

Section 4: Multiple Linear Regression

Tengyuan Liang, Chicago Booth

<https://tyliang.github.io/BUS41000/>
OpenIntro Statistics, Chapter 8
Statistics for Business, Part IV

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!
- ▶ Multi-factor asset pricing models (beyond CAPM)
- ▶ 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:

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

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

The MLR Model

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

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

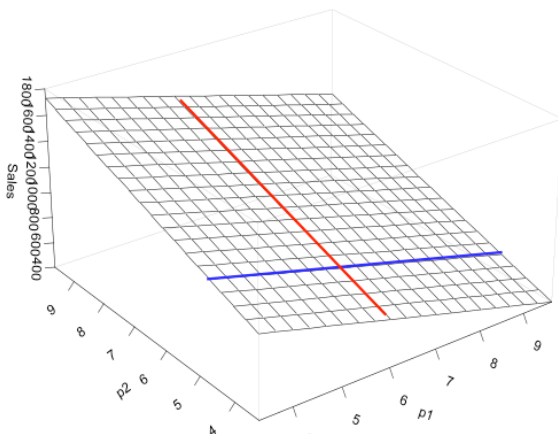
Holding all other variables constant, β_j is the average change in Y per unit change in X_j .

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 ($P2$).

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



Least Squares

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

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

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

| 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,

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

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

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

$$381 < \text{Sales} < 493$$

Least Squares

Just as before, each b_i is our estimate of β_i

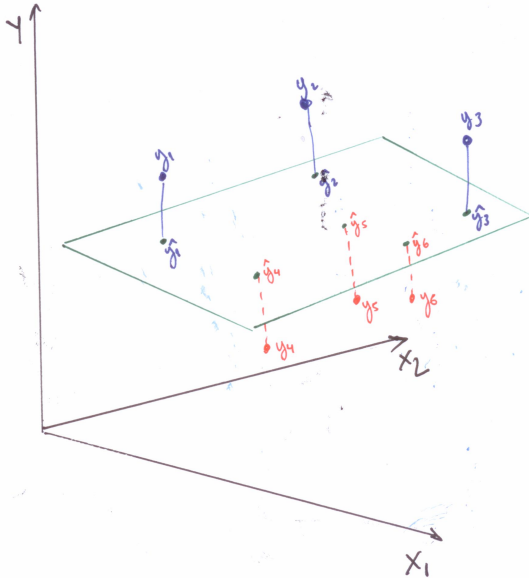
Fitted Values: $\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} \dots + b_pX_p.$

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

Least Squares: Find $b_0, b_1, b_2, \dots, b_p$ 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



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.

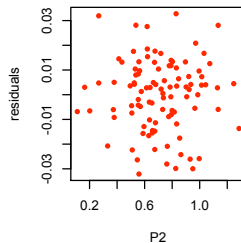
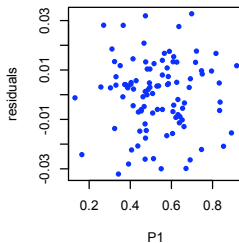
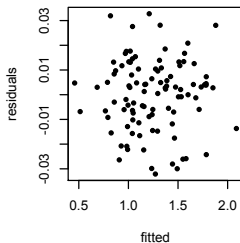
We decompose Y into the part predicted by X and the part due to idiosyncratic error.

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

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

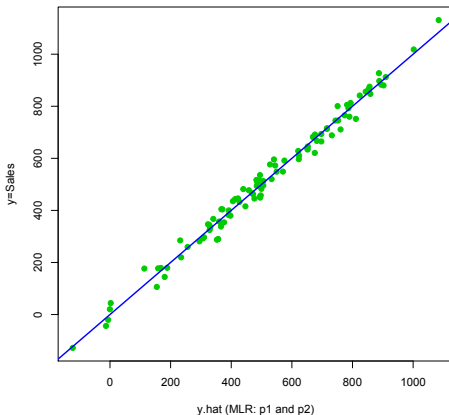
Residuals in MLR

Consider the residuals from the Sales data:



Fitted Values in MLR

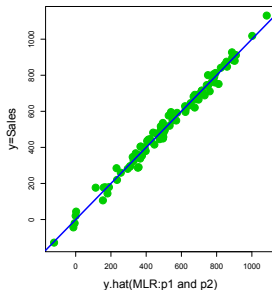
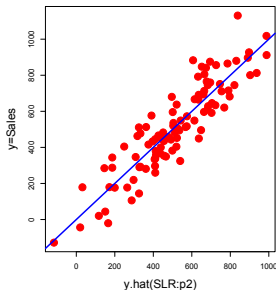
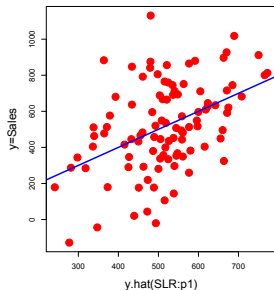
Another great plot for MLR problems is to look at \hat{Y} (true values) against $\hat{\hat{Y}}$ (fitted values).



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

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 .

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 | | | | | |
|------------|-----------|-----------|-----------|-----------|-----------------------|
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| Regression | 2 | 6.28747 | 3.143735 | 142.31576 | 4.56302E-15 |
| Residual | 27 | 0.596426 | 0.02209 | | |
| Total | 29 | 6.883896 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| 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 |

$$R^2 = 0.913$$

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

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 $j = 0, \dots, d$.

- 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 $|t_j| > 2$

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

| <i>Regression Statistics</i> | |
|------------------------------|--------|
| Multiple R | 0.99 |
| R Square | 0.99 |
| Adjusted R Square | 0.99 |
| Standard Error | 28.42 |
| Observations | 100.00 |

ANOVA

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|------------|------------|----------|-----------------------|
| Regression | 2.00 | 6004047.24 | 3002023.62 | 3717.29 | 0.00 |
| Residual | 97.00 | 78335.60 | 807.58 | | |
| Total | 99.00 | 6082382.84 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| 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 $\beta_1 \approx b_1 \pm 2 \times s_{b_1}$

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

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 the Overall Test of Significance... the F-test.
It will help us determine whether or not our regression is worth anything!

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

SUMMARY OUTPUT

| <i>Regression Statistics</i> | |
|------------------------------|-------------|
| Multiple R | 0.855921721 |
| R Square | 0.732601993 |
| Adjusted R Square | 0.662845991 |
| Standard Error | 7.067993765 |
| Observations | 30 |

ANOVA

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

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> | <i>Lower 99.0%</i> | <i>Upper 99.0%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|--------------------|--------------------|
| 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
- ▶ The Adjusted R^2 (but not a great one!) is an attempt at fixing the problem. . .

Garbage Data

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

SUMMARY OUTPUT

| <i>Regression Statistics</i> | |
|------------------------------|-------------|
| Multiple R | 0.516876852 |
| R Square | 0.26716168 |
| Adjusted R Square | 0.075986466 |
| Standard Error | 11.70095097 |
| Observations | 30 |

ANOVA

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-----------|-----------|----------|-----------------------|
| Regression | 6 | 1147.985 | 191.3308 | 1.39747 | 0.257927747 |
| Residual | 23 | 3148.982 | 136.9123 | | |
| Total | 29 | 4296.967 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> | <i>Lower 99.0%</i> | <i>Upper 99.0%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|--------------------|--------------------|
| 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:

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

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

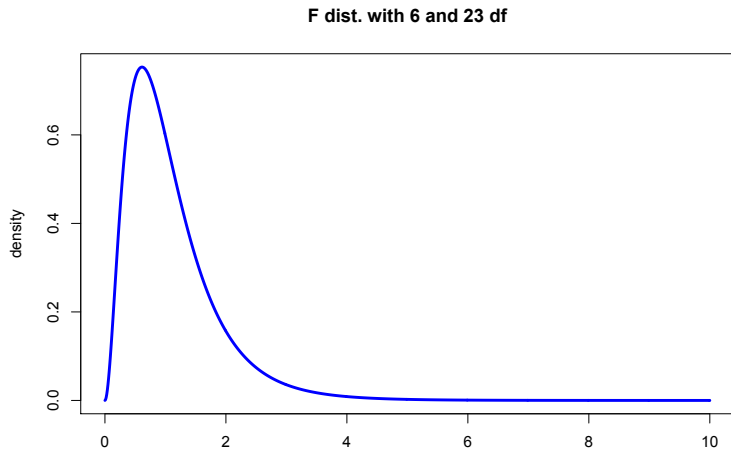
This is the F -test of overall significance. Under the null hypothesis f is distributed:

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

- 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

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-----------|-----------|----------|-----------------------|
| 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

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-------------|-----------|----------|-----------------------|
| 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

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-----------|-----------|----------|-----------------------|
| Regression | 6 | 1147.985 | 191.3308 | 1.39747 | 0.257927747 |
| Residual | 23 | 3148.982 | 136.9123 | | |
| Total | 29 | 4296.967 | | | |

ANOVA

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

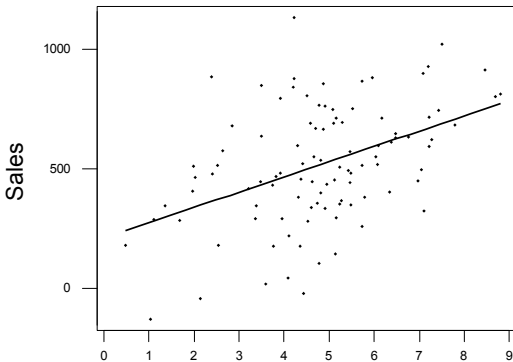
Understanding Multiple Regression

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 (P_1) is:

$$Sales = 211 + 63.7P_1$$

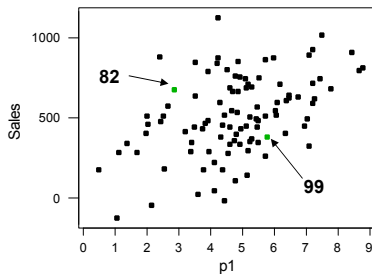
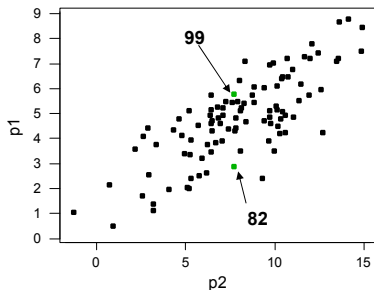
- ▶ If now we add the competitors price to the regression we get

$$Sales = 116 - 97.7P_1 + 109P_2$$

- ▶ Does this look better? How did it happen?
- ▶ Remember: -97.7 is the affect on sales of a change in P_1 **with P_2 held fixed!!**

Understanding Multiple Regression

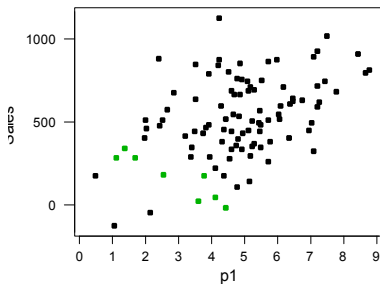
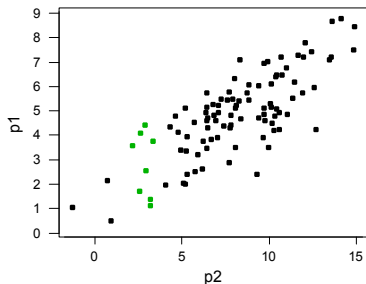
- ▶ How can we see what is going on? Let's compare Sales in two different observations: weeks 82 and 99.
- ▶ We see that an **increase** in $P1$, holding $P2$ **constant**, corresponds to a drop in Sales!



- ▶ 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. . .

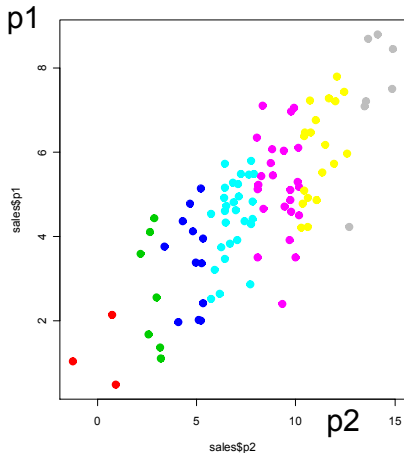


- ▶ For a fixed level of $P2$, variation in $P1$ is negatively correlated with Sales!!

Understanding Multiple Regression

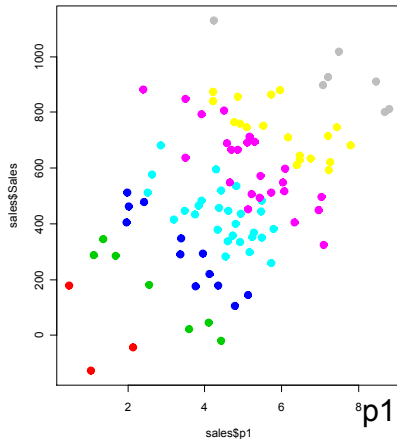
- Below, different colors indicate different ranges for $P2$...

larger $p1$ are associated with
larger $p2$



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

Sales



Understanding Multiple Regression

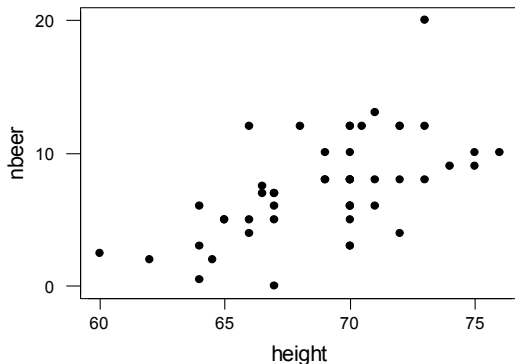
► Summary:

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

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

| <i>Regression Statistics</i> | |
|------------------------------|-------|
| Multiple R | 0.58 |
| R Square | 0.34 |
| Adjusted R Square | 0.33 |
| Standard Error | 3.11 |
| Observations | 50.00 |

| ANOVA | | | | | |
|------------|-----------|-----------|-----------|----------|-----------------------|
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| Regression | 1.00 | 237.77 | 237.77 | 24.60 | 0.00 |
| Residual | 48.00 | 463.86 | 9.66 | | |
| Total | 49.00 | 701.63 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| 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

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

| <i>Regression Statistics</i> | |
|------------------------------|-------|
| Multiple R | 0.69 |
| R Square | 0.48 |
| Adjusted R Square | 0.46 |
| Standard Error | 2.78 |
| Observations | 50.00 |

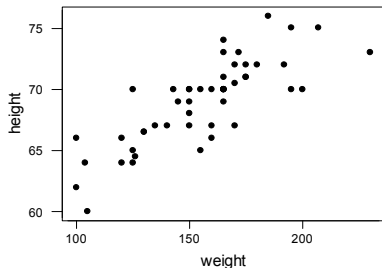
ANOVA

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-----------|-----------|----------|-----------------------|
| Regression | 2.00 | 337.24 | 168.62 | 21.75 | 0.00 |
| Residual | 47.00 | 364.38 | 7.75 | | |
| Total | 49.00 | 701.63 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| 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



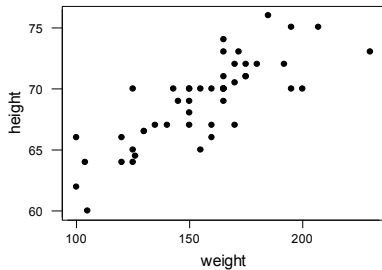
The correlations:

| | | |
|--------|-------|--------|
| | 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}



The correlations:

| | 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

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

| <i>Regression Statistics</i> | |
|------------------------------|------|
| Multiple R | 0.69 |
| R Square | 0.48 |
| Adjusted R | 0.47 |
| Standard E | 2.76 |
| Observatio | 50 |

ANOVA

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|-----------|-----------|-------------|-----------|----------|-----------------------|
| Regressor | 1 | 336.0317807 | 336.0318 | 44.11878 | 2.60227E-08 |
| Residual | 48 | 365.5932193 | 7.616525 | | |
| Total | 49 | 701.625 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| 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??

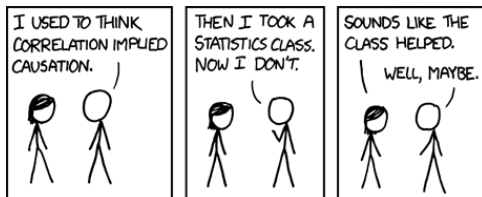
Understanding Multiple Regression

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

| <i>Regression Statistics</i> | |
|------------------------------|----------|
| Multiple R | 0.948136 |
| R Square | 0.898961 |
| Adjusted R Square | 0.891477 |
| Standard Error | 0.160502 |
| Observations | 30 |

ANOVA

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-----------|-----------|-------------|-----------------------|
| Regression | 2 | 6.188355 | 3.094177 | 120.1119098 | 3.63577E-14 |
| Residual | 27 | 0.695541 | 0.025761 | | |
| Total | 29 | 6.883896 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| 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 | Standard Error | 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. . .)

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. . . remember the trade-off between complexity and generalization (more on this later)