Section 3.2: Multiple Linear Regression II

Jared S. Murray
The University of Texas at Austin
McCombs School of Business

Multiple Linear Regression: Inference and Understanding

We can answer new questions with MLR:

- Are *any* of the independent variables predictive of the response?
- What's the effect of X_j *controlling for* other factors (other X's)?

Interpreting and understanding MLR is a little more complicated than SLR...

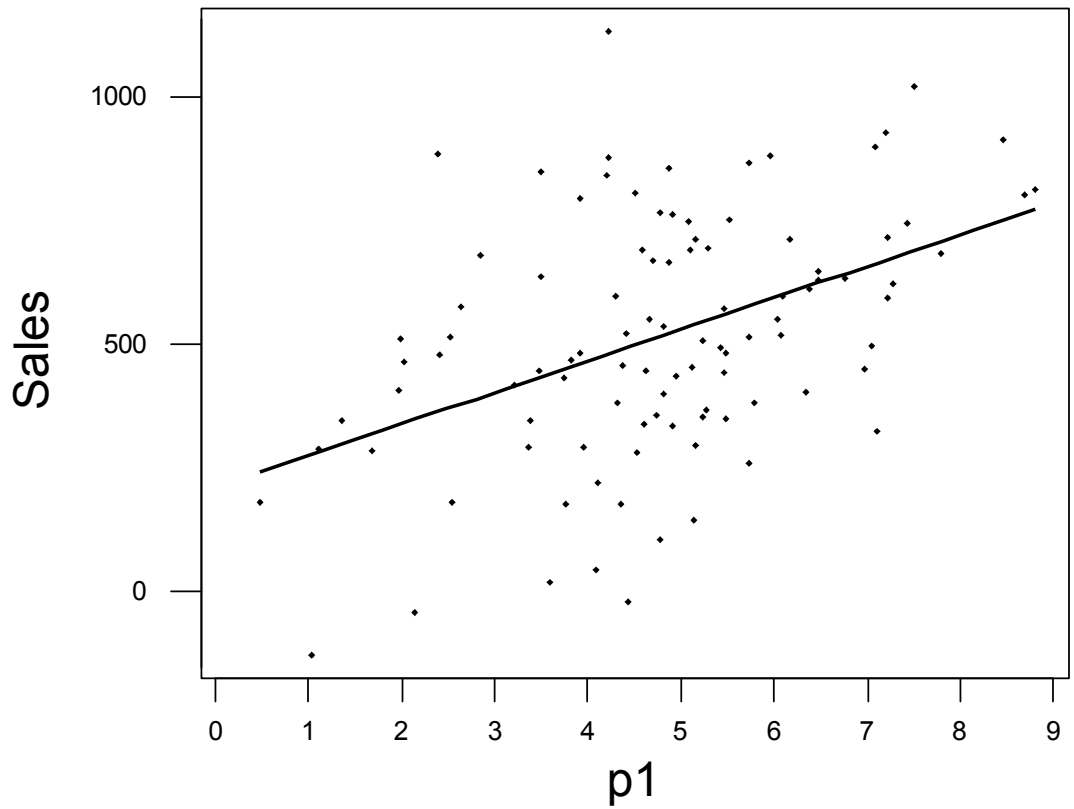
Understanding Multiple Regression

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 alone, we obtain a 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:

$$Sales = 211 + 63.7P1$$

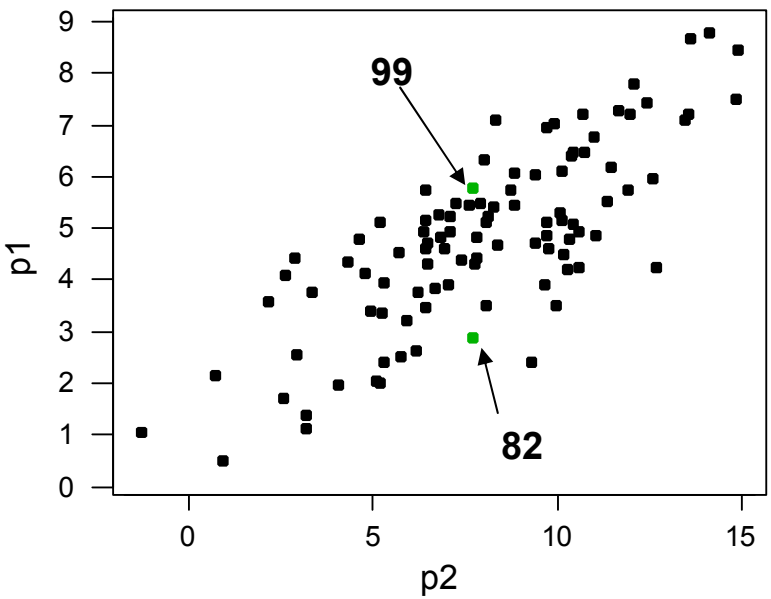
- If now we add the competitors price to the regression we get

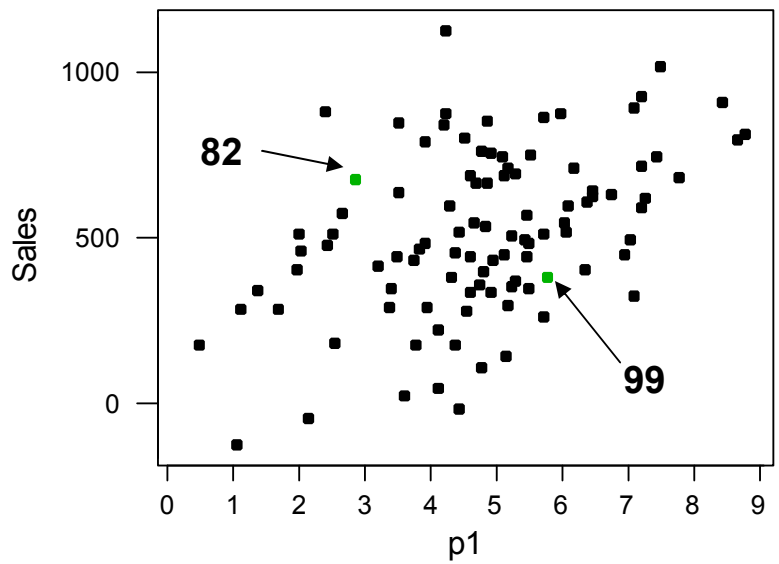
$$Sales = 116 - 97.7P1 + 109P2$$

- 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!

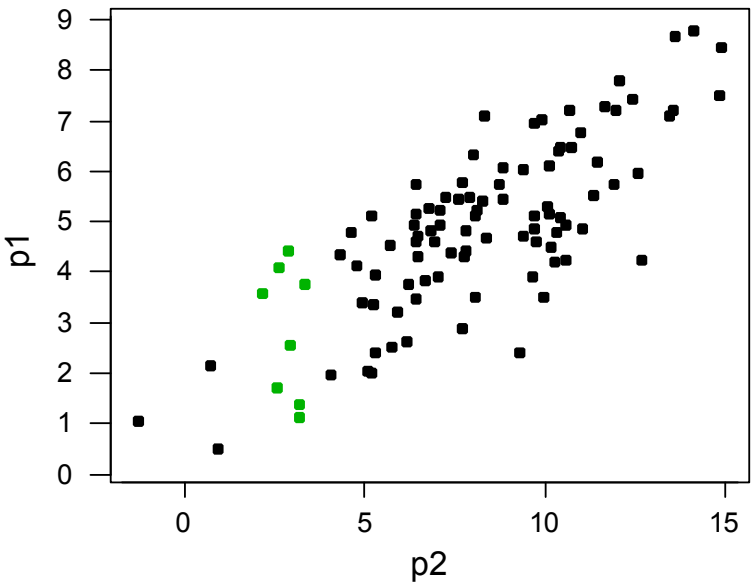


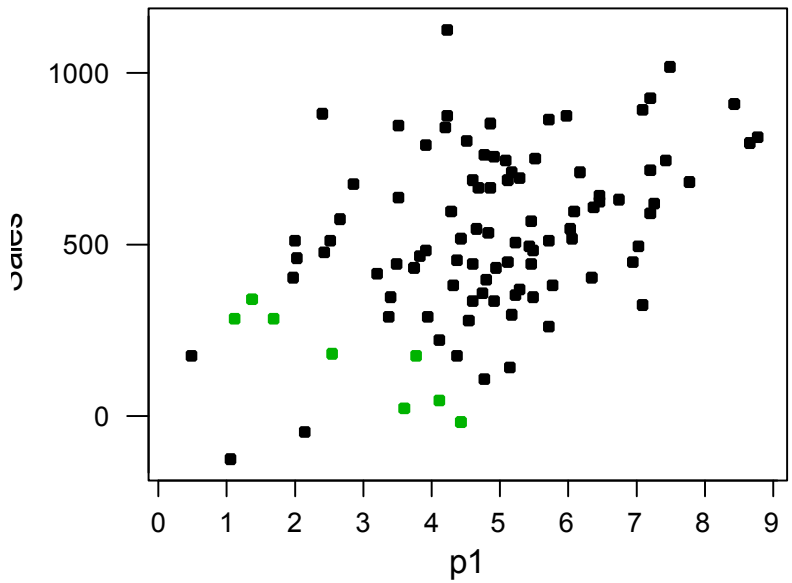


- Note the strong relationship (dependence) between $P1$ and $P2$!

Understanding Multiple Regression

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

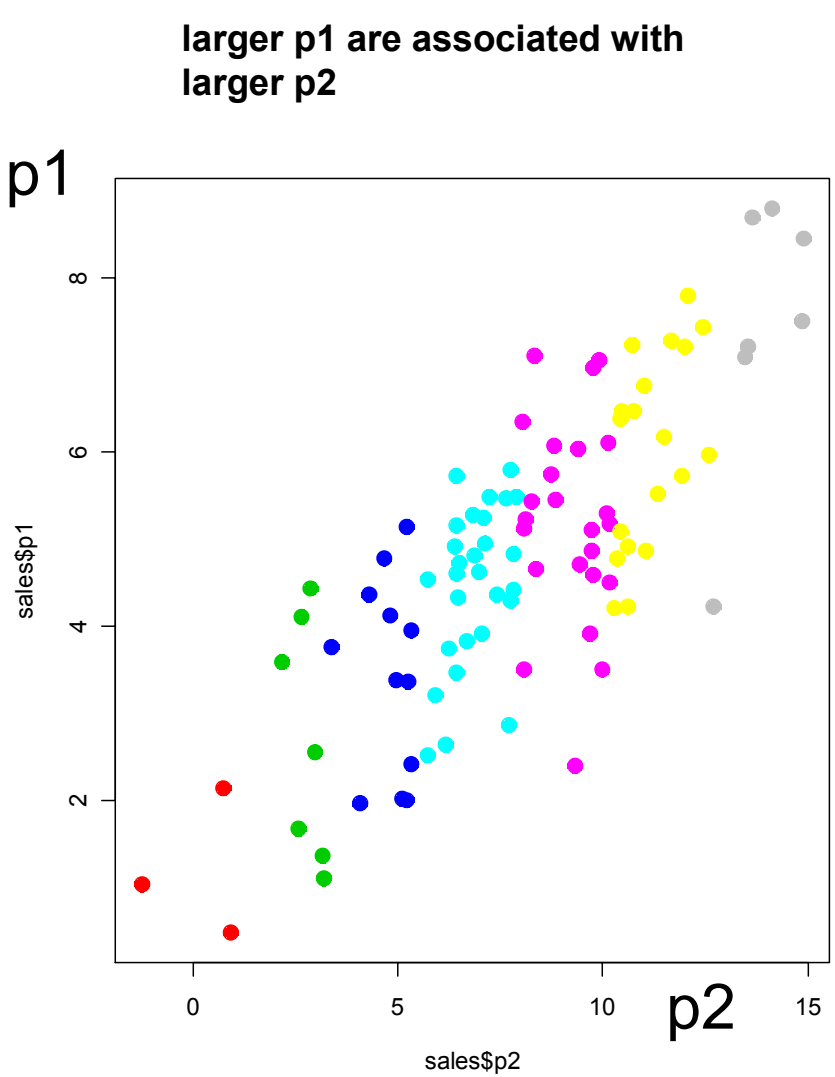


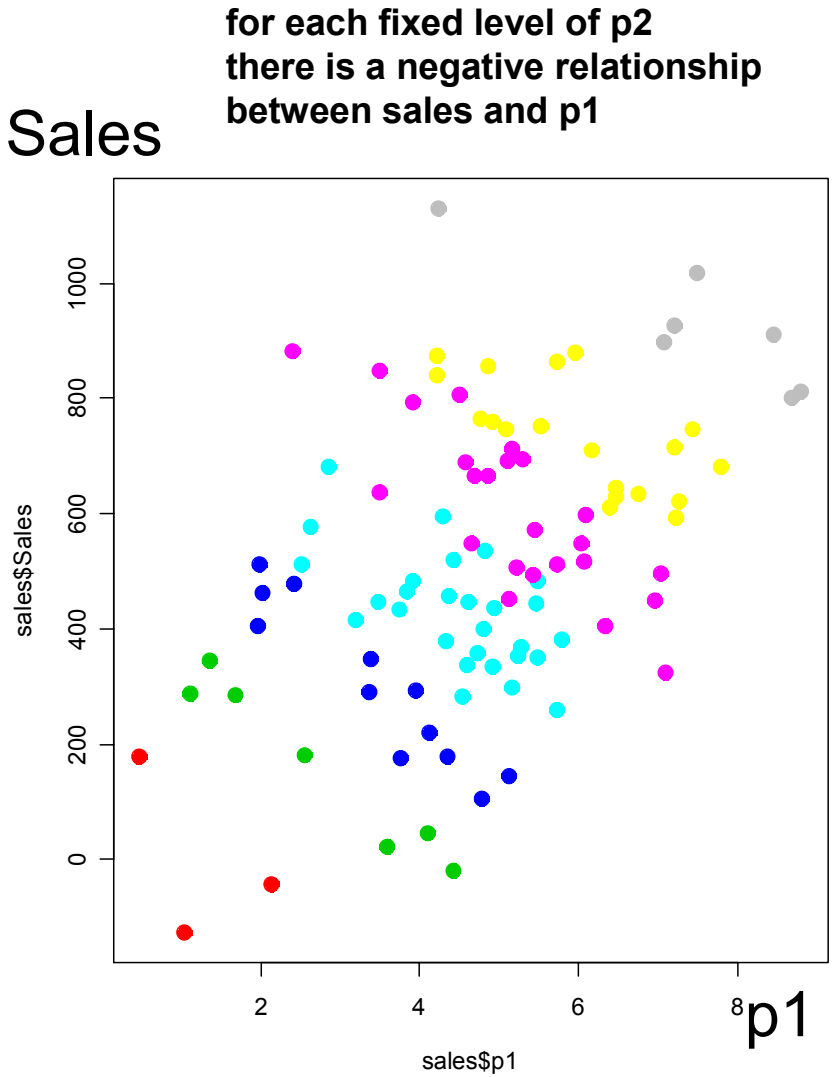


- For a fixed level of $P2$, variation in $P1$ is negatively correlated with Sales!!

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!

Confidence Intervals for Individual Coefficients

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

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

- The sampling distribution of each coefficient's estimator is

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

Confidence Intervals for Individual Coefficients

Computing confidence 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: $t_j = \dfrac{(b_j - \beta_j^0)}{s_{b_j}}$ 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 R... Do we know all of these numbers?

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  115.717      8.548   13.54   <2e-16 ***
## p1           -97.657      2.669  -36.59   <2e-16 ***
## p2           108.800      1.409   77.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.42 on 97 degrees of freedom
## Multiple R-squared:  0.9871,Adjusted R-squared:  0.9869
## F-statistic:  3717 on 2 and 97 DF,  p-value: < 2.2e-16
```

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

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

Confidence Intervals for Individual Coefficients

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

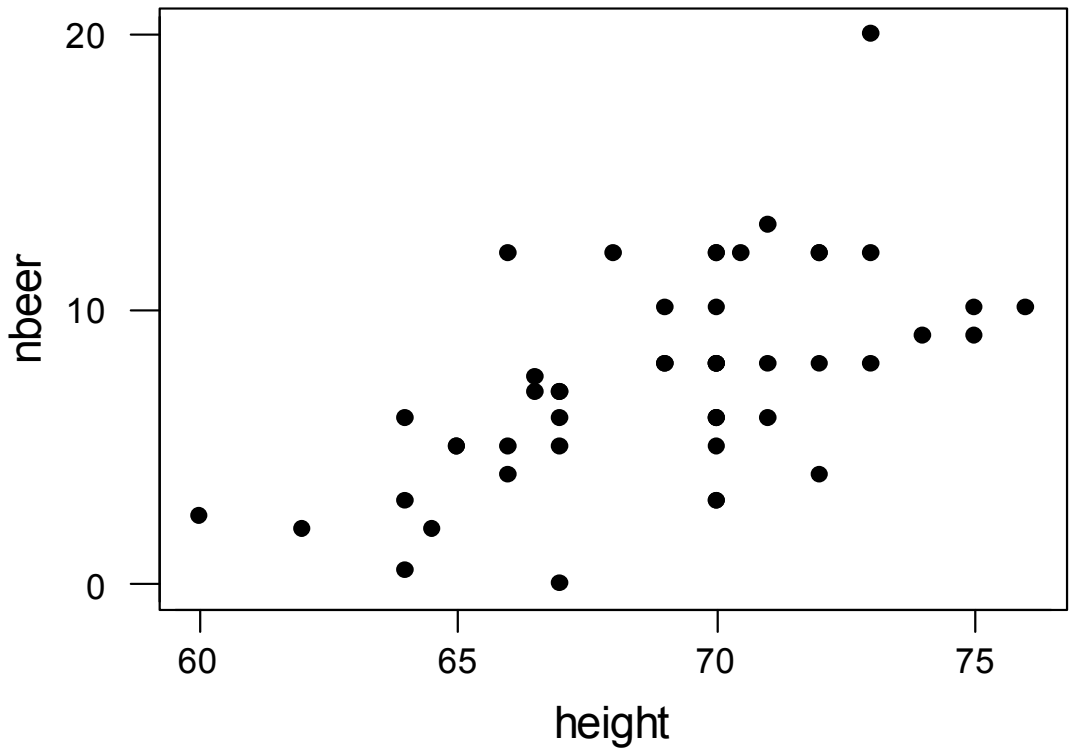
- You are evaluating the j^{th} coefficient conditional on the other X's being in the model, but regardless of the values you've estimated for the other b's.

Remember: β_j gives us the effect of a one-unit change in X_j, **holding the other X's in the model constant**.

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

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

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9200     8.9560  -4.122 0.000148 ***
## height        0.6430     0.1296   4.960 9.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.109 on 48 degrees of freedom
## Multiple R-squared:  0.3389,Adjusted R-squared:  0.3251
## F-statistic:  24.6 on 1 and 48 DF,  p-value: 9.23e-06
```

Yes! Beers and height are related...

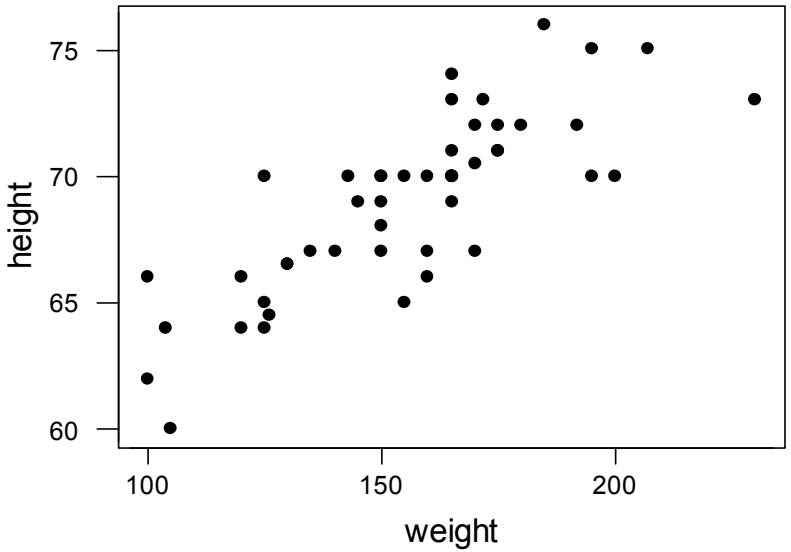
Understanding Multiple Regression

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

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.18709   10.76821  -1.039 0.304167
## height        0.07751    0.19598   0.396 0.694254
## weight        0.08530    0.02381   3.582 0.000806 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.784 on 47 degrees of freedom
## Multiple R-squared:  0.4807,Adjusted R-squared:  0.4586
## F-statistic: 21.75 on 2 and 47 DF,  p-value: 2.056e-07
```

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

Understanding Multiple Regression



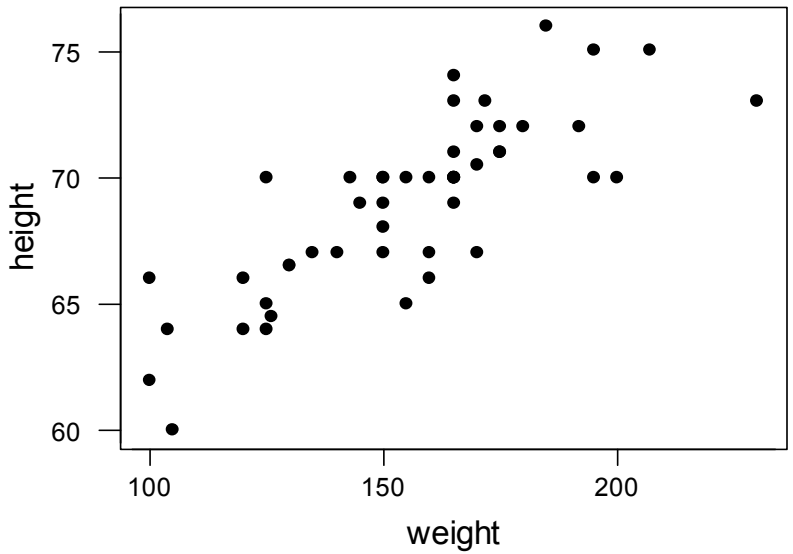
The correlations:

	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 $\rightarrow$ more beers, on average.
- 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 relevant 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

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

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.02070    2.21329  -3.172  0.00264 **
## weight       0.09289    0.01399   6.642  2.6e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.76 on 48 degrees of freedom
## Multiple R-squared:  0.4789,Adjusted R-squared:  0.4681
## F-statistic: 44.12 on 1 and 48 DF,  p-value: 2.602e-08
```

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

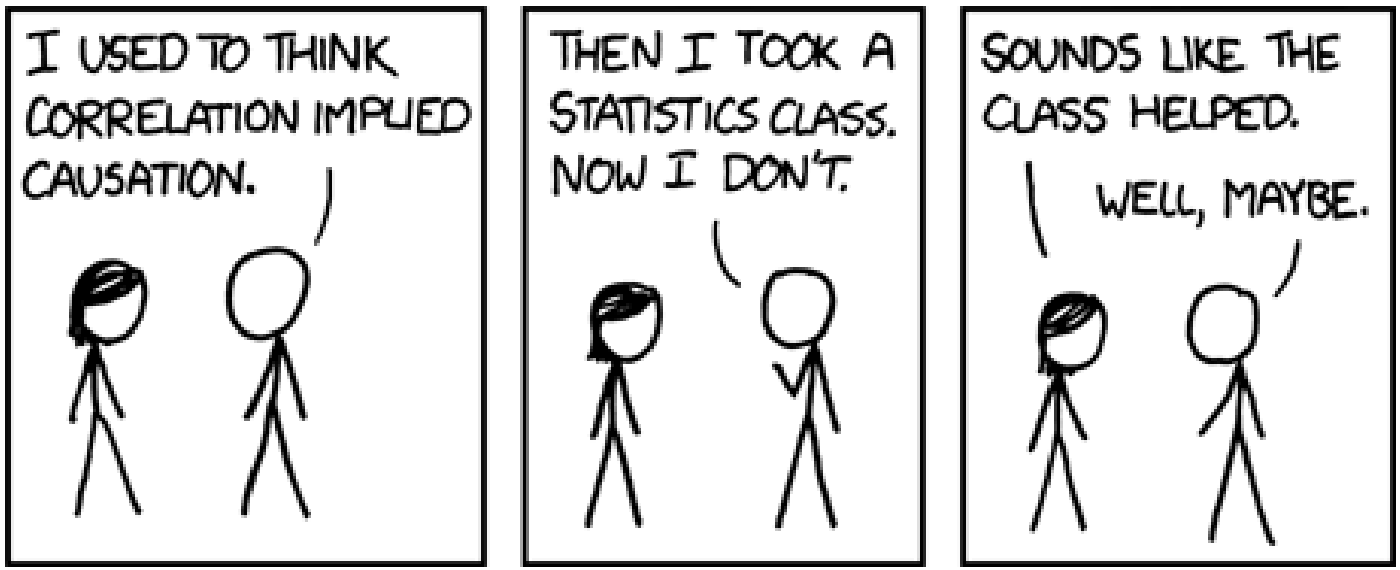
Understanding Multiple Regression

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`

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

$$s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2}$$

- The more variation in X (the larger s_x^2) the more "we know" about β_1... ie, our error $(b_1 - \beta_1)$ tends to be smaller.

Understanding Multiple Regression

- In MLR we relate the variation in Y to the variation in an X holding the other X's fixed. So, we need to know how much each X varies on its own.
- We can relate the standard errors in MLR to the standard errors from SLR: With two Xs,

$$s^2_{b_j} = \frac{1}{1 - r^2_{x_1 x_2}} \times \frac{s^2}{(n-1)s^2_{x_j}}$$

where $r_{x_1 x_2} = \text{cor}(x_1, x_2)$. The SE in MLR increases by a factor of $\frac{1}{1-r^2_{x_1 x_2}}$ relative to simple linear regression.

Understanding Multiple Regression

- In MLR we relate the variation in Y to the variation in an X holding the other X's fixed. So, we need to know how much each X varies on its own.
- In general, with p covariates,

$$s_{b_j}^2 = \frac{1}{1 - R_j^2} \times \frac{s^2}{(n-1)s_{x_j}^2}$$

  where R_j^2 is the R^2 from regressing X_j on the other X's.
- When there are strong dependencies between the covariates (known as **multicollinearity**), it is hard to attribute predictive ability to any of them individually.

Back to Baseball

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

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.0143     0.8199  -8.555 3.61e-09 ***
## OBP          27.5929     4.0032   6.893 2.09e-07 ***
## SLG           6.0311     2.0215   2.983  0.00598 **
```

Compare the std error to the model with OBP alone:

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.7816     0.8816  -8.827 1.40e-09 ***
## OBP          37.4593     2.5544  14.665 1.15e-14 ***
```

Even though s^2 is smaller in the MLR model (check it out!), the SE on OBP is higher than in SLR, since

```
cor(baseball$OBP, baseball$SLG)


## [1] 0.8261033
```

F-tests

- In many situations, 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

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.78708   11.58926   0.931 0.361634
## X1           0.61319    0.16098   3.809 0.000903 ***
## X2          -0.07305    0.13572  -0.538 0.595594
## X3           0.32033    0.16852   1.901 0.069925 .
## X4           0.08173    0.22148   0.369 0.715480
## X5           0.03838    0.14700   0.261 0.796334
## X6          -0.21706    0.17821  -1.218 0.235577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.068 on 23 degrees of freedom
## Multiple R-squared:  0.7326,Adjusted R-squared:  0.6628
## F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

Is there any relationship here at all? Which b_j's are significant?

Why not look at R^2

- R^2 in MLR ALWAYS grows as we increase the number of explanatory variables.
- Even if there is no relationship between the $X's$ and Y, $R^2 > 0$!!
- Adjusted R^2 is a (not great) attempt at fixing the problem
- 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...

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 63.95079    2.56337  24.948   <2e-16 ***
## G.1         -3.30589    2.28921  -1.444   0.1622
## G.2          2.82356    2.73411   1.033   0.3125
## G.3         -1.67550    2.20049  -0.761   0.4541
## G.4         -0.08067    2.74747  -0.029   0.9768
## G.5          3.61861    2.04390   1.770   0.0899 .
## G.6         -0.93827    2.27453  -0.413   0.6838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.81 on 23 degrees of freedom
## Multiple R-squared:  0.2536,Adjusted R-squared:  0.05889
## F-statistic: 1.302 on 6 and 23 DF,  p-value: 0.2955
```

Garbage Data

- R^2 is 0.25 !!
- We need to develop a way to see whether a R^2 of 0.25 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

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

Big $f \rightarrow$ big R^2 but we know what kind of f we are likely to get when all the coefficients are indeed zero (i.e., we know the probability distribution of f when all $\beta_j = 0$). We use this to decide if "big" is "big enough".

The F-test

We are testing:

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

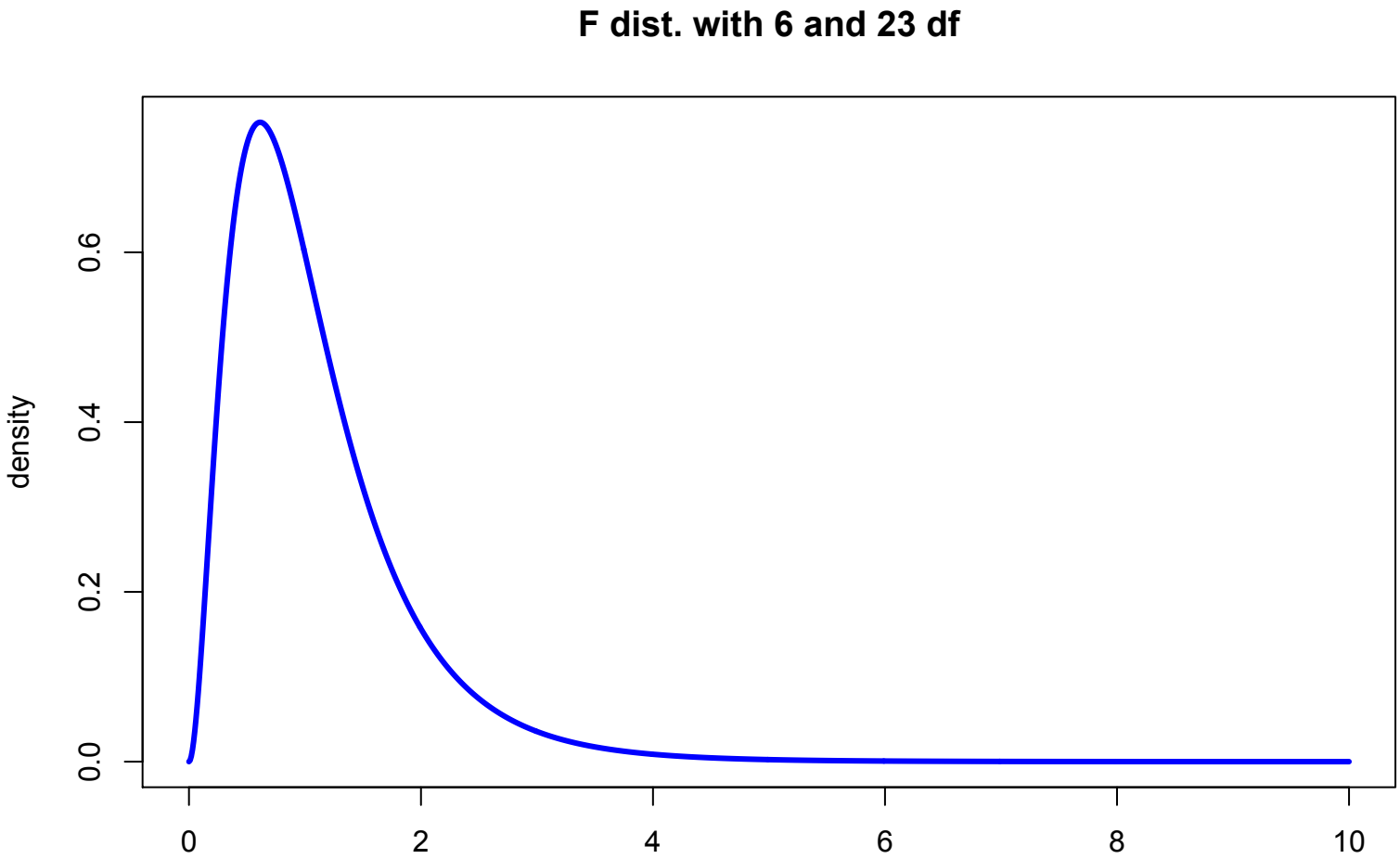
$$H_1 : \text{at least one } \beta_j \neq 0.$$

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

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

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 "degrees of freedom").

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!

In R, the last line of `summary` gives the F statistic and $p-$value.

The F-test

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

```
## Residual standard error: 11.81 on 23 degrees of freedom
## Multiple R-squared:  0.2536,Adjusted R-squared:  0.05889
## F-statistic: 1.302 on 6 and 23 DF,  p-value: 0.2955
```

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

```
## Residual standard error: 7.068 on 23 degrees of freedom
## Multiple R-squared:  0.7326,Adjusted R-squared:  0.6628
## F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

MLR: 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 the question... "is there an alternative explanation for this result"?
- Simple models are often better than very complex alternatives