

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*.

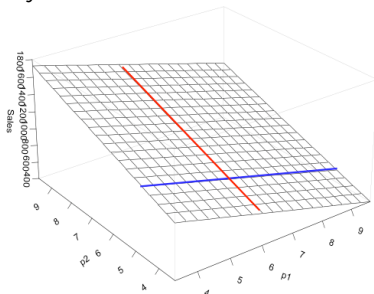
Q: What do the β s mean?

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

$$\text{Model: } \text{Sales}_i = \beta_0 + \beta_1 P1_i + \beta_2 P2_i + \epsilon_i, \epsilon \sim N(0, \sigma^2)$$

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

Plug-in Prediction in MLR

Suppose that by using advanced corporate espionage tactics, I discover that my competitor will charge \$10 the next quarter. After some marketing analysis I decided to charge \$8. **How much will I sell?**

Our model is

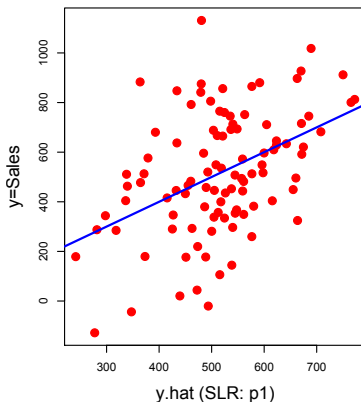
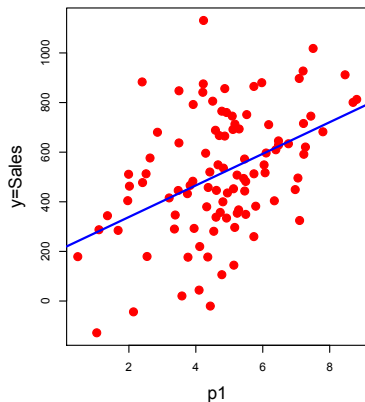
$$\text{Sales} = \beta_0 + \beta_1 P1 + \beta_2 P2 + \epsilon$$

Our estimates are $b_0 = 115$, $b_1 = -97$, $b_2 = 109$ and $s = 28$; i.e., $\epsilon \sim N(0, 28^2)$

Q: How will you estimate of sales when $P1=8$, $P2=10$ (95% confidence interval)?

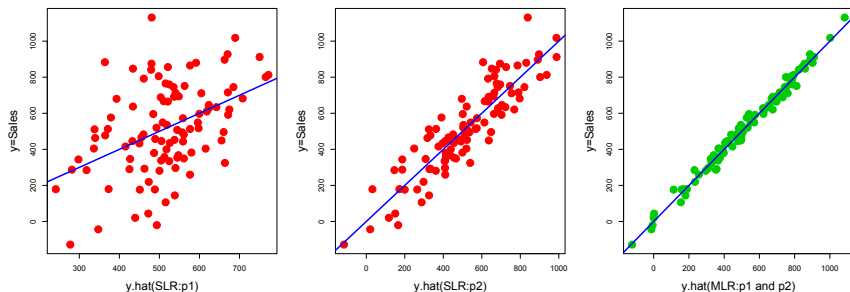
Fitted Values in MLR

With just P_1 ...



- ▶ Left plot: *Sales* vs P_1 (something odd?)
- ▶ Right plot: *Sales* vs. \hat{y} (only P_1 as a regressor)

Fitted Values in MLR

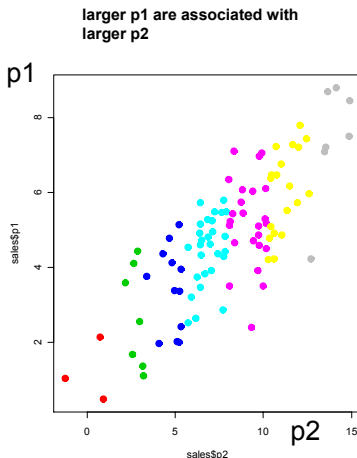


- ▶ First plot: *Sales* regressed on *P1* alone..
- ▶ Second plot: *Sales* regressed on *P2* alone...
- ▶ Third plot: *Sales* regressed on *P1* and *P2*

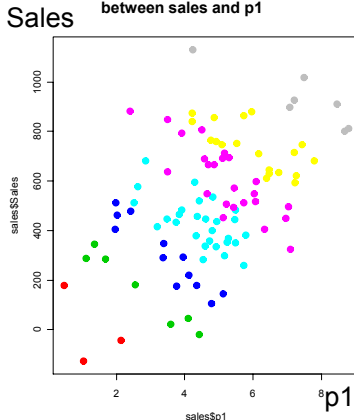
Also look at residuals

Solving the Puzzle

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



for each fixed level of $p2$
there is a negative relationship
between sales and $p1$



Intervals for Individual Coefficients

We denote the **sampling distribution** of each estimator as

$$b_j \sim N(\beta_j, s_{b_j}^2)$$

- ▶ A 95% C.I. for β_j is approximately $b_j \pm 2s_{b_j}$
- ▶ The t-stat: $t_j = \frac{(b_j - \beta_j^0)}{s_{b_j}}$ is the number of standard errors between the LS estimate and the null value (β_j^0)
- ▶ As before, we reject the null when t-stat is greater than 2 in absolute value
- ▶ Also as before, a small p-value leads to a rejection of the null
- ▶ Rejecting when the p-value is less than 0.05 is equivalent to rejecting when the $|t_j| > 2$

Understanding Multiple Regression: Summary

- ▶ There are two, very important things we need to understand about the MLR model:
 1. How dependencies between the X 's **affect our interpretation** of a multiple regression;
 2. How dependencies between the X 's **inflate standard errors** (aka multicollinearity)
 3. Correlation does not imply causation
 4. Succinct models with the "right" predictors are superior

Understanding Multiple Regression

A relationship between y and x (or x 's) may be driven by variables “lurking” in the background which are related to your current x 's.

correlation is NOT (necessarily) causation

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. Multiple regression allows us to **control** for all important variables by including them into the regression.

- ▶ **Example: “Once we control for weight, height and beers are NOT related” !!**
- ▶ **Why is it better to model beer vs. weight rather than beer vs. both height and weight?**

Understanding Multiple Regression: Collinearity

- ▶ Previous Example: relationship amongst the X 's can **affect our interpretation** of a multiple regression
- ▶ These dependencies also **inflate the standard errors** for the regression coefficients, and hence our uncertainty about them.
- ▶ In simple linear regression our uncertainty about b_1 is measured by

$$s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2} = \frac{s^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- ▶ The more variation in X (the larger s_x^2) the more “we know” about β_1 ... ie, $(b_1 - \beta_1)$ is smaller.

Understanding Multiple Regression: Collinearity

- ▶ In Multiple Regression we seek to relate the variation in Y to the variation in an X holding the other X 's fixed. So, we need to see how much each X varies on its own.
- ▶ in MLR, the standard errors are defined by the following formula:

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

- ▶ How do we measure the bottom part of the equation? We regress X_j on all the other X 's and compute the residual sum of squares (call it SSE_j) so that $s_{b_j}^2 = \frac{s^2}{SSE_j}$
- ▶ SSE_j is (much) lower if independent variables are (highly) correlated!

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

