Section 4: Multiple Linear Regression

Jared S. Murray

The Multiple Regression Model

Many problems involve more than one independent variable or factor which affects the dependent or response variable.

- More than size to predict house price!
- Demand for a product given prices of competing brands, advertising,house hold attributes, etc.

In SLR, the conditional mean of Y depends on X. The Multiple Linear Regression (MLR) model extends this idea to include more than one independent variable.

The MLR Model

Same as always, but with more covariates.

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

Recall the key assumptions of our linear regression 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)

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

The MLR Model

Our interpretation of regression coefficients can be extended from the simple single covariate regression case:

$$\beta_j = \frac{\partial E[Y|X_1, \ldots, X_p]}{\partial X_j}$$

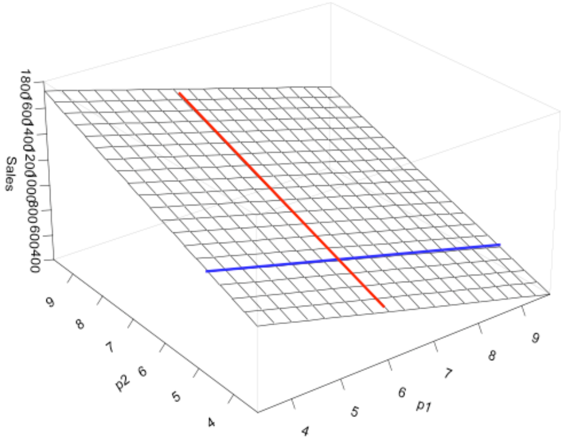
Holding all other variables constant, β_j is the average change in Y per unit change in X_j.

The MLR Model

If $p = 2$, we can plot the regression surface in 3D.

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

$$Sales = \beta_0 + \beta_1 P1 + \beta_2 P2 + \epsilon$$



Least Squares

The data...

p1	p2	Sales
5.1356702	5.2041860	144.48788
3.4954600	8.0597324	637.24524
7.2753406	11.6759787	620.78693
4.6628156	8.3644209	549.00714
3.5845370	2.1502922	20.42542
5.1679168	10.1530371	713.00665
3.3840914	4.9465690	346.70679
4.2930636	7.7605691	595.77625
4.3690944	7.4288974	457.64694
7.2266002	10.7113247	591.45483
...	...	...

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

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

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

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

$$Sales = \beta_0 + \beta_1 P1 + \beta_2 P2 + \epsilon$$

with $\epsilon \sim N(0, \sigma^2)$

Our estimates are $b_0 = 115$, $b_1 = -97$, $b_2 = 109$ and $s = 28$ which leads to

$$Sales = 115 + -97 * P1 + 109 * P2 + \epsilon$$

with $\epsilon \sim N(0, 28^2)$

Plug-in Prediction in MLR

By plugging-in the numbers,

$$\begin{aligned} Sales &= 115 + -97 * 8 + 109 * 10 + \epsilon \\ &= 437 + \epsilon \end{aligned}$$

$$Sales | P1 = 8, P2 = 10 \sim N(437, 28^2)$$

and the 95% Prediction Interval is $(437 \pm 2 * 28)$

$$381 < Sales < 493$$

Residual Standard Error

The calculation for s^2 is exactly the same:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p-1}$$

- $\hat{Y}_i = b_0 + b_1 X_{1i} + \cdots + b_p X_{pi}$
- The residual "standard error" is the estimate for the standard deviation of ϵ,i.e,

$$\hat{\sigma} = s = \sqrt{s^2}.$$

In Excel... Do we know all of these numbers?

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

95% C.I. for $\beta_1 \approx b1 \pm 2 \times s_{b_1}$

$$[-97.66 - 2 \times 2.67; -97.66 + 2 \times 2.67] = [-102.95; -92.36]$$

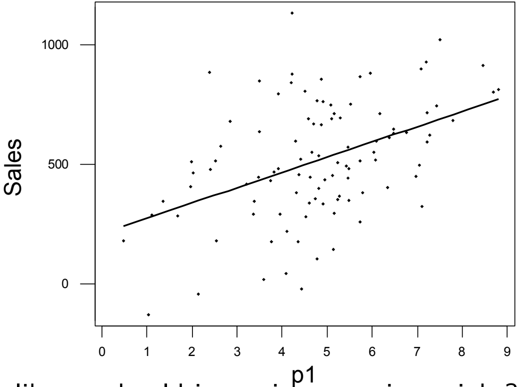
Understanding Multiple Regression

The Sales Data:

- *Sales* : units sold in excess of a baseline
- *P1*: our price in \$ (in excess of a baseline price)
- *P2*: competitors price (again, over a baseline)

Understanding Multiple Regression

- If we regress Sales on our own price, we obtain a somewhat surprising conclusion... the higher the price the more we sell!!



- It looks like we should just raise our prices, right? NO, not if you have taken this statistics class!

Understanding Multiple Regression

- The regression equation for Sales on own price (P1) is:

$$Sales = 211 + 63.7P1$$

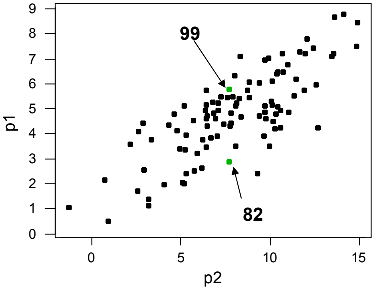
- If now we add the competitors price to the regression we get

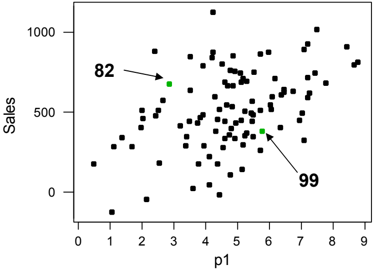
$$Sales = 116 - 97.7P1 + 109P2$$

- Does this look better? How did it happen?
- Remember: -97.7 is the affect on sales of a change in $P1$ with $P2$ held fixed!!

Understanding Multiple Regression

- How can we see what is going on? Let's compare Sales in two different observations: weeks 82 and 99.
- We see that an increase in $P1$, holding $P2$ constant, corresponds to a drop in Sales!

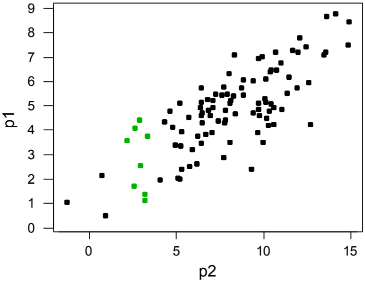


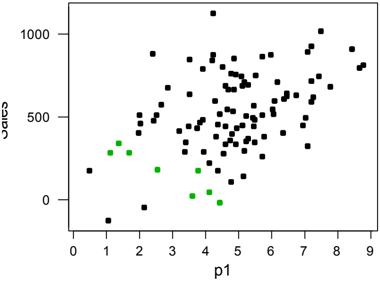


- Note the strong relationship (dependence) between $P1$ and $P2$!!

Understanding Multiple Regression

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



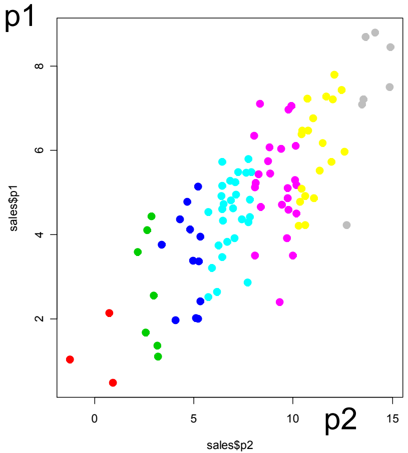


- For a fixed level of $P2$, variation in $P1$ is negatively correlated with Sales!!

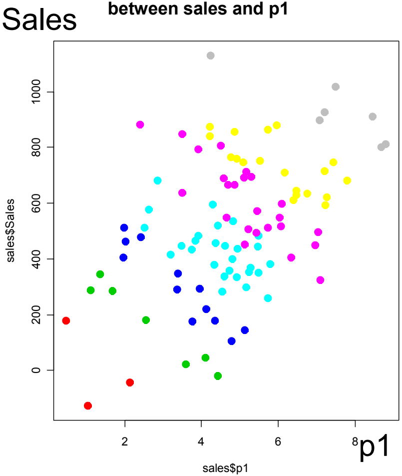
Understanding Multiple Regression

- Below, different colors indicate different ranges for $P2$...

larger p1 are associated with larger p2



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



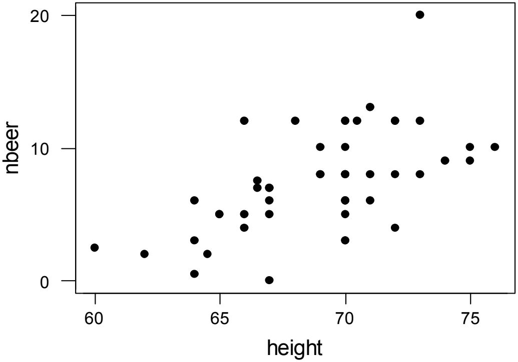
Understanding Multiple Regression

- Summary:
  1. A larger $P1$ is associated with larger $P2$ and the overall effect leads to bigger sales
  2. With $P2$ held fixed, a larger $P1$ leads to lower sales
  3. MLR does the trick and unveils the "correct" economic relationship between Sales and prices!

Understanding Multiple Regression

Beer Data (from an MBA class)

- *nbeer* – number of beers before getting drunk
- *height and weight*



Is number of beers related to height?

Understanding Multiple Regression

$$nbeers = \beta_0 + \beta_1 height + \epsilon$$

Regression Statistics	
Multiple R	0.58
R Square	0.34
Adjusted R Square	0.33
Standard Error	3.11
Observations	50.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.00	237.77	237.77	24.60	0.00
Residual	48.00	463.86	9.66		
Total	49.00	701.63			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-36.92	8.96	-4.12	0.00	-54.93	-18.91
height	0.64	0.13	4.96	0.00	0.38	0.90

Yes! Beers and height are related...

Understanding Multiple Regression

$$nbeers = \beta_0 + \beta_1 weight + \beta_2 height + \epsilon$$

Regression Statistics	
Multiple R	0.69
R Square	0.48
Adjusted R Square	0.46
Standard Error	2.78
Observations	50.00

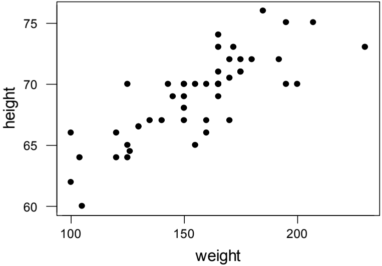
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.00	337.24	168.62	21.75	0.00
Residual	47.00	364.38	7.75		
Total	49.00	701.63			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-11.19	10.77	-1.04	0.30	-32.85	10.48
weight	0.09	0.02	3.58	0.00	0.04	0.13
height	0.08	0.20	0.40	0.69	-0.32	0.47

What about now?? Height is not necessarily a factor...

Understanding Multiple Regression



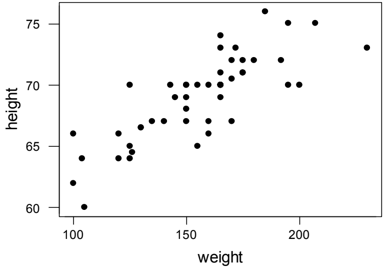
<u>The correlations:</u>

	nbeer	weight
weight	0.692	
height	0.582	0.806

The two x's are highly correlated !!

- If we regress "beers" only on height we see an effect. Bigger heights go with more beers.
- However, when height goes up weight tends to go up as well... in the first regression, height was a proxy for the real *cause* of drinking ability. Bigger people can drink more and weight is a more accurate measure of "bigness".

Understanding Multiple Regression



<u>The correlations:</u>

	nbeer	weight
weight	0.692	
height	0.582	0.806

The two x's are highly correlated !!

- In the multiple regression, when we consider only the variation in height that is not associated with variation in weight, we see no relationship between height and beers.

Understanding Multiple Regression

$$nbeers = \beta_0 + \beta_1 weight + \epsilon$$

Regression Statistics	
Multiple R	0.69
R Square	0.48
Adjusted R	0.47
Standard E	2.76
Observatio	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	336.0317807	336.0318	44.11878	2.60227E-08
Residual	48	365.5932193	7.616525		
Total	49	701.625			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.021	2.213	-3.172	0.003	-11.471	-2.571
weight	0.093	0.014	6.642	0.000	0.065	0.121

Why is this a better model than the one with weight and height??

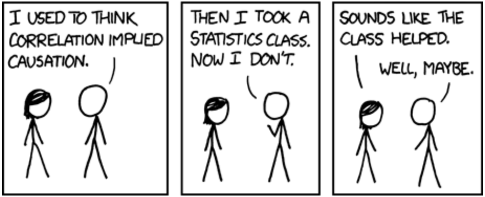
Understanding Multiple Regression

In general, when we see a relationship between y and x (or x's), that relationship may be driven by variables "lurking" in the background which are related to your current x's.

This makes it hard to reliably find "causal" relationships. Any correlation (association) you find could be caused by other variables in the background... correlation is NOT 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. "Once we control for weight, height and beers are NOT related"!!

correlation is NOT causation



also...

- http://www.tylervigen.com/spurious-correlations

Back to Baseball – Let's try to add AVG on top of OBP

Regression Statistics	
Multiple R	0.948136
R Square	0.898961
Adjusted R Square	0.891477
Standard Error	0.160502
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	6.188355	3.094177	120.1119098	3.63577E-14
Residual	27	0.695541	0.025761		
Total	29	6.883896			

	Coefficients	*andard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.933633	0.844353	-9.396107	5.30996E-10	-9.666102081	-6.201163
AVG	7.810397	4.014609	1.945494	0.062195793	-0.426899658	16.04769
OBP	31.77892	3.802577	8.357205	5.74232E-09	23.9766719	39.58116

$$R/G = \beta_0 + \beta_1 AVG + \beta_2 OBP + \epsilon$$

Is AVG any good?

Back to Baseball - Now let's add SLG

Regression Statistics	
Multiple R	0.955698
R Square	0.913359
Adjusted R Square	0.906941
Standard Error	0.148627
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	6.28747	3.143735	142.31576	4.56302E-15
Residual	27	0.596426	0.02209		
Total	29	6.883896			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.014316	0.81991	-8.554984	3.60968E-09	-8.69663241	-5.332
OBP	27.59287	4.003208	6.892689	2.09112E-07	19.37896463	35.80677
SLG	6.031124	2.021542	2.983428	0.005983713	1.883262806	10.17899

$$R/G = \beta_0 + \beta_1 OBP + \beta_2 SLG + \epsilon$$

What about now? Is SLG any good

Back to Baseball

Correlations

AVG	1		
OBP	0.77	1	
SLG	0.75	0.83	1

- When AVG is added to the model with OBP, no additional information is conveyed. AVG does nothing "on its own" to help predict Runs per Game...
- SLG however, measures something that OBP doesn't (power!) and by doing something "on its own" it is relevant to help predict Runs per Game. (Okay, but not much...)

Detecting Sex Discrimination

$$Salary_i = \beta_0 + \beta_1 Sex_i + \epsilon_i$$

Regression Statistics	
Multiple R	0.346541
R Square	0.120091
Adjusted R Square	0.115819
Standard Error	10.58426
Observations	208

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3149.634	3149.6	28.1151	2.93545E-07
Residual	206	23077.47	112.03		
Total	207	26227.11			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	37.20993	0.894533	41.597	3E-102	35.44631451	38.9735426
Gender	8.295513	1.564493	5.3024	2.9E-07	5.211041089	11.3799841

$\hat{\beta}_1 = b_1 = 8.29...$ on average, a male makes approximately \$8,300 more than a female in this firm.

How should the plaintiff's lawyer use the confidence interval in his presentation?

Detecting Sex Discrimination

How can the defense attorney try to counteract the plaintiff's argument?

Perhaps, the observed difference in salaries is related to other variables in the background and NOT to policy discrimination...

Obviously, there are many other factors which we can legitimately use in determining salaries:

- education
- job productivity
- experience

How can we use regression to incorporate additional information?

Detecting Sex Discrimination

Let's add a measure of experience...

$$Salary_i = \beta_0 + \beta_1 Sex_i + \beta_2 Exp_i + \epsilon_i$$

What does that mean?

$$\begin{aligned} E[Salary|Sex = 0, Exp] &= \beta_0 + \beta_2 Exp \\ E[Salary|Sex = 1, Exp] &= (\beta_0 + \beta_1) + \beta_2 Exp \end{aligned}$$

Detecting Sex Discrimination

	Exp	Gender	Salary	Sex
1	3	Male	32.00	1
2	14	Female	39.10	0
3	12	Female	33.20	0
4	8	Female	30.60	0
5	3	Male	29.00	1
...	...	...		
208	33	Female	30.00	0

Detecting Sex Discrimination

$$Salary_i = \beta_0 + \beta_1 Sex_i + \beta_2 Exp + \epsilon_i$$

Regression Statistics	
Multiple R	0.701
R Square	0.491
Adjusted R Square	0.486
Standard Error	8.070
Observations	208

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.000	12876.269	6438.134	98.857	0.000
Residual	205.000	13350.839	65.126		
Total	207.000	26227.107			

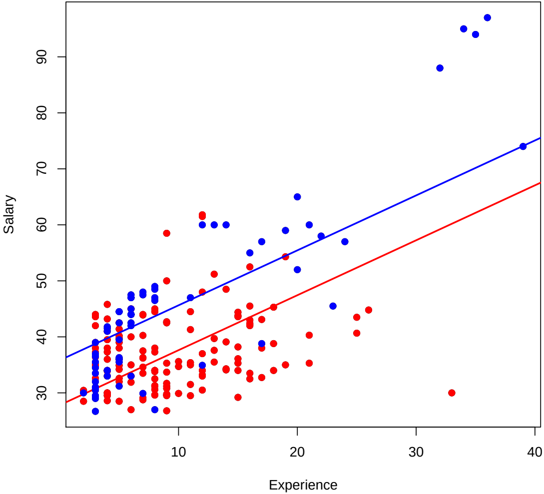
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	27.812	1.028	27.057	0.000	25.785	29.839
Sex	8.012	1.193	6.715	0.000	5.660	10.364
Exp	0.981	0.080	12.221	0.000	0.823	1.139

$$Salary_i = 27 + 8 Sex_i + 0.98 Exp_i + \epsilon_i$$

Is this good or bad news for the defense?

Detecting Sex Discrimination

$$Salary_i = \begin{cases} 27 + 0.98 Exp_i + \epsilon_i & \text{females} \\ 35 + 0.98 Exp_i + \epsilon_i & \text{males} \end{cases}$$



More than Two Categories

We can use dummy variables in situations in which there are more than two categories. Dummy variables are needed for each category except one, designated as the “base” category.

Why? *Remember that the numerical value of each category has no quantitative meaning!*

Example: House Prices

We want to evaluate the difference in house prices in a couple of different neighborhoods.

	Nbhd	SqFt	Price
1	2	1.79	114.3
2	2	2.03	114.2
3	2	1.74	114.8
4	2	1.98	94.7
5	2	2.13	119.8
6	1	1.78	114.6
7	3	1.83	151.6
8	3	2.16	150.7
...	...	...	...

Example: House Prices

Let's create the dummy variables $dn1$, $dn2$ and $dn3$...

	Nbhd	SqFt	Price	dn1	dn2	dn3
1	2	1.79	114.3	0	1	0
2	2	2.03	114.2	0	1	0
3	2	1.74	114.8	0	1	0
4	2	1.98	94.7	0	1	0
5	2	2.13	119.8	0	1	0
6	1	1.78	114.6	1	0	0
7	3	1.83	151.6	0	0	1
8	3	2.16	150.7	0	0	1

... ...

Example: House Prices

$$Price_i = \beta_0 + \beta_1 dn1_i + \beta_2 dn2_i + \beta_3 Size_i + \epsilon_i$$

$$\begin{aligned}
E[Price|dn1 = 1, Size] &= \beta_0 + \beta_1 + \beta_3 Size \quad \text{(Nbhd 1)} \\
E[Price|dn2 = 1, Size] &= \beta_0 + \beta_2 + \beta_3 Size \quad \text{(Nbhd 2)} \\
E[Price|dn1 = 0, dn2 = 0, Size] &= \beta_0 + \beta_3 Size \quad \text{(Nbhd 3)}
\end{aligned}$$

Example: House Prices

$$Price = \beta_0 + \beta_1 dn1 + \beta_2 dn2 + \beta_3 Size + \epsilon$$

Regression Statistics	
Multiple R	0.828
R Square	0.685
Adjusted R Square	0.677
Standard Error	15.260
Observations	128

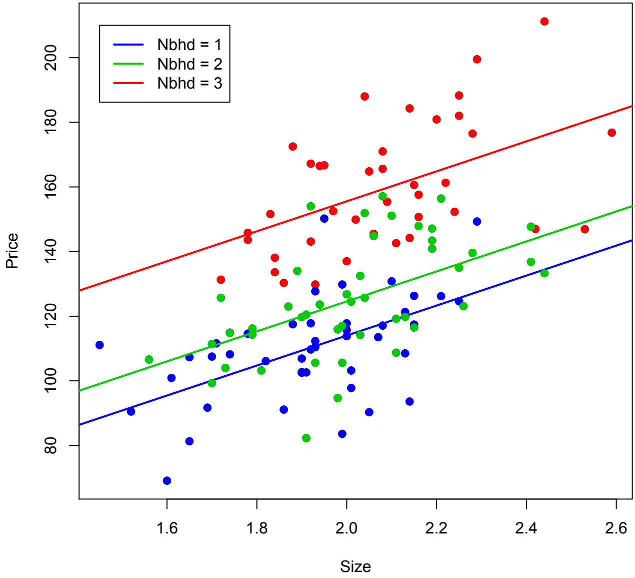
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	3	62809.1504	20936	89.9053	5.8E-31
Residual	124	28876.0639	232.87		
Total	127	91685.2143			

	Coefficients	*Standard Erroi*	*t Stat*	*P-value*	*.ower 95%*	*Jpper 95%*
Intercept	62.78	14.25	4.41	0.00	34.58	90.98
dn1	-41.54	3.53	-11.75	0.00	-48.53	-34.54
dn2	-30.97	3.37	-9.19	0.00	-37.63	-24.30
size	46.39	6.75	6.88	0.00	33.03	59.74

$$Price = 62.78 - 41.54 dn1 - 30.97 dn2 + 46.39 Size + \epsilon$$

Example: House Prices



Example: House Prices

$$Price = \beta_0 + \beta_1 Size + \epsilon$$

Regression Statistics	
Multiple R	0.553
R Square	0.306
Adjusted R Square	0.300
Standard Error	22.476
Observations	128

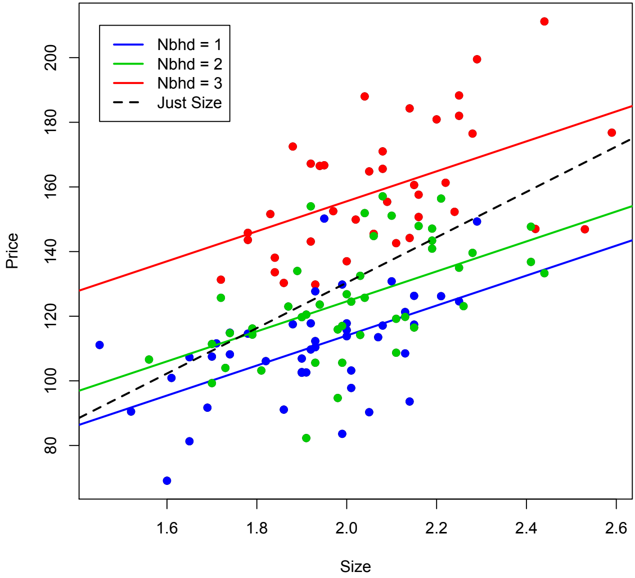
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	28036.4	28036.36	55.501	1E-11
Residual	126	63648.9	505.1496		
Total	127	91685.2			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*ower 95%*	*pper 95%*
Intercept	-10.09	18.97	-0.53	0.60	-47.62	27.44
size	70.23	9.43	7.45	0.00	51.57	88.88

$$Price = -10.09 + 70.23 Size + \epsilon$$

Example: House Prices



Things to remember:

- Intervals are your friend! Understanding uncertainty is a key element for sound business decisions.
- Correlation is NOT causation!
- When presented with a analysis from a regression model or any analysis that implies a causal relationship, skepticism is always a good first response! Ask question... "is there an alternative explanation for this result"?
- Simple models are often better than very complex alternatives...