

Intro/Recap of Multiple Linear Regression (MLR)

AKA Ordinary Least Squares
See LinearRegression in scikit-learn

The MLR Model

Note: I will use typical **statistics notation**:

coefficients are called β s, the dependent variable is Y, and estimates are indicated by “hats”.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Key General Issues

- ▶ What are the model assumptions? Are they valid?
- ▶ How do you estimate the parameters from data?
 - (i) cost function (true or surrogate?)
 - (ii) optimization method
- ▶ How do you evaluate your model?
 - (i) training/validation/test/scoring error
 - (ii) performance measures

Assumptions behind the MLR Model

- (i) The conditional mean of Y is linear in the X_j variables.
- (ii) The error term (deviations from line)
 - ▶ are normally distributed
 - ▶ independent from each other
 - ▶ identically distributed (i.e., they have constant variance)

$$Y|X_1 \dots X_p \sim N(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p, \sigma^2)$$

Then minimizing Mean Squared Error (MSE) on the training data yields the Maximum Likelihood Estimate (MLE) solution of the assumed *generative model*.

Q: What do the β s mean?

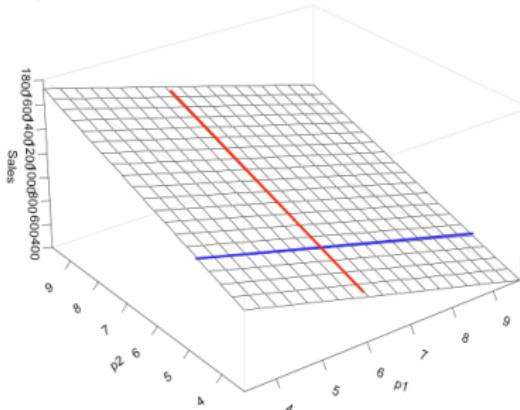
See Bishop page
140 for “why”

MLR On Sales Data

Consider sales of a product as predicted by price of this product (**P1**) and the price of a competing product (**P2**).

$$\text{Sales} = \beta_0 + \beta_1 P1 + \beta_2 P2 + \epsilon ; \text{ Thus } \beta_j = \frac{\partial E[Y|X_1, \dots, X_p]}{\partial X_j}$$

Holding all other variables constant, β_j is the average change in Y per unit change in X_j .



Q: Will your sales go up if you reduce the price?

Least Squares

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

$$b_0 = \hat{\beta}_0 = 115.72, b_1 = \hat{\beta}_1 = -97.66, b_2 = \hat{\beta}_2 = 108.80,$$

$$s = \hat{\sigma} = 28.42$$

Note that, for simple linear regression solution using OLS,

$$R^2 = \text{corr}(Y, \hat{Y})^2$$

Regression Summary Output

(Multiple R) - Absolute value of the correlation coefficient.

Values range from 0 to 1. Zero means that X & Y aren't correlated at all, 1 means that the variables are 100% correlated. Since it's not positive or negative (because of absolute value), it only tells us the strength of the relationship, not the direction.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

(R Squared) - Tells us how much of the variance in our predicted variable (Y) is explained by our independent variables (X) (in this case, 83%). Which means that 17% of the variance is explained by other factors. Want this to be high!

(Adjusted R Squared) - Used when analyzing multiple regression output (when you have more than one X). When we have more than one independent variable, the computation process inflates the R-squared, so this number is adjusted for that inflation.

(RMSE aka Root Mean Squared Error) - Measured in units of Y (eg \$, units sold), it gives you an average error amount. In this case, the RMSE is around 14. Since that represents 1 standard deviation, values within that range account for 68% of the population. Two standard deviations ($2 \times 14 = 28$) would cover 95% of values. So we can say that we're 68% confident that a predicated value is within +/-14 units of the actual value, and 95% confident that a predicted value is within +/- 28 units of the actual value.

ANOVA

	df	SS	MS	F	Significance F
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	b0 38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763
Size	b1 35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846

The **intercept** tells us the kind of 'base' value for Y.

The **coefficient** for each X variable tells us that for each 1 unit increase in X (all other things fixed), we expect to see Y go up by 35 units.

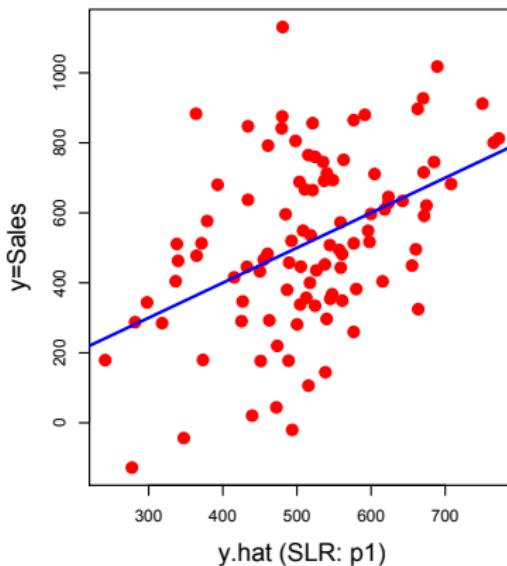
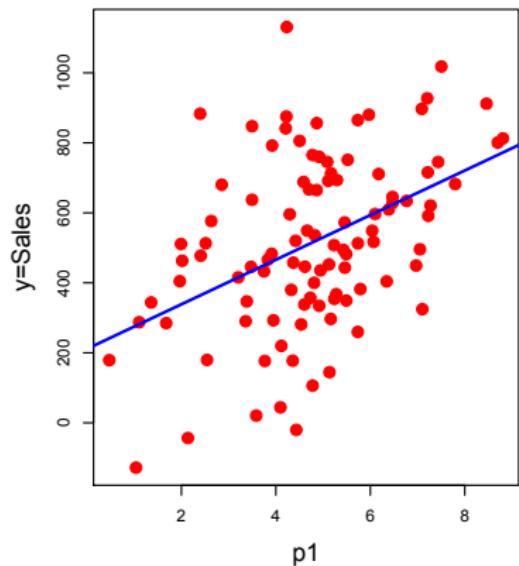
(Standard Error) - Like RMSE but for individual variables, it tells you the average error amount. We want the standard error to be small in relation to its coefficient.

(P-value) - Tells you which predictor variables are significant. A p-value less than 0.05 suggests that changes in your predictor variable are related to changes in Y. Helps us decide which variables to keep in the model.

(Lower & Upper 95%) - You can be 95% confident that the real value of the coefficient that you are estimating falls between these two numbers. If the interval does not contain 0 (meaning a change in X has zero impact on changes in Y), your P value will be .05 or less.

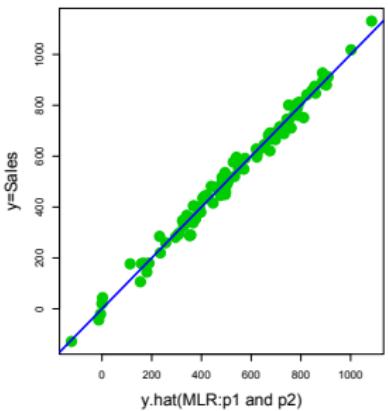
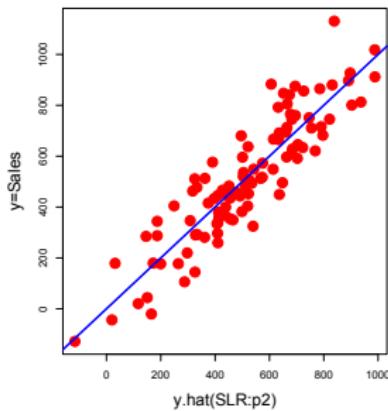
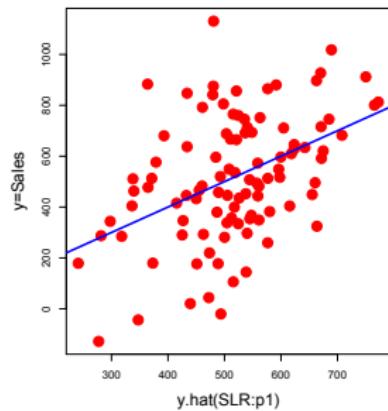
Fitted Values in MLR

With just $P1\dots$



- ▶ Left plot: *Sales vs $P1$ (something odd?)*
- ▶ Right plot: *Sales vs. \hat{y} (only $P1$ as a regressor)*

Fitted Values in MLR

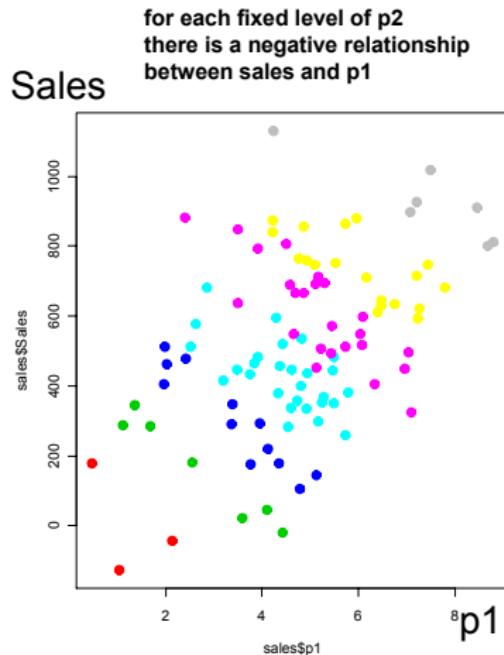
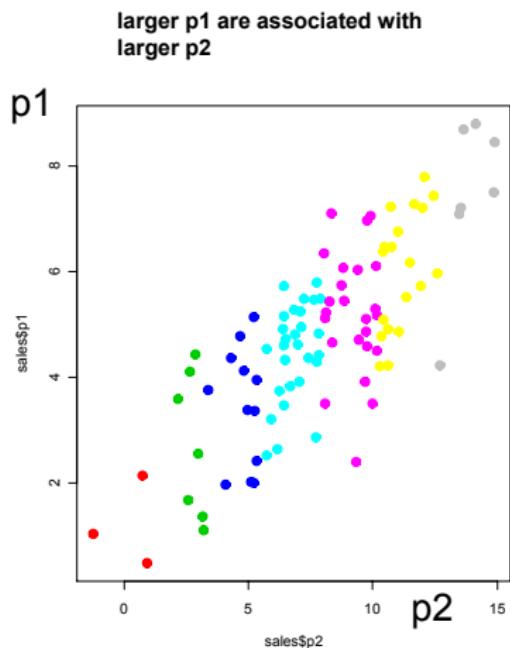


- ▶ First plot: *Sales regressed on P1 alone..*
- ▶ Second plot: *Sales regressed on P2 alone...*
- ▶ Third plot: *Sales regressed on P1 and P2*

Also look at residuals

Solving the Puzzle

- ▶ Let's look at a subset of points where $P1$ varies and $P2$ is held approximately constant...



Key Points to Remember

1. How dependencies between the X's affect our interpretation of a multiple regression.

Any time a report says two variables are related and there's a suggestion of a "causal" relationship, ask yourself whether or not other variables might be the real reason for the effect.

- ▶ Example: Why is it better to model beer vs. weight rather than beer vs. both height and weight?

2. How dependencies between the X's inflate standard errors (aka multicollinearity)

- ▶ in MLR, the standard errors are defined by the following formula:

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

3. Correlation does not imply causation
4. Succinct models with the "right" predictors are superior

- ▶ Reject predictor when the p-value is less than 0.05 (i.e. when the $|t_j| > 2$)

More Decisions

- ▶ How many X's do you have and what are they?
 - ▶ Bank Example: dummy coding and interaction effects
 - ▶ What if number of (potential) predictors is very large (p vs. n)
- ▶ Outliers in X or in Y
- ▶ Transformation of Variables (look at residuals!)
 - ▶ Non-constant residuals may suggest log transform

