Division of
Statistics + Scientific Computation

THE UNIVERSITY OF TEXAS AT AUSTIN

Advanced Regression
Summer Statistics Institute

Day 2: MLR and Dummy Variables

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:

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

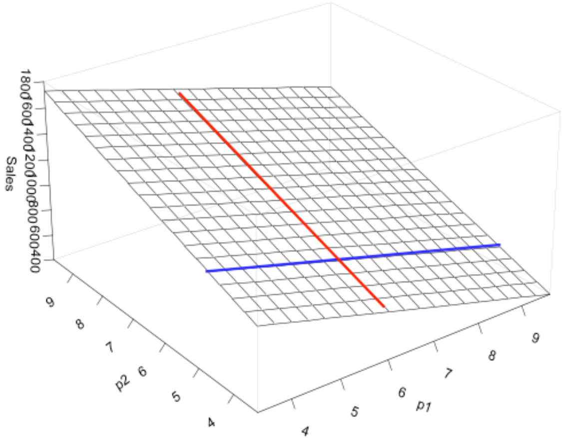
Holding all other variables constant, β_j is the average change in Y per unit change in X_j.

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

$$Y = \beta_0 + \beta_1 X_1 \ldots + \beta_p X_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

How do we estimate the MLR model parameters?

The principle of Least Squares is exactly the same as before:

- Define the fitted values
- Find the best fitting plane by minimizing the sum of squared residuals.

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 $(437 \pm 2 * 28)$

$$381 < Sales < 493$$

Least Squares

Just as before, each b_i is our estimate of β_i

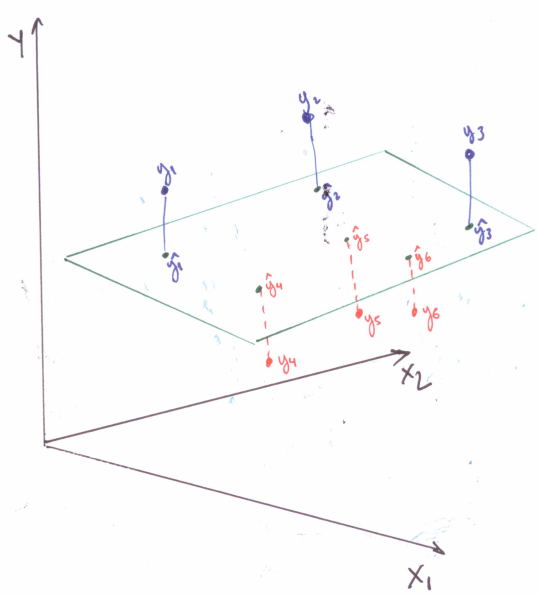
Fitted Values: $\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \ldots + b_p X_p$.

Residuals: $e_i = Y_i - \hat{Y}_i$.

Least Squares: Find $b_0, b_1, b_2, \ldots, b_p$ to minimize $\sum_{i=1}^{n} e_i^2$.

In MLR the formulas for the b_i's are too complicated so we won't talk about them...

Least Squares



Least Squares

p1	p2	Sales		yhat	residuals	b0	b1	b2
5.13567	5.204186	144.4879		184.0963	-39.60838	115	-97	109
3.49546	8.059732	637.2452		654.4512	-17.20597			
7.275341	11.67598	620.7869		681.9736	-61.18671			
4.662816	**8.364421**	**549.0071**		**574.4288**	**-25.42163**			
3.584537	2.150292	20.42542		1.681753	18.74367			
5.167917	10.15304	713.0067		720.3931	-7.386461			
3.384091	4.946569	346.7068		325.9192	20.78763			
4.293064	7.760569	595.7762		544.4749	51.30139			
4.369094	7.428897	457.6469		500.9477	-43.30072			
7.2266	10.71132	591.4548		581.5542	9.900659			

Excel break: fitted values, residuals,...

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,

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

Residuals in MLR

As in the SLR model, the residuals in multiple regression are purged of any linear relationship to the independent variables. Once again, they are on average zero.

Because the fitted values are an exact linear combination of the X's they are not correlated to the residuals.

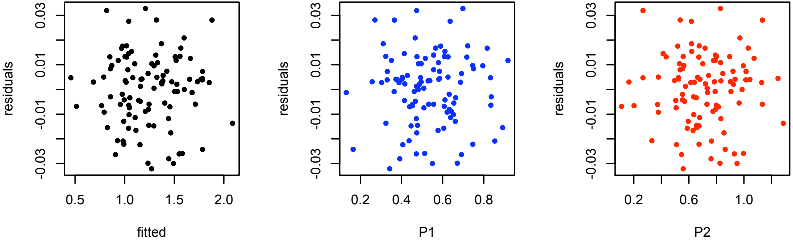
We decompose Y into the part predicted by X and the part due to idiosyncratic error.

$$Y = \hat{Y} + e$$

$$\bar{e} = 0; \quad \text{corr}(X_j, e) = 0; \quad \text{corr}(\hat{Y}, e) = 0$$

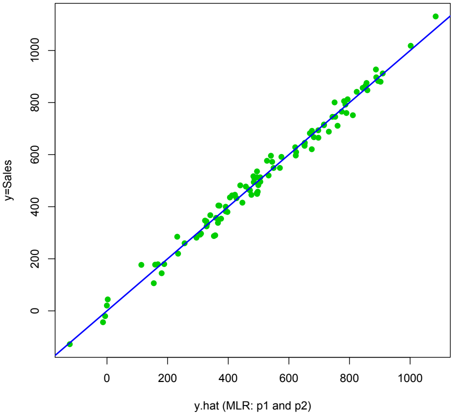
Residuals in MLR

Consider the residuals from the Sales data:



Fitted Values in MLR

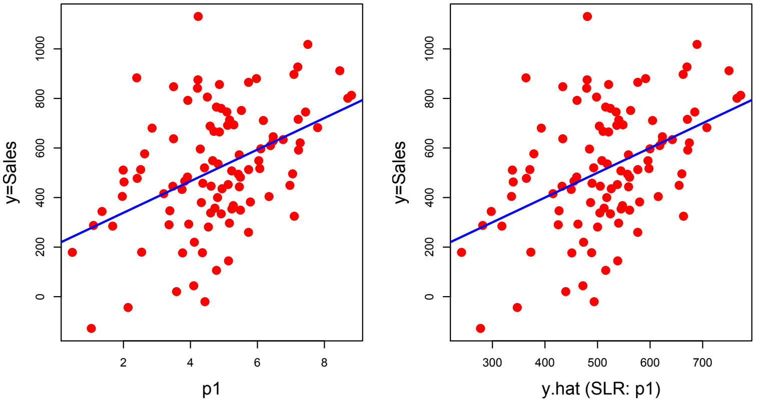
Another great plot for MLR problems is to look at Y (true values) against $\hat{Y}$ (fitted values).



If things are working, these values should form a nice straight line. Can you guess the slope of the blue line?

Fitted Values in MLR

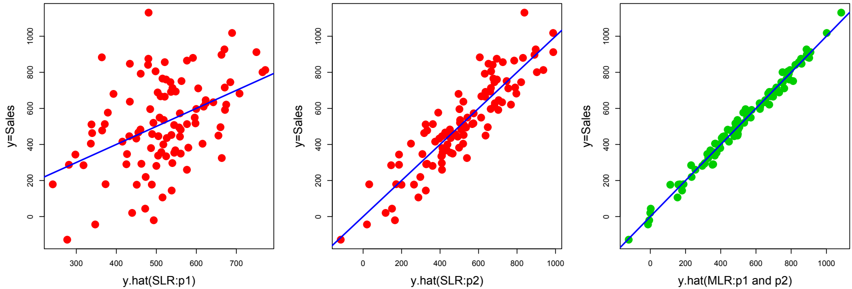
With just $P1$...



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

Fitted Values in MLR

Now, with $P1$ and $P2$...



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

R-squared

- We still have our old variance decomposition identity...

$$SST = SSR + SSE$$

- ... and R^2 is once again defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

  telling us the percentage of variation in Y explained by the X's.

- In Excel, R^2 is in the same place and "Multiple R" refers to the correlation between $\hat{Y}$ and Y.

Least Squares

$$Sales_i = \beta_0 + \beta_1 P1_i + \beta_2 P2_i + \epsilon_i$$

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

$$R^2 = 0.99$$

$$\text{Multiple R} = r_{Y,\hat{Y}} = \text{corr}(Y, \hat{Y}) = 0.99$$

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

Back to Baseball

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

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 Errc*	*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

$R^2 = 0.913$

Multiple R $= r_{Y,\hat{Y}} = \text{corr}(Y, \hat{Y}) = 0.955$

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

Intervals for Individual Coefficients

As in SLR, the sampling distribution tells us how close we can expect b_j to be from β_j

The LS estimators are unbiased: $E[b_j] = \beta_j$ for $j = 0, \ldots, d$.

- We denote the sampling distribution of each estimator as

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

Intervals for Individual Coefficients

Intervals and t-statistics are exactly the same as in SLR.

- 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 $|t_j| > 2$

Intervals for Individual Coefficients

IMPORTANT: Intervals and testing via b_j & s_{b_j} are one-at-a-time procedures:

- You are evaluating the j^{th} coefficient conditional on the other X's being in the model, but regardless of the values you've estimated for the other b's.

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

- 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 multicolinearity)
- We will look at a few examples to illustrate the ideas...

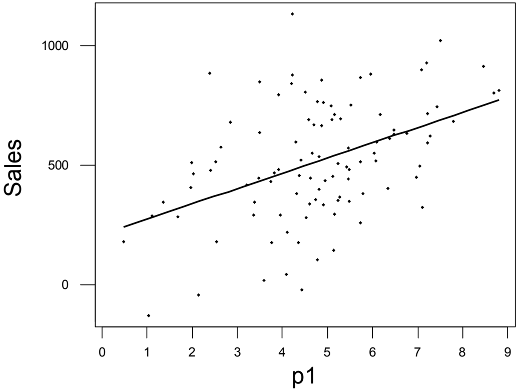
Understanding Multiple Regression

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:

$$Sales = 211 + 63.7P1$$

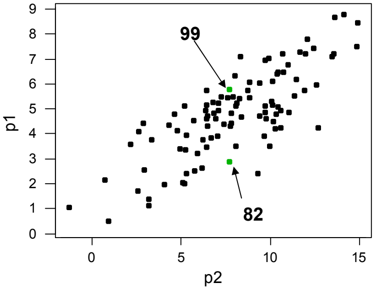
- If now we add the competitors price to the regression we get

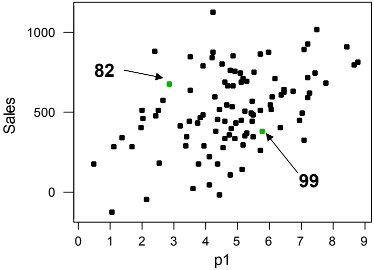
$$Sales = 116 - 97.7P1 + 109P2$$

- 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!

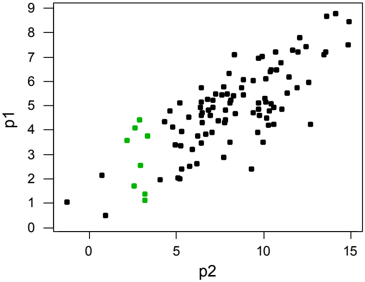


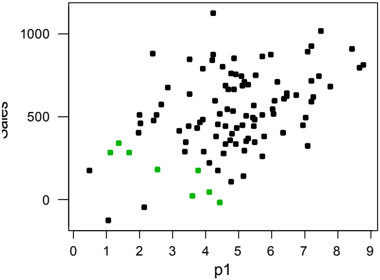


- Note the strong relationship (dependence) between $P1$ and $P2$!!

Understanding Multiple Regression

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

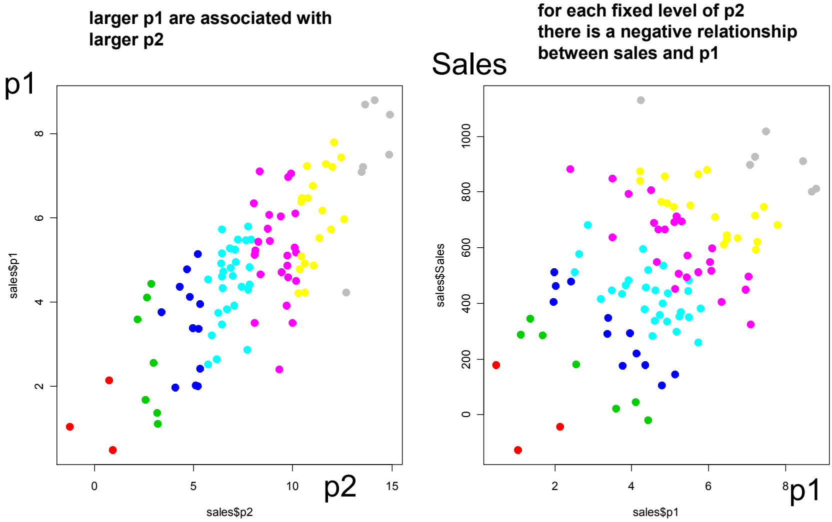




- For a fixed level of $P2$, variation in $P1$ is negatively correlated with Sales!!

Understanding Multiple Regression

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



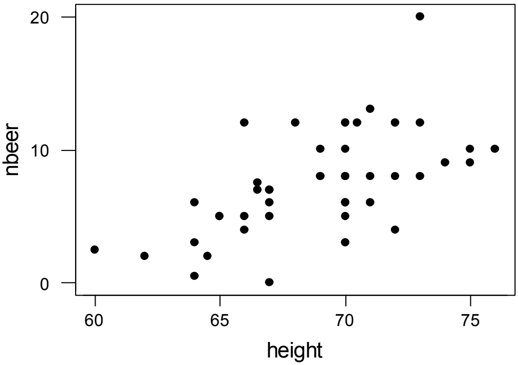
Understanding Multiple Regression

- 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

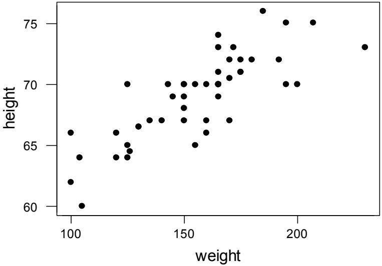
ANOVA

	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



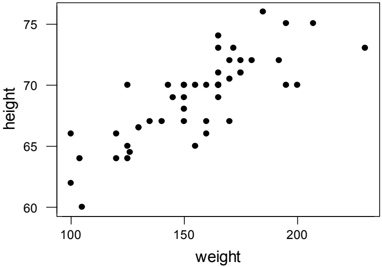
The correlations:

	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"!!

Understanding Multiple Regression

- With the above examples we saw how the relationship amongst the X's can affect our interpretation of a multiple regression... we will now look at how these dependencies will inflate the standard errors for the regression coefficients, and hence our uncertainty about them.
- Remember that 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

- 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

$$s_{b_j}^2 = \frac{s^2}{SSE_j}$$

Understanding Multiple Regression

In the "number of beers example"... $s = 2.78$ and the regression on height on weight gives...

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.806
R Square	0.649
Adjusted R Squar	0.642
Standard Error	2.051
Observations	50.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.000	373.148	373.148	88.734	0.000
Residual	48.000	201.852	4.205		
Total	49.000	575.000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	53.751	1.645	32.684	0.000	50.444	57.058
weight	0.098	0.010	9.420	0.000	0.077	0.119

$$SSE_2 = 201.85$$

$$s_{b_2} = \sqrt{\frac{2.78^2}{201.85}} = 0.20$$ Is this right?

Understanding Multiple Regression

- What happens if we are regressing Y on X's that are highly correlated. SSE_j goes down and the standard error s_{b_j} goes up!
- What is the effect on the confidence intervals ($b_j \pm 2 \times s_{b_j}$)? They get wider!
- This situation is called Multicolinearity
- If a variable X does nothing "on its own" we can't estimate its effect on Y.

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...)

F-tests

- In many situation, we need a testing procedure that can address *simultaneous* hypotheses about more than one coefficient
- Why not the t-test?
- We will look at two important types of simultaneous tests
  (i) Overall Test of Significance
  (ii) Partial F-test

The first test will help us determine whether or not our regression is worth anything... the second will allow us to compare different models.

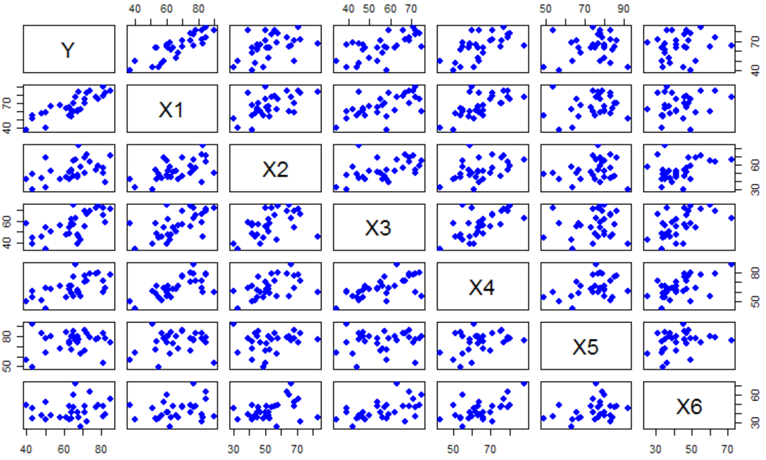
Supervisor Performance Data

Suppose you are interested in the relationship between the overall performance of supervisors to specific activities involving interactions between supervisors and employees (from a psychology management study)

The Data

- Y = Overall rating of supervisor
- X_1 = Handles employee complaints
- X_2 = Does not allow special privileges
- X_3 = Opportunity to learn new things
- X_4 = Raises based on performance
- X_5 = Too critical of poor performance
- X_6 = Rate of advancing to better jobs

Supervisor Performance Data



Supervisor Performance Data

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.855921721
R Square	0.732601993
Adjusted R Square	0.662845991
Standard Error	7.067993765
Observations	30

ANOVA

	df	SS	MS	F	Significance F
Regression	6	3147.966342	524.6611	10.50235	1.24041E-05
Residual	23	1149.000325	49.95654		
Total	29	4296.966667			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	10.78707639	11.58925724	0.930782	0.361634	-13.18712868	34.76128	-21.747859	43.32201173
X1	0.613187608	0.160983115	3.809018	0.000903	0.280168664	0.946207	0.161254	1.06512125
X2	-0.073050143	0.13572469	-0.538223	0.595594	-0.353818055	0.207718	-0.4540749	0.307974622
X3	0.320332116	0.168520319	1.900852	0.069925	-0.028278721	0.668943	-0.152761	0.793425219
X4	0.081732134	0.221477677	0.369031	0.71548	-0.376429347	0.539894	-0.5400301	0.703494319
X5	0.038381447	0.146995442	0.261106	0.796334	-0.265701791	0.342465	-0.3742841	0.451046997
X6	-0.217056682	0.178209471	-1.217986	0.235577	-0.585711058	0.151598	-0.7173505	0.283237125

Is there any relationship here? Are all the coefficients significant?
What about all of them together?

Why not look at R^2

- R^2 in MLR is still a measure of goodness of fit.
- However it ALWAYS grows as we increase the number of explanatory variables.
- Even if there is no relationship between the $X's$ and Y, $R^2 > 0$!!
- To see this let's look at some "Garbage" Data

Garbage Data

I made up 6 “garbage” variables that have nothing to do with Y...

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.516876852
R Square	0.26716168
Adjusted R Square	0.075986466
Standard Error	11.70095097
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1147.985	191.3308	1.39747	0.257927747
Residual	23	3148.982	136.9123		
Total	29	4296.967			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	94.8053024	38.6485	2.453014	0.022169	14.85478564	174.7558	-13.6940154	203.3046202
G1	0.241049359	0.369932	0.651605	0.521115	-0.524213203	1.006312	-0.79747383	1.279572553
G2	-0.739495869	0.341006	-2.168569	0.040705	-1.444921431	-0.03407	-1.69681541	0.217823675
G3	-0.564272368	0.463453	-1.217539	0.235744	-1.522998304	0.394454	-1.86534101	0.736796272
G4	0.156297568	0.291278	0.536592	0.596702	-0.446257444	0.758853	-0.66141832	0.974013455
G5	-0.267328742	0.266723	-1.002269	0.326642	-0.819088173	0.284431	-1.01611092	0.481453434
G6	0.441170035	0.329715	1.338034	0.193965	-0.240897504	1.123238	-0.48445078	1.366790852

Garbage Data

- R^2 is 26% !!
- We need to develop a way to see whether a R^2 of 26% can happen by chance when all the true β's are zero.
- It turns out that if we transform R^2 we can solve this.

Define

$$f = \frac{R^2/p}{(1 - R^2)/(n - p - 1)}$$

A big f corresponds to a big R^2 but there is a distribution that tells what kind of f we are likely to get when all the coefficients are indeed zero... The f statistic provides a scale that allows us to decide if "big" is "big enough".

The F-test

We are testing:

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_p = 0$$

$$H_1 : \text{at least one } \beta_j \neq 0.$$

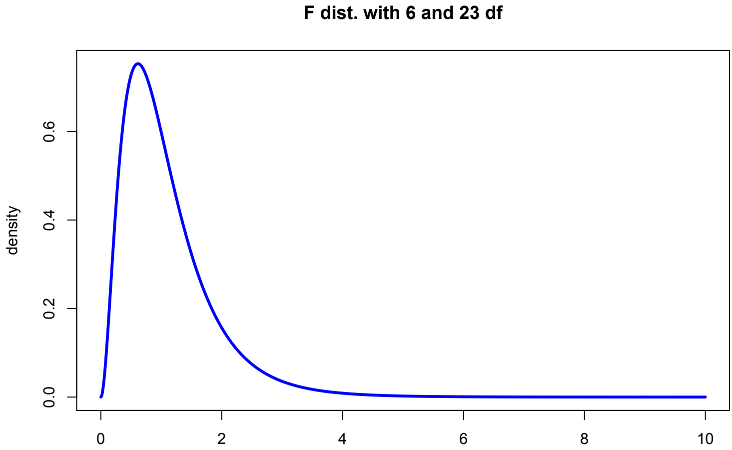
This is the F-test of overall significance. Under the null hypothesis f is distributed:

$$f \sim F_{p,n-p-1}$$

- Generally, $f > 4$ is very significant (reject the null).

The F-test

What kind of distribution is this?



It is a right skewed, positive valued family of distributions indexed by two parameters (the two df values).

The F-test

Let's check this test for the "garbage" data...

ANOVA

	df	SS	MS	F	Significance F
Regression	6	1147.985	191.3308	1.39747	0.257927747
Residual	23	3148.982	136.9123		
Total	29	4296.967			

How about the original analysis (survey variables)...

ANOVA

	df	SS	MS	F	Significance F
Regression	6	3147.966342	524.6611	10.50235	1.24041E-05
Residual	23	1149.000325	49.95654		
Total	29	4296.966667			

F-test

The *p*-value for the F-test is

$$\text{p-value} = Pr(F_{p,n-p-1} > f)$$

- We usually reject the null when the p-value is less than 5%.
- Big $f \rightarrow$ REJECT!
- Small p-value $\rightarrow$ REJECT!

The F-test

In Excel, the p-value is reported under "Significance F"

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1147.985	191.3308	1.39747	0.257927747
Residual	23	3148.982	136.9123		
Total	29	4296.967			

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	3147.966342	524.6611	10.50235	1.24041E-05
Residual	23	1149.000325	49.95654		
Total	29	4296.966667			

The F-test

Note that f is also equal to (you can check the math!)

$$f = \frac{SSR/p}{SSE/(n-p-1)}$$

In Excel, the values under MS are SSR/p and $SSE/(n-p-1)$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1147.985	191.3308	1.39747	0.257927747
Residual	23	3148.982	136.9123		
Total	29	4296.967			

$$f = \frac{191.33}{136.91} = 1.39$$

Partial F-tests

- What about fitting a reduced model with only a couple of X's? In other words, do we need all of the X's to explain Y?
- For example, in the Supervisor data we could argue that X_1 and X_3 were the most important variables in predicting Y.
- The full model (6 covariates) has $R^2_{full} = 0.733$ while the model with only X_1 and X_3 has $R^2_{rest} = 0.708$ (check that!)
- Can we make a decision based only in the R^2 calculations? NO!!

Partial F-test

With the total F-test, we were asking

"Is this regression worthwhile?"

Now, we're asking

"Is is useful to add these extra covariates to the regression?"

You **always** want to use the simplest model possible.

- Only add covariates if they are truly informative.

Partial F-test

Consider the regression model

$$Y = \beta_0 + \beta_1 X_1 \ldots + \beta_{p_{base}} X_{p_{base}} + \beta_{p_{base}+1} X_{p_{base}+1} \ldots \beta_{p_{full}} X_{p_{full}} + \varepsilon$$

Such that d_{base} is the number of covariates in the base (small) model and $p_{full} > p_{base}$ is the number in the full (larger) model.

The Partial F-test is concerned with the hypotheses

$H_0 : \beta_{p_{base}+1} = \beta_{p_{base}+2} = \ldots = \beta_{p_{full}} = 0$

$H_1 :$ at least one $\beta_j \neq 0$ for $j > p_{base}$.

Partial F-test

It turns out that under the null H_0 (i.e. base model is true),

$$
\begin{aligned}
f &= \frac{(R^2_{full} - R^2_{base})/(p_{full} - p_{base})}{(1 - R^2_{full})/(n - p_{full} - 1)} \\
&\sim F_{p_{full} - p_{base}, n - p_{full} - 1}
\end{aligned}
$$

That is, under the null hypothesis, the ratio of normalized $R^2_{full} - R^2_{base}$ (increase in R^2) and $1 - R^2_{full}$ has F-distribution with $p_{full} - p_{base}$ and $n - p_{full} - 1$ df.

- Big f means that $R^2_{full} - R^2_{base}$ is statistically significant.
- Big f means that at least one of the added X's is useful.

Supervisor Performance: Partial F-test

Back to our supervisor data; we want to test

$$H_0 : \beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$$
$$H_1 : \text{at least one } \beta_j \neq 0 \text{ for } j \in \{2, 4, 5, 6\}.$$

The F-stat is $f = \dfrac{(0.733 - .708)/(6-2)}{(1 - .733)/(30 - 6 - 1)} = \dfrac{0.00625}{0.0116} = 0.54$

This leads to a p-value of 0.71 ... What do we conclude?

Example: Detecting Sex Discrimination

Imagine you are a trial lawyer and you want to file a suit against a company for salary discrimination... you gather the following data...

```
 Gender    Salary
1    Male   32.0
2  Female   39.1
3  Female   33.2
4  Female   30.6
5    Male   29.0
... ... ...
208 Female  30.0
```

Detecting Sex Discrimination

You want to relate salary(Y) to gender(X)... how can we do that?

Gender is an example of a categorical variable. The variable gender separates our data into 2 groups or categories. The question we want to answer is: *"how is your salary related to which group you belong to..."*

Could we think about additional examples of categories potentially associated with salary?

- MBA education vs. not
- legal vs. illegal immigrant
- quarterback vs wide receiver

Detecting Sex Discrimination

We can use regression to answer these question but we need to recode the categorical variable into a dummy variable

	Gender	Salary	Sex
1	Male	32.00	1
2	Female	39.10	0
3	Female	33.20	0
4	Female	30.60	0
5	Male	29.00	1
...	...	...	
208	Female	30.00	0

Note: In Excel you can create the dummy variable using the formula:

=IF(Gender="Male",1,0)

Detecting Sex Discrimination

Now you can present the following model in court:

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

How do you interpret β_1?

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

β_1 is the male/female difference

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

The data gives us the "year hired" as a measure of experience...

```
   YrHired Gender Salary  Sex
1        92   Male  32.00   1
2        81 Female  39.10   0
3        83 Female  33.20   0
4        87 Female  30.60   0
5        92   Male  29.00   1
... ... ...
208      62 Female  30.00   0
```

Detecting Sex Discrimination

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

Regression Statistics	
Multiple R	0.700680156
R Square	0.490952681
Adjusted R	0.485986366
Standard E	8.070070757
Observatioı	208

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	12876.27	6438	98.8565	8.7642E-31
Residual	205	13350.84	65.13		
Total	207	26227.11			

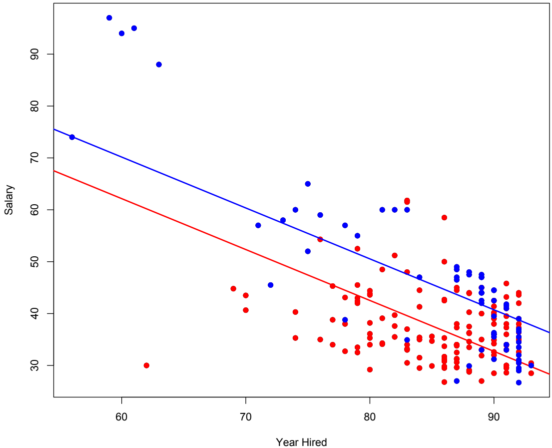
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	121.0212441	6.891851	17.56	9.8E-43	107.433246	134.6092
Gender	8.011885777	1.193089	6.715	1.8E-10	5.65958805	10.36418
YrHired	-0.981150947	0.080285	-12.22	3.7E-26	-1.1394402	-0.822862

$$Salary_i = 121 + 8 Sex_i - 0.98 Exp_i + \epsilon_i$$

Is this good or bad news for the defense?

Detecting Sex Discrimination

$$Salary_i = \begin{cases} 121 - 0.98 Exp_i + \epsilon_i & \text{females} \\ 129 - 0.98 Exp_i + \epsilon_i & \text{males} \end{cases}$$



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 *dn*1, *dn*2 and *dn*3...

	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

$$Price_i = \beta_0 + \beta_1 dn1 + \beta_2 dn2 + \beta_3 Size + \epsilon_i$$

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

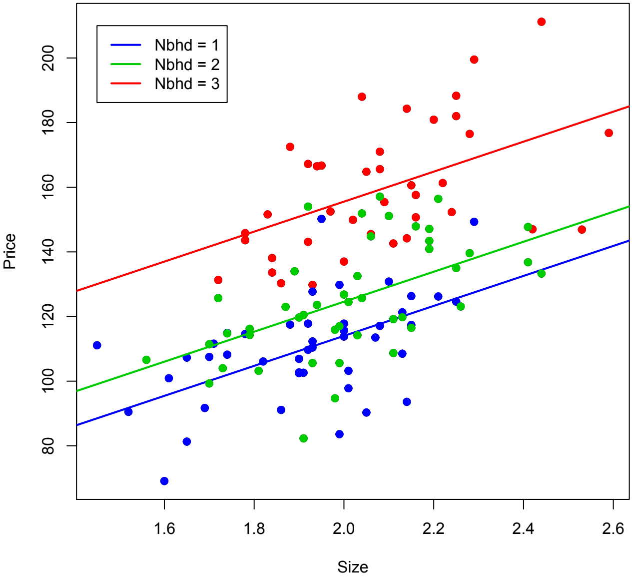
ANOVA

	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

$$Price_i = 62.78 - 41.54 dn1 - 30.97 dn2 + 46.39 Size + \epsilon_i$$

Example: House Prices



Example: House Prices

$$Price_i = \beta_0 + \beta_1 Size + \epsilon_i$$

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

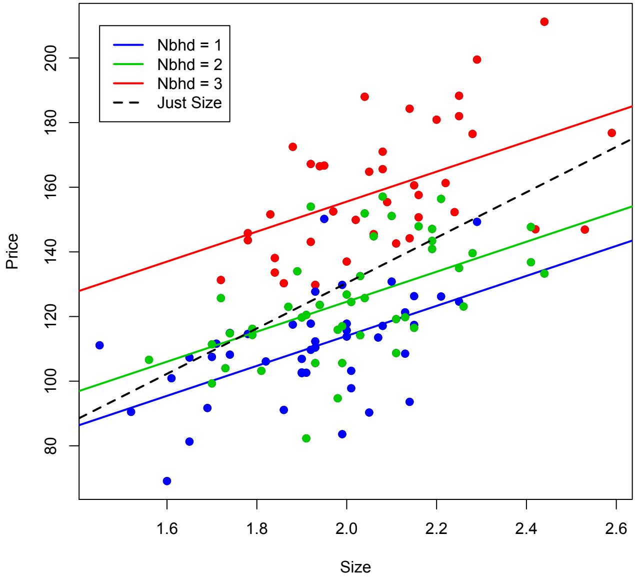
ANOVA

	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			

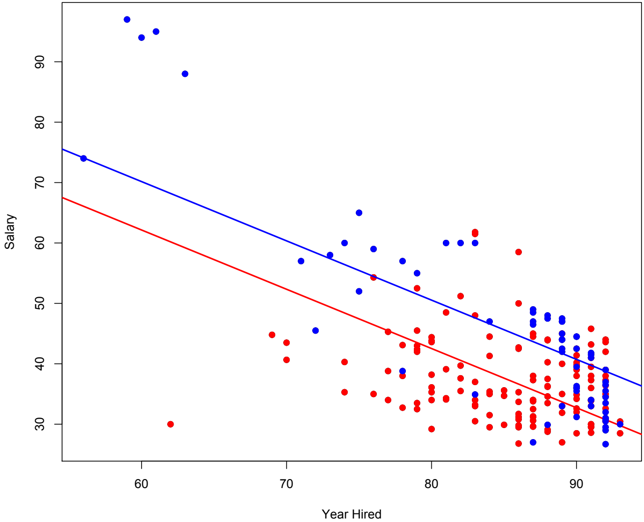
	Coefficients	*andard Err*	*t Stat*	*P-value*	*ower 95%*	*pper 95%*
Intercept	-10.09	18.97	-0.53	0.60	-47.62	27.44
size	70.23	9.43	7.45	0.00	51.57	88.88

$$Price_i = -10.09 + 70.23 Size + \epsilon_i$$

Example: House Prices



Back to the Sex Discrimination Case



Does it look like the effect of experience on salary is the same for males and females?

Back to the Sex Discrimination Case

Could we try to expand our analysis by allowing a different slope for each group?

Yes... Consider the following model:

$$Salary_i = \beta_0 + \beta_1 Exp_i + \beta_2 Sex_i + \beta_3 Exp_i \times Sex_i + \epsilon_i$$

For Females:

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

For Males:

$$Salary_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Exp_i + \epsilon_i$$

Sex Discrimination Case

How does the data look like?

	YrHired	Gender	Salary	Sex	SexExp
1	92	Male	32.00	1	92
2	81	Female	39.10	0	0
3	83	Female	33.20	0	0
4	87	Female	30.60	0	0
5	92	Male	29.00	1	92
...	...	...			
208	62	Female	30.00	0	62

Sex Discrimination Case

$$Salary_i = \beta_0 + \beta_1 Sex_i + \beta_2 Exp + \beta_3 Exp * Sex + \epsilon_i$$

Regression Statistics	
Multiple R	0.799130351
R Square	0.638609318
Adjusted R S	0.63329475
Standard Err	6.816298288
Observations	208

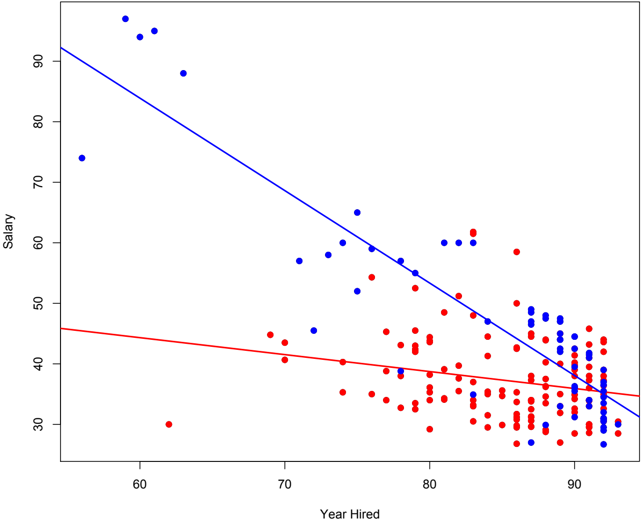
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	16748.88	5582.96	120.16	7.513E-45
Residual	204	9478.232	46.4619		
Total	207	26227.11			

	Coefficients	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	61.12479795	8.770854	6.96908	4E-11	43.831649	78.41795
Gender	114.4425931	11.7012	9.78041	9E-19	91.371794	137.5134
YrHired	-0.279963351	0.102456	-2.7325	0.0068	-0.4819713	-0.077955
GenderExp	-1.247798369	0.136676	-9.1296	7E-17	-1.5172765	-0.97832

$$Salary_i = 61 + 114 Sex_i + -0.27 Exp + -1.24 Exp * Sex + \epsilon_i$$

Sex Discrimination Case



Is this good or bad news for the plaintiff?

Variable Interaction

So, the effect of experience on salary is different for males and females... in general, when the effect of the variable X_1 onto Y depends on another variable X_2 we say that X_1 and X_2 interact with each other.

We can extend this notion by the inclusion of multiplicative effects through interaction terms.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} X_{2i}) + \varepsilon$$

$$\frac{\partial \mathbb{E}[Y|X_1, X_2]}{\partial X_1} = \beta_1 + \beta_3 X_2$$

We will pick this up in our next section...