Prediction

David Puelz
The University of Austin

Regression

Regression analysis is the most widely used statistical tool for understanding relationships among variables

It provides a conceptually simple method for investigating functional relationships between one or more factors and an outcome of interest

The relationship is expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory or predictor variables

Why?

Straight-up **prediction**:

- How much will I sell my house for?

Explanation and understanding:

- What is the impact of adding a bedroom to my sale price?

Example 1: Predicting house prices

Problem:

- Predict market price based on observed characteristics

Solution:

- Look at property sales data where we know the price and some observed characteristics.
- Build a decision rule that predicts price as a function of the observed characteristics.

Predicting house prices

Q: What characteristics do we use?

Predicting house prices

Q: What characteristics do we use?

We have to define the variables of interest and develop a specific quantitative measure of these variables ...

Many factors or variables affect the price of a house:

- size
- number of baths
- garage, air conditioning, etc
- neighborhood

Predicting house prices

To keep things super simple, let's focus only on size. The value

that we seek to predict is called the
dependent (or output) variable, and we denote this:

- Y = price of house (e.g. thousands of dollars)

The variable that we use to guide prediction is the
explanatory (or input) variable, and this is labeled

- X = size of house (e.g. thousands of square feet)

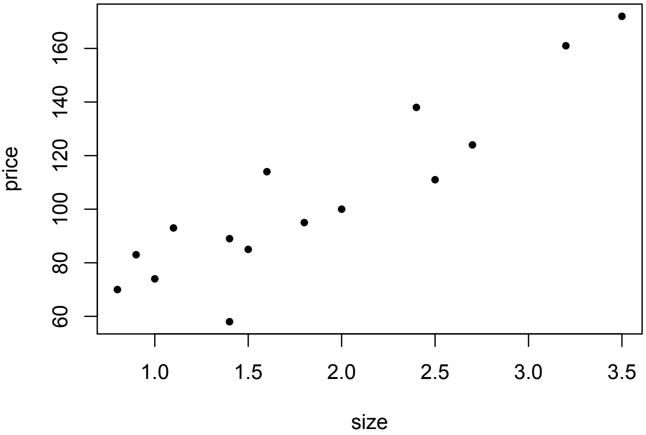
Predicting house prices

What does this data look like?

Size	Price
0.80	70
0.90	83
1.00	74
1.10	93
1.40	89
1.40	58
1.50	85
1.60	114
1.80	95
2.00	100
2.40	138
2.50	111
2.70	124
3.20	161
3.50	172

Predicting house prices

It is much more useful to look at a scatterplot

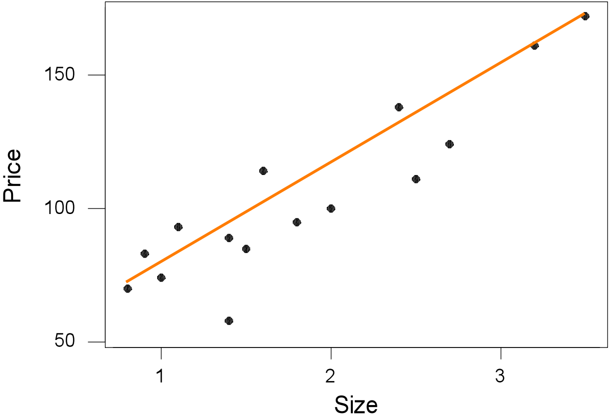


In other words, view the data as points in the $X \times Y$ plane.

Linear prediction

There seems to be a linear relationship between price and size:

As size goes up, price goes up.



Linear prediction

Recall that the equation of a line is:

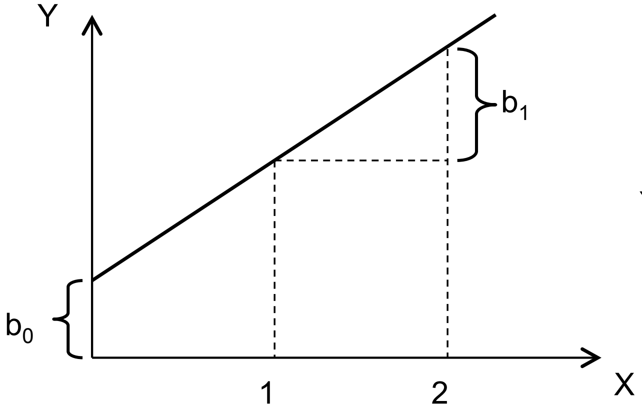
$$Y = b_0 + b_1 X$$

Where b_0 is the **intercept** and b_1 is the **slope**.

$\rightarrow$ The **intercept** value is in units of Y (\$1,000)

$\rightarrow$ The **slope** is in units of Y *per* units of X (\$1,000/1,000 sq ft)

Linear prediction



$Y = b_0 + b_1X$

Linear prediction

Q: How to find the "best line"?

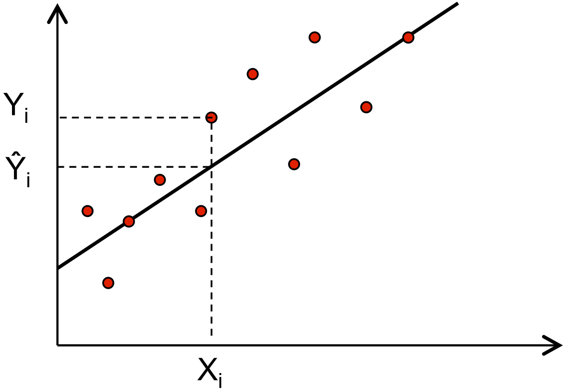
We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

A reasonable way to fit a line is to minimize the amount by which the fitted value differs from the actual value.

This amount is called the residual.

Linear prediction

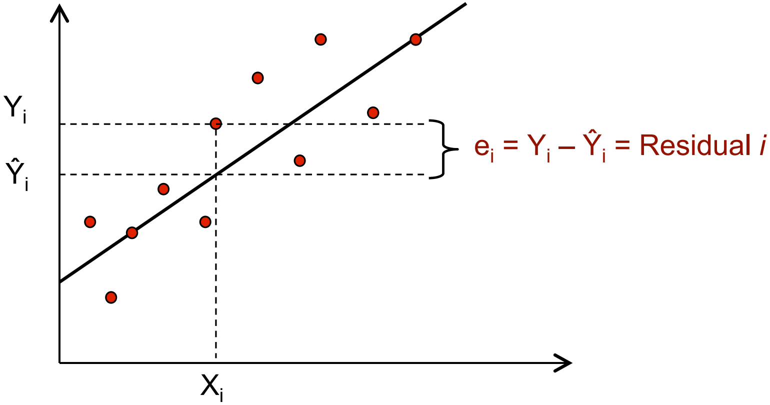
What is the "fitted value"?



The dots are the observed values and the line represents our fitted values given by $\hat{Y}_i = b_0 + b_1 X_1$.

Linear prediction

What is the "residual"' for the *i*th observation?



We can write $Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$.

Least squares

Ideally, we want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

Least squares

Ideally, we want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

The line fitting process:

- Minimize the “total” of residuals to get best fit.

Least squares

Ideally, we want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

The line fitting process:

- Minimize the "total" of residuals to get best fit.

  Least Squares chooses b_0 and b_1 to minimize $\sum_{i=1}^{N} e_i^2$

$$\sum_{i=1}^{N} e_i^2 = e_1^2 + e_2^2 + \cdots + e_N^2 = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \cdots + (Y_N - \hat{Y}_N)^2$$

Example 2: Offensive performance in baseball

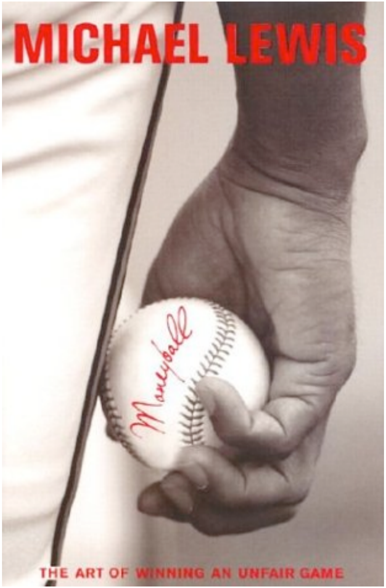
Problems:

- Evaluate/compare traditional measures of offensive performance
- Help evaluate the worth of a player

Solutions:

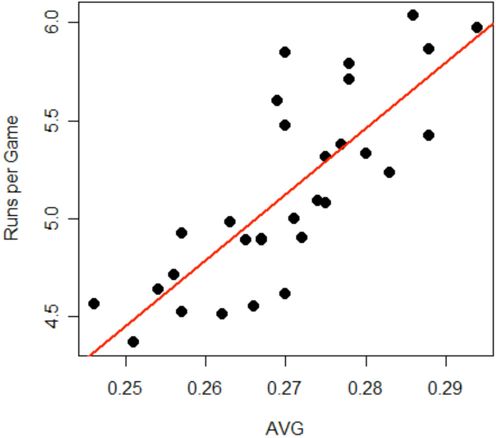
- Compare *prediction rules* that forecast runs as a function of either AVG (batting average), SLG (slugging percentage – total bases divided by at bats) or OBP (on base percentage)

Example 2: Offensive performance in baseball



Baseball data – using AVG

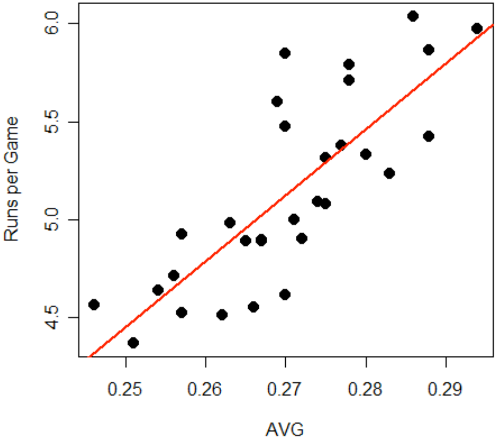
Each observation corresponds to a team in MLB. Each quantity is the average over a season.



Y = runs per game; X = AVG (average)

LS fit: Runs/Game = -3.93 + 33.57 AVG

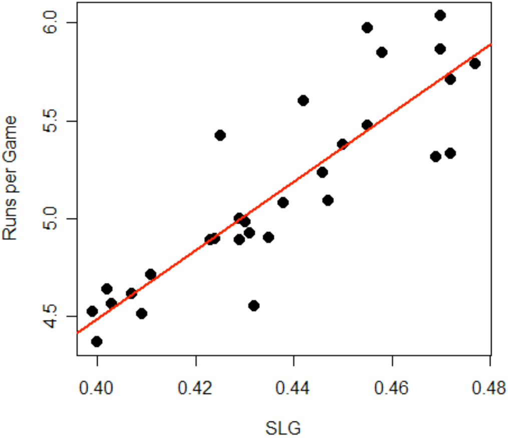
Baseball data – using AVG



Y = runs per game; X = AVG (average)

LS fit: Runs/Game = -3.93 + 33.57 AVG

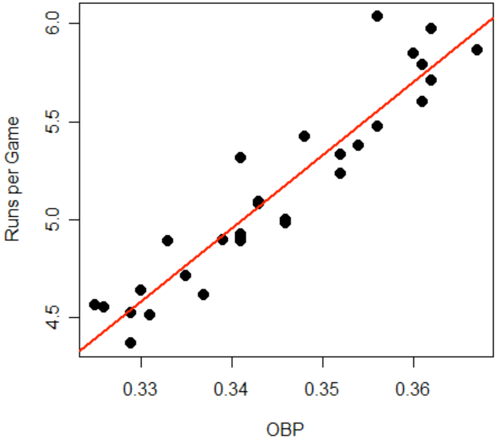
Baseball data – using SLG



Y = runs per game; X = SLG (slugging percentage)

LS fit: Runs/Game = -2.52 + 17.54 SLG

Baseball data – using OBP



Y = runs per game; X = OBP (on base percentage)

LS fit: Runs/Game = -7.78 + 37.46 OBP

Baseball data

- What is the best prediction rule?

- Let's compare the predictive ability of each model using the average squared error

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} = \left(\frac{\sum_{i=1}^{N}\left(\widehat{\text{Runs}_i} - \text{Runs}_i\right)^2}{N}\right)^{\frac{1}{2}}$$

Place your money on OBP!

	Root Mean Squared Error
AVG	0.29
SLG	0.23
OBP	0.16

More on least squares

Remember how we get the slope (b_1) and intercept (b_0). We minimize the sum of squared prediction errors.

The formulas for b_0 and b_1 that minimize the least squares criterion are:

$$b_1 = r_{xy} \times \frac{s_y}{s_x} \qquad b_0 = \bar{Y} - b_1 \bar{X}$$

where,

- $\bar{X}$ and $\bar{Y}$ are the sample mean of X and Y
- $\text{corr}(x, y) = r_{xy}$ is the sample correlation
- s_x and s_y are the sample standard deviation of X and Y

What are these numbers in the formula?

- Sample Mean: measure of centrality

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- Sample Variance: measure of spread

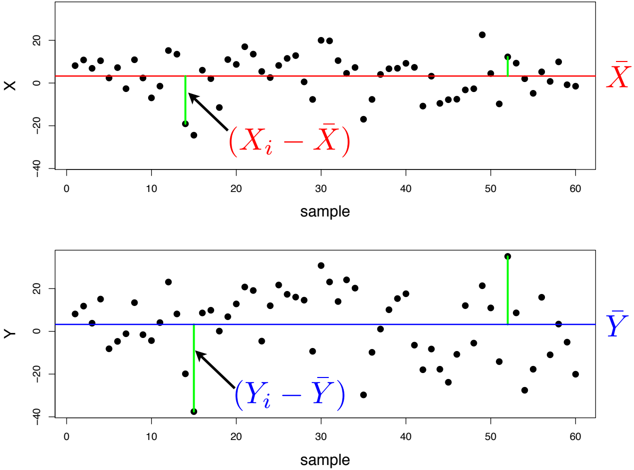
$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

- Sample Standard Deviation:

$$s_y = \sqrt{s_y^2}$$

Visual: standard deviation

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2$$

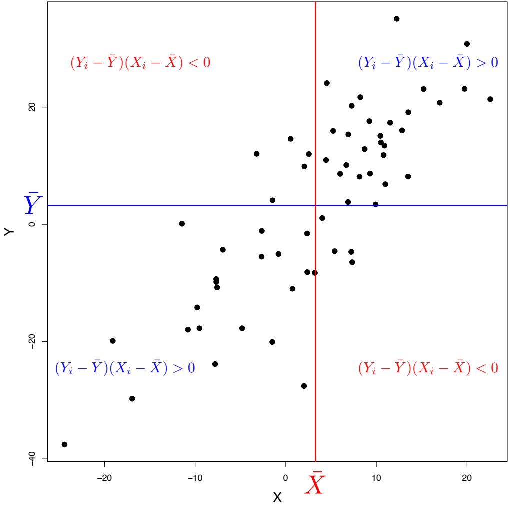


$s_x = 9.7$ $s_y = 15.98$

Visual: Covariance

Measure the **direction** and **strength** of the linear relationship between Y and X

$$\mathrm{cov}(Y,X) = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{n-1}$$



- $s_y = 15.98,\ s_x = 9.7$
- $\mathrm{cov}(X, Y) = 125.9$

How do we interpret that?

A standardized measure: Correlation

Correlation is the standardized covariance:

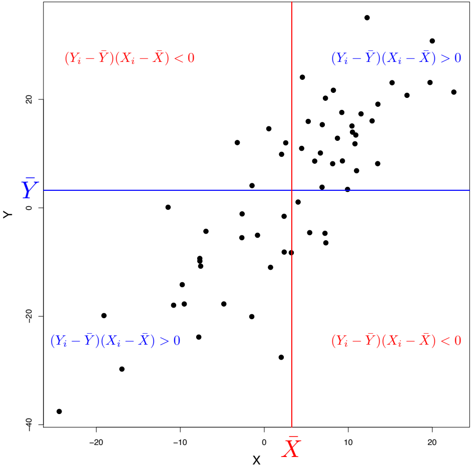
$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\mathrm{cov}(X, Y)}{s_x s_y}$$

The correlation is scale invariant and the units of measurement don't matter: It is always true that $-1 \leq \mathrm{corr}(X, Y) \leq 1$.

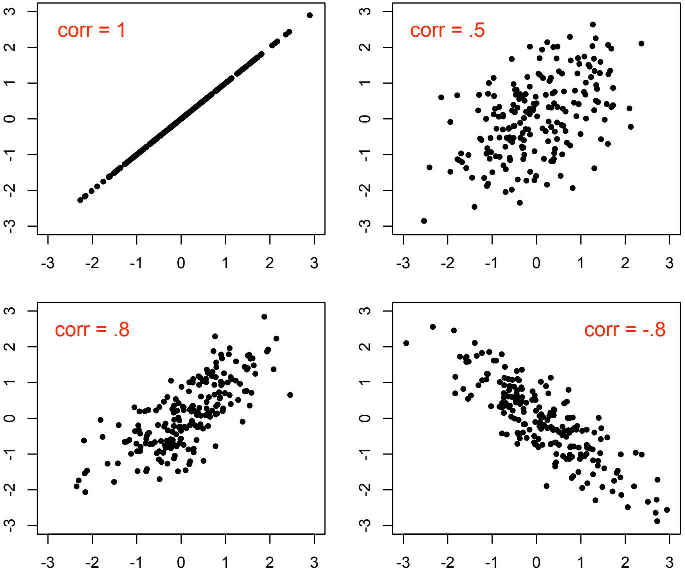
This gives the direction (negative or positive) and strength ($0 \rightarrow 1$) of the linear relationship between X and Y.

Correlation

$$\text{corr}(Y, X) = \frac{\text{cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\text{cov}(X, Y)}{s_x s_y} = \frac{125.9}{15.98 \times 9.7} = 0.812$$



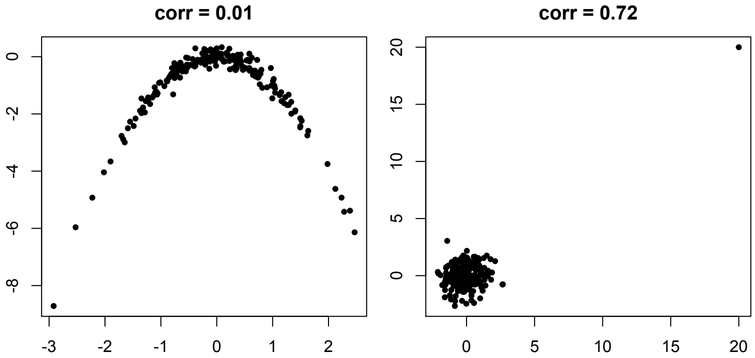
Correlation



Correlation

Only measures linear relationships:

$\text{corr}(X, Y) = 0$ does not mean the variables are not related!



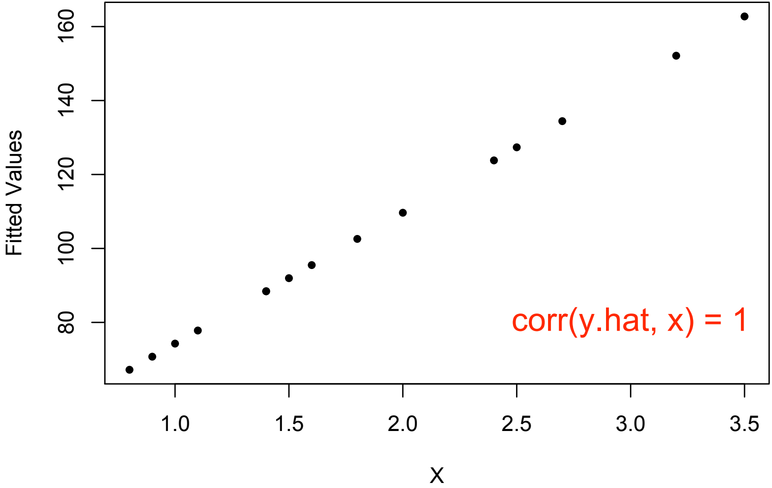
Also be careful with influential observations. Check out `cor()` in `R`.

More on least squares

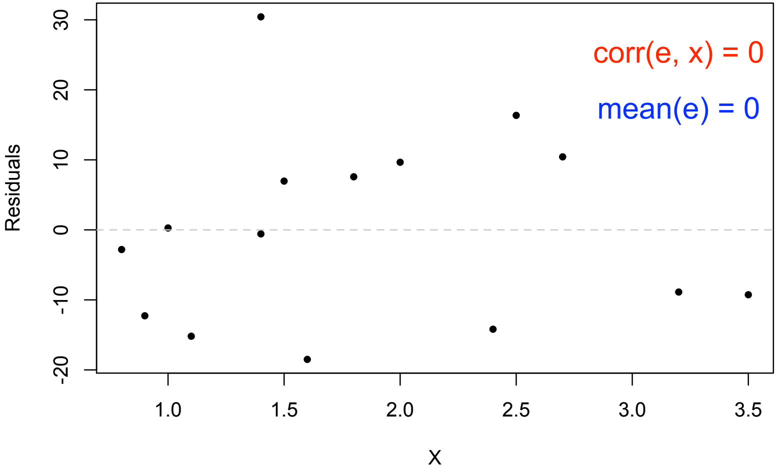
From now on, terms "fitted values" ($\hat{Y}_i$) and "residuals" (e_i) refer to those obtained from the least squares line.

The fitted values and residuals have some special properties. Let's look at the housing data analysis to figure out what these properties are...

The fitted values and X

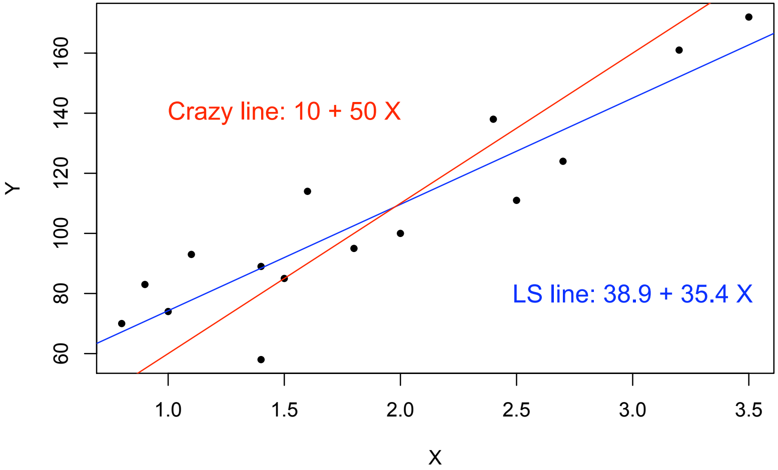


The residuals and X



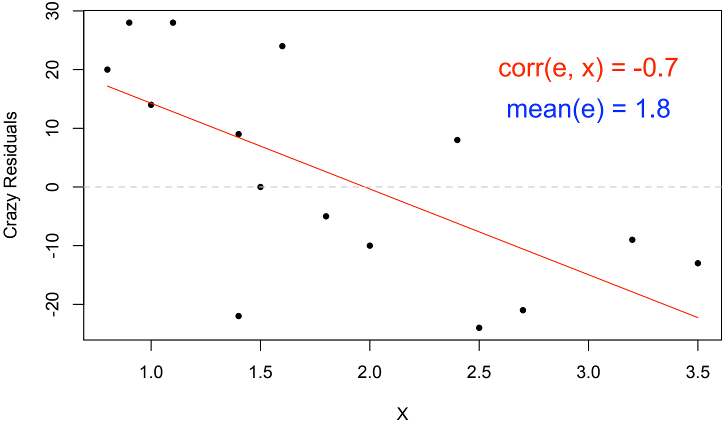
Why?

What is the intuition for the relationship between $\hat{Y}$ and e and X? Lets consider some "crazy"alternative line:



Fitted values and residuals

This is a bad fit! We are underestimating the value of small houses and overestimating the value of big houses.



Clearly, we have left some predictive ability on the table!

Fitted values and residuals

As long as the correlation between e and X is non-zero, we could always adjust our prediction rule to do better.

We need to exploit all of the predictive power in the X values and put this into $\hat{Y}$, leaving no "*Xness*" in the residuals.

In summary: $Y = \hat{Y} + e$ where:

- $\hat{Y}$ is "made from X"; $\text{corr}(X, \hat{Y}) = 1$.
- e is unrelated to X; $\text{corr}(X, e) = 0$.

Multiple linear regression

Multiple linear regression

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, household 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 structural assumption of our linear regression model:

$\rightarrow$ The conditional mean of Y is linear in the X_j variables.

The MLR model

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

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

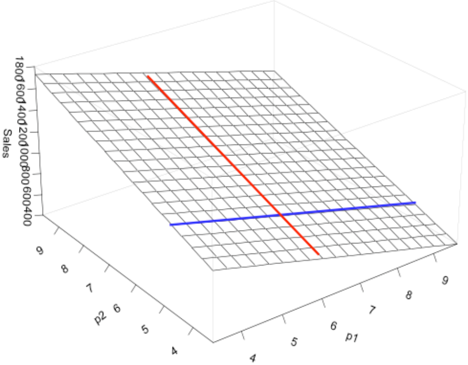
Holding all other variables constant, β_j is the average change in Y per unit change in X_j.

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 again!

$$Y = \beta_0 + \beta_1 X_1 \ldots + \beta_p X_p + \varepsilon$$

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 again!

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
...	...	...

The model: $\text{Sales}_i = \beta_0 + \beta_1 P1_i + \beta_2 P2_i + \epsilon_i$

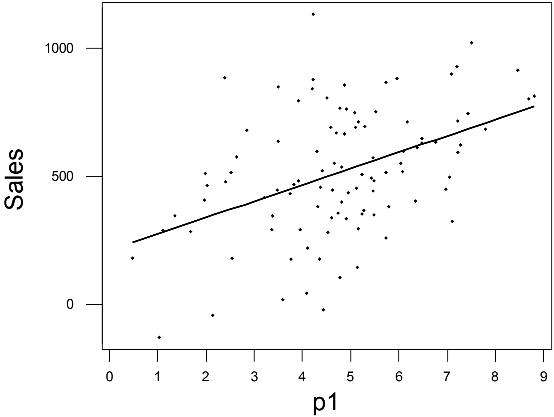
But what about the right-hand-side?

It is also important to understand and interpret the coefficients, i.e., what is is happening on the "right-hand-side" of our model ...

- **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)

But what about the right-hand-side?

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?

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

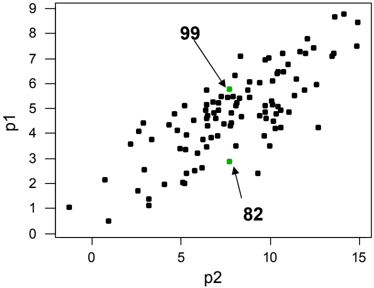
$$Sales = 116 - 97.7P1 + 109P2$$

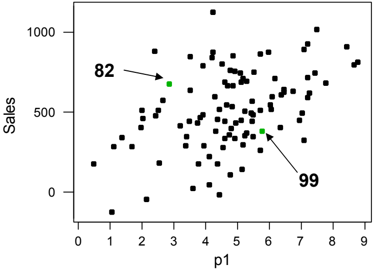
Does this look better? How did it happen? Remember: -97.7 is the affect on sales of a change in $P1$ **with $P2$ held fixed!**

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!

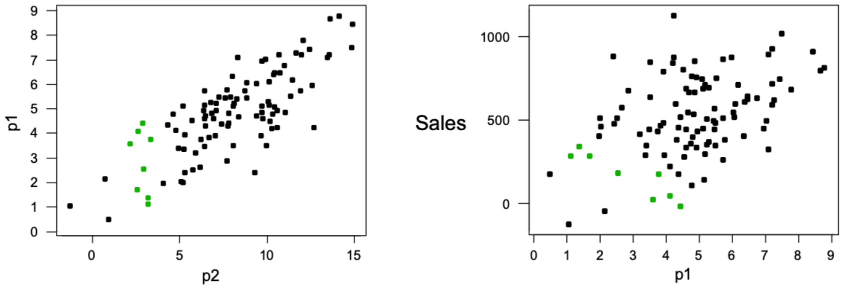




Note the strong relationship (dependence) between $P1$ and $P2$!

Understanding multiple regression

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

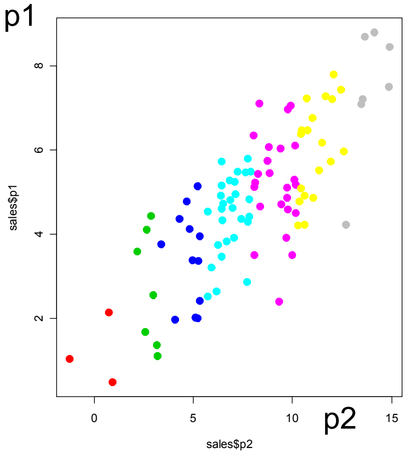


For a fixed level of $P2$, variation in $P1$ is negatively correlated with Sales!

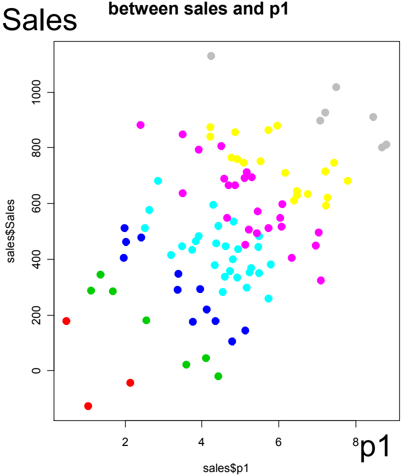
Understanding multiple regression

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:

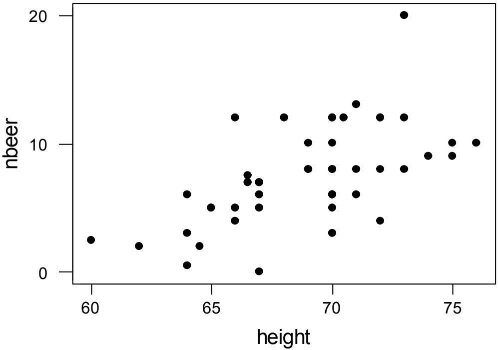
→ A larger $P1$ is associated with larger $P2$ and the overall effect leads to bigger sales

→ With $P2$ held fixed, a larger $P1$ leads to lower sales

→ MLR does the trick and unveils the **correct** economic relationship between Sales and prices!

Example: Beers, height, weight, and getting drunk

Beer data (from PRL class last year)
– **nbeer** – number of beers before getting drunk
– **height** and **weight**



Is number of beers related to height?

R output: Yes!

```
data = read.csv('nbeer.csv')
fit = lm(nbeer~height,data)
summary(fit)
```

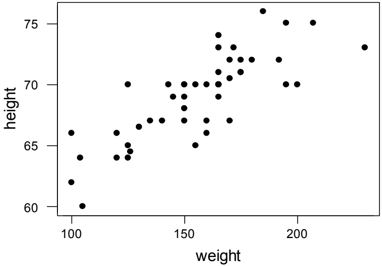
```
##
## Call:
## lm(formula = nbeer ~ height, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.164 -2.005 -0.093  1.738  9.978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9200     8.9560  -4.122 0.000148 ***
## height        0.6430     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
```

R output: What about now?

```
data = read.csv('nbeer.csv')
fit = lm(nbeer~height+weight,data)
summary(fit)
```

```
##
## Call:
## lm(formula = nbeer ~ height + weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5080 -2.0269  0.0652  1.5576  5.9087
##
## 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
```

Understanding multiple regression



<u>The correlations:</u>

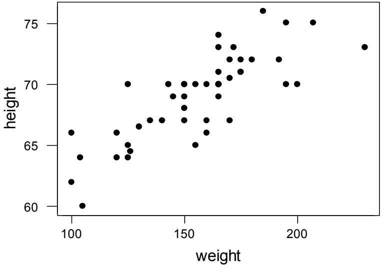
	nbeer	weight
weight	0.692	
height	0.582	0.806

The two x's are highly correlated !!

If we regress "beers" only on height we see an effect. Taller 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.

R output: Why is this a better model than height + weight?

```
data = read.csv('nbeer.csv')
fit = lm(nbeer~weight,data)
summary(fit)
```

```
##
## Call:
## lm(formula = nbeer ~ weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7709 -2.0304 -0.0742  1.6580  5.6556
##
## 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
```

Summary slide

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