have the model: $y_{Me} = (5 + 10^6) * X_1 + (1 - 2 * 10^6)X_2$

With some simple algebra, one can tell that they are essentially the same, but that also implies that in $y = \beta_1 * X_1 + \beta_2 * X_2$, β_1 and β_2 have infinite combinations. Therefore, the model has infinite uncertainty. (One can also mathmatically tell that in the equation, the term "Variation in X_j not associated with other X's goes to 0, making $s^2_{b_j} \to \infty$).