Prediction

Outline

Regression: General introduction

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 variable

Why?

Straight-up **prediction**:

- How much will I sell my house for?

Explanation and understanding:

- What is the impact of economic freedom on growth?

Example: 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?

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)

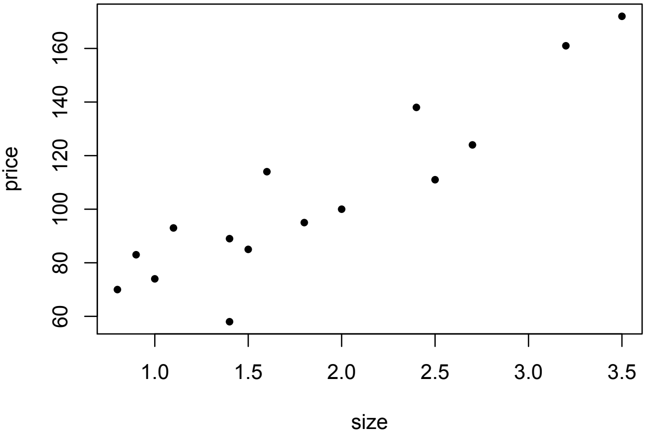
Predicting house prices

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.

Regression model

Y = response or outcome variable
X = explanatory or input variables

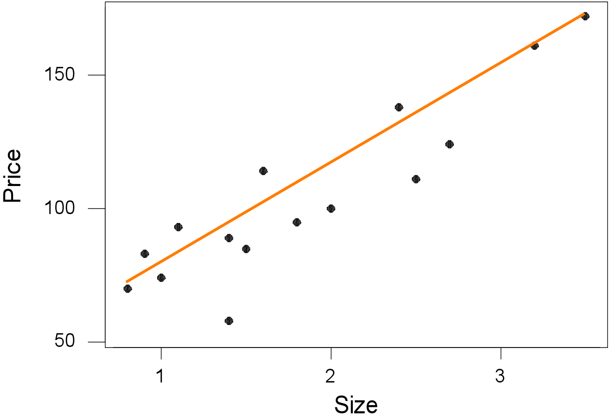
A linear relationship is written

$$Y = b_0 + b_1 X + e$$

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:

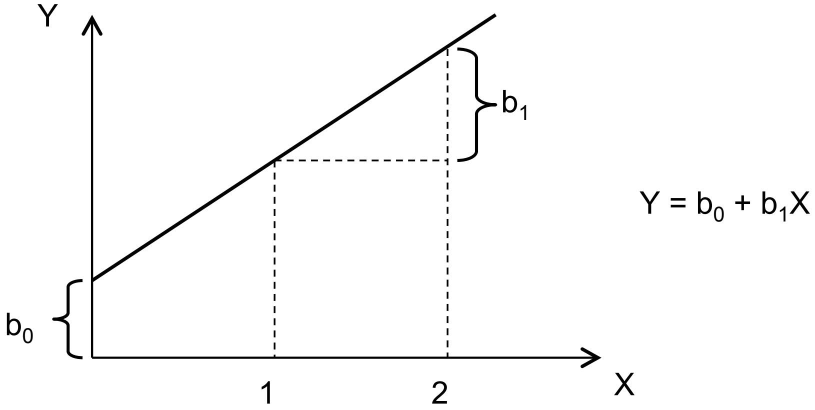
$$Y = b_0 + b_1 X$$

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

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

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

Linear prediction



Linear prediction

Q: How to find the "best line"?

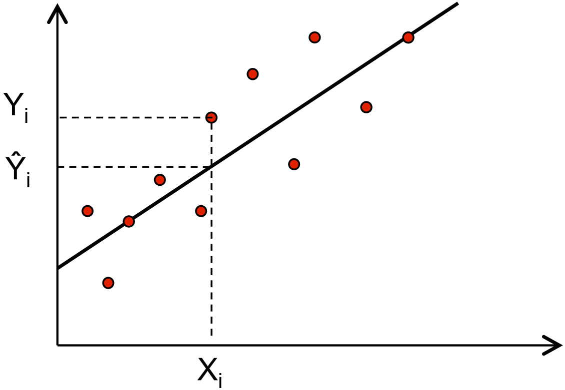
We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

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

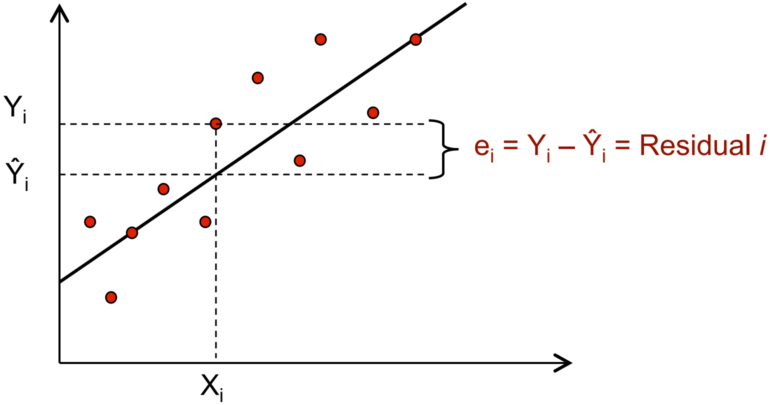
What is the "fitted value"?



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:

- Give weights to all of the residuals.
- 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:

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

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

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

Least squares – R output

```r
data = read.csv('housedata.csv')
fit = lm(Price~Size,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ Size, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.425  -8.618   0.575  10.766  18.498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.885      9.094   4.276 0.000903 ***
## Size          35.386      4.494   7.874 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 13 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.8133
## F-statistic:    62 on 1 and 13 DF,  p-value: 2.66e-06
```

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

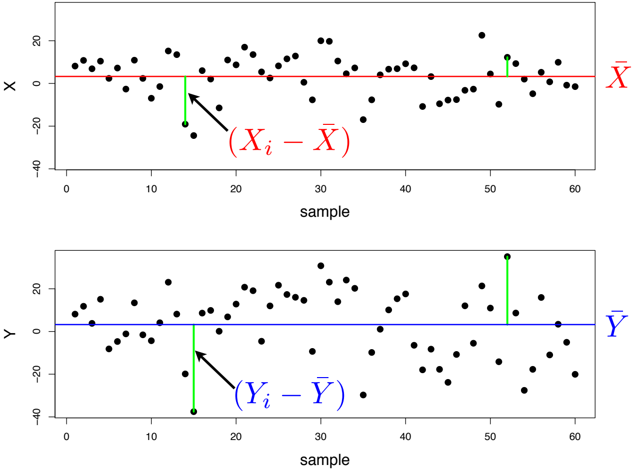
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

- Sample Standard Deviation:

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

Visual: standard deviation

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

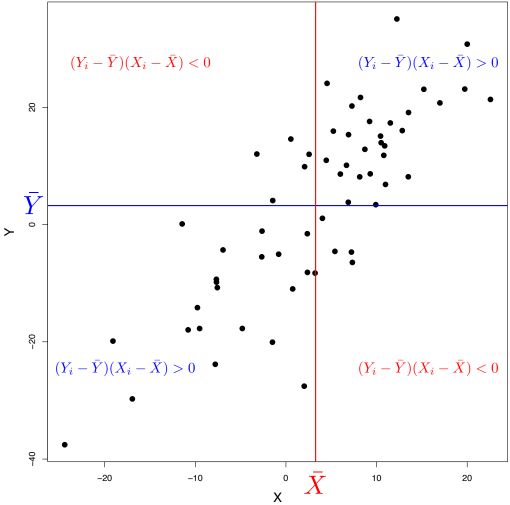


$s_x = 9.7$ $s_y = 15.98$

Visual: Covariance

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

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



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

How do we interpret that?

A standardized measure: Correlation

Correlation is the standardized covariance:

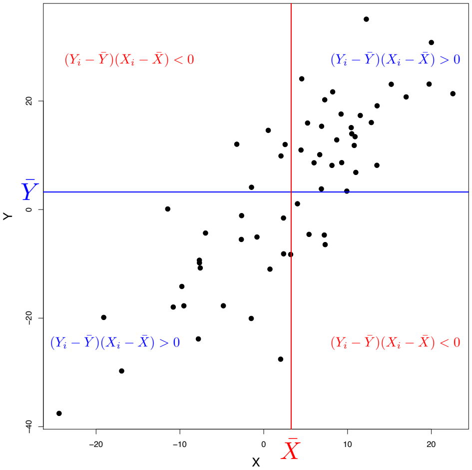
$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\operatorname{cov}(X, Y)}{s_x s_y}$$

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

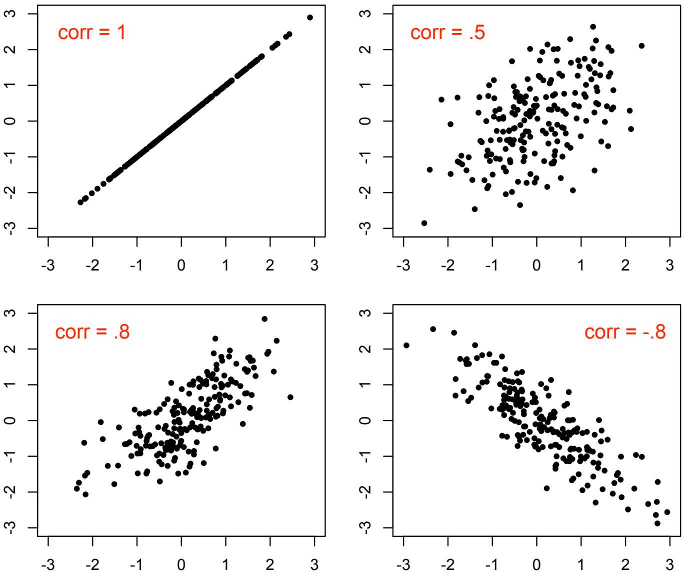
This gives the direction (negative or positive) and strength ($0 \rightarrow 1$) of the linear relationship between X and Y.

Correlation

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



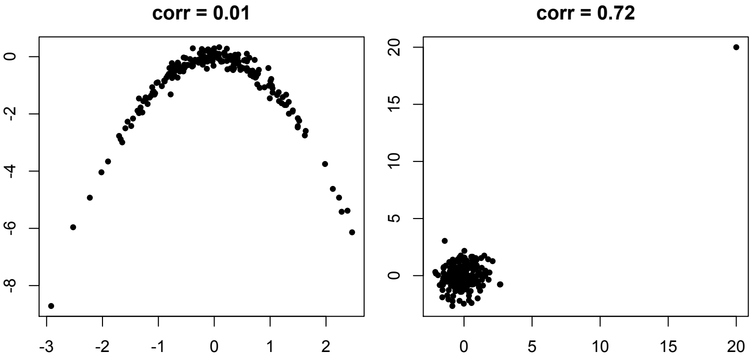
Correlation



Correlation

Only measures linear relationships:

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



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

Back to least squares

Intercept:

$$b_0 = \bar{Y} - b_1\bar{X} \Rightarrow \bar{Y} = b_0 + b_1\bar{X}$$

The point $(\bar{X}, \bar{Y})$ is on the regression line!

Least squares finds the point of means and rotates the line through that point until getting the "right" slope

Slope:

$$\begin{aligned} b_1 = \text{corr}(X, Y) \times \frac{s_Y}{s_X} &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \frac{\text{cov}(X, Y)}{\text{var}(X)} \end{aligned}$$

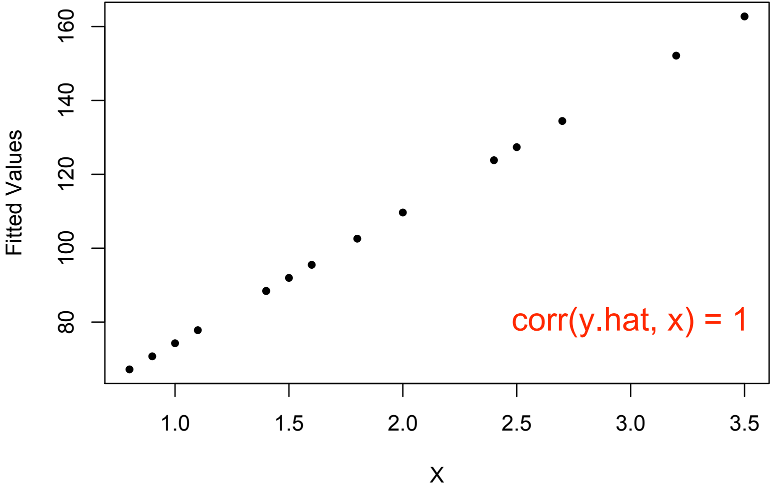
So, the right slope is the **correlation coefficient** times a **scaling factor** that ensures the proper units for b_1

More on least squares

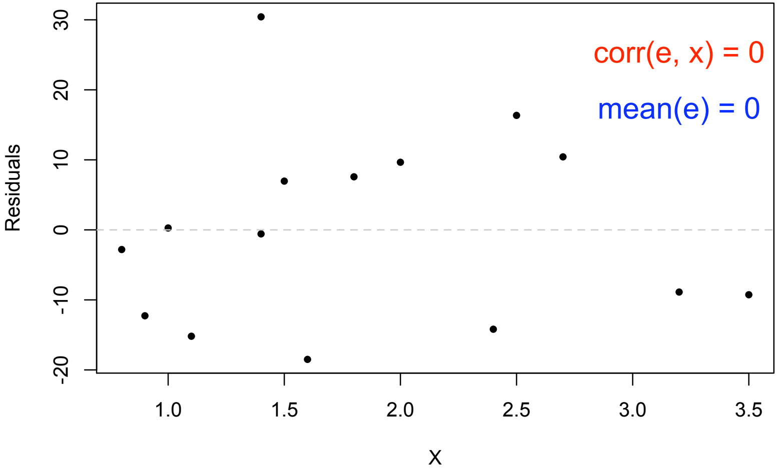
From now on, terms "fitted values" ($\hat{Y}_i$) and "residuals" (e_i) refer to those obtained from the least squares line.

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

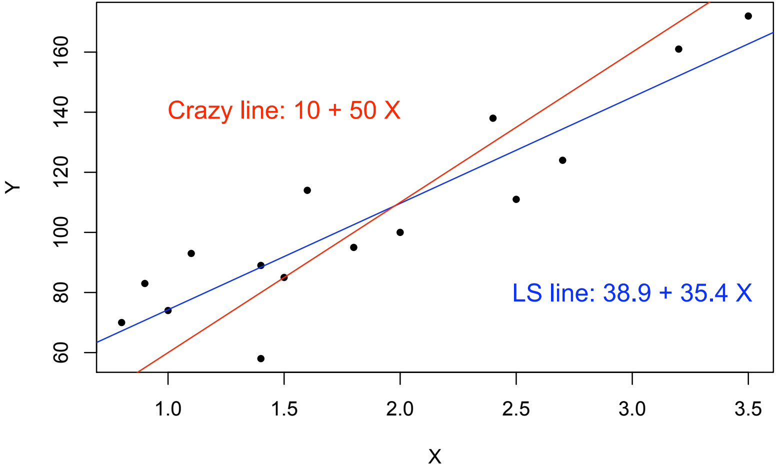


The residuals and X



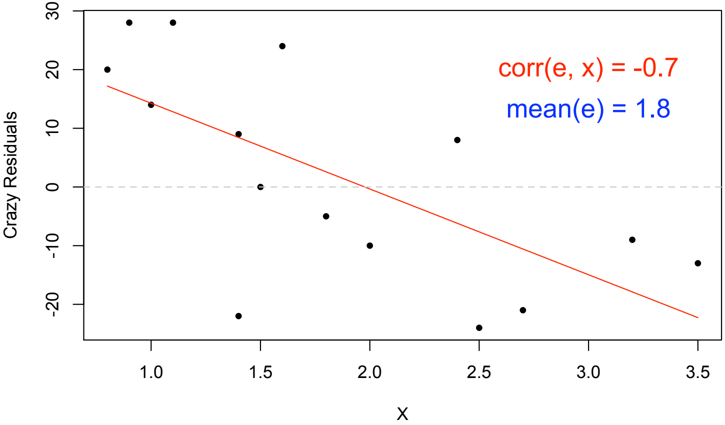
Why?

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; $\text{corr}(X, e) = 0$.

Example 2: Offensive performance in baseball

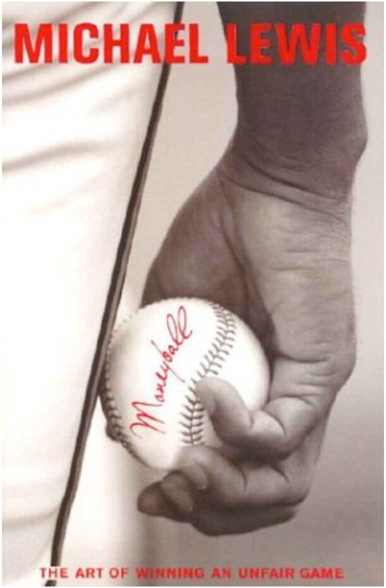
Problems:

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

Solutions:

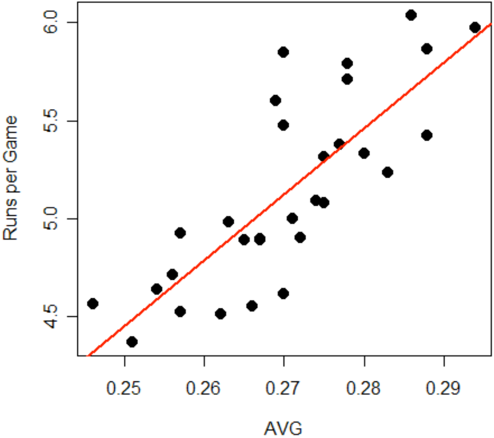
- Compare *prediction rules* that forecast runs as a function of either AVG (batting average), SLG (slugging percentage – total bases divided by at bats) or OBP (on base percentage)

Example 2: Offensive performance in baseball



Baseball data – using AVG

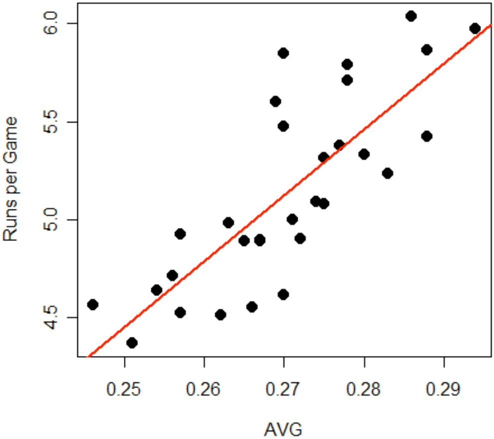
Each observation corresponds to a team in MLB. Each quantity is the average over a season.



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

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

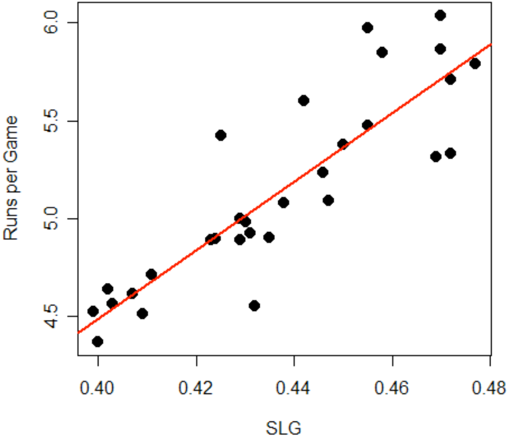
Baseball data – using AVG



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

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

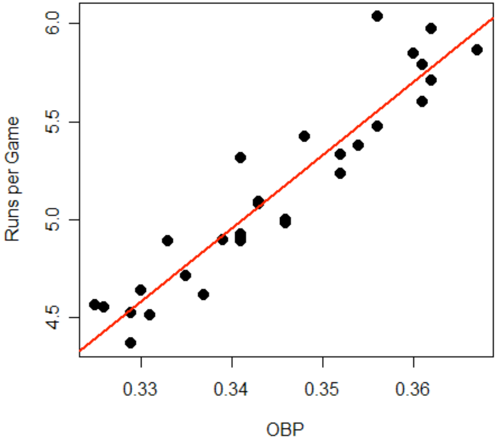
Baseball Data – using SLG



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

Decomposing the variance

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{\substack{\text{Total Sum of}\\ \text{Squares}\\ \text{SST}}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\substack{\text{Regression SS}\\ \text{SSR}}} + \underbrace{\sum_{i=1}^{n} e_i^{\,2}}_{\substack{\text{Error SS}\\ \text{SSE}}}$$

SSR: Variation in Y explained by the regression line.
SSE: Variation in Y that is left unexplained.

Decomposing the variance

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{\boxed{\substack{\textbf{Total Sum of} \\ \textbf{Squares} \\ \textbf{SST}}}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\boxed{\substack{\textbf{Regression SS} \\ \textbf{SSR}}}} + \underbrace{\sum_{i=1}^{n}e_i^{\,2}}_{\boxed{\substack{\textbf{Error SS} \\ \textbf{SSE}}}}$$

SSR: Variation in Y explained by the regression line.
SSE: Variation in Y that is left unexplained.

$$\text{SSR} = \text{SST} \Rightarrow \text{perfect fit.}$$

Be careful of similar acronyms; e.g. SSR for "residual" SS.

A goodness of fit measure: R^2

The coefficient of determination, denoted by R^2, measures goodness of fit:

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

- $0 < R^2 < 1$.
- The closer R^2 is to 1, the better the fit.

Back to baseball

Three very similar, related ways to look at a simple linear regression... with only one X variable, life is easy!

	R^2	corr	SSE
OBP	0.88	0.94	0.79
SLG	0.76	0.87	1.64
AVG	0.63	0.79	2.49

Prediction and regression + probability

Prediction and the modeling goal

A prediction rule is any function where you input X and it outputs $\hat{Y}$ as a predicted response at X.

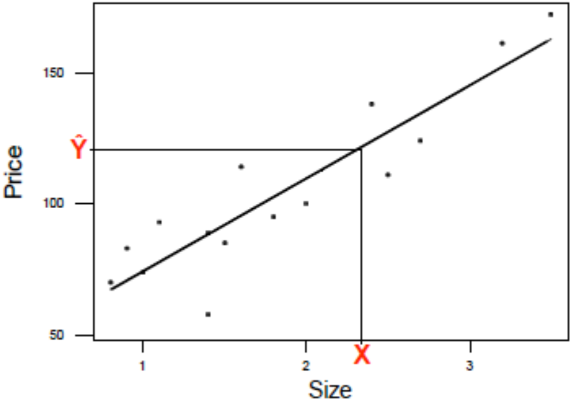
The least squares line is a prediction rule:

$$\hat{Y} = f(X) = b_0 + b_1 X$$

Prediction and the modeling goal

$\hat{Y}$ is not going to be a perfect prediction.

We need to devise a notion of forecast accuracy.



Prediction and the modeling goal

There are two things that we want to know:

- What value of Y can we expect for a given X?
- How sure are we about this forecast? Or how different could Y be from what we expect?

Our goal is to measure the accuracy of our forecasts or how much uncertainty there is in the forecast. One method is to specify a range of Y values that are likely, given an X value.

Prediction Interval: probable range for Y-values given X

Prediction and modeling

Key Insight: To construct a prediction interval, we will have to assess the likely range of error values corresponding to a Y value that has not yet been observed!

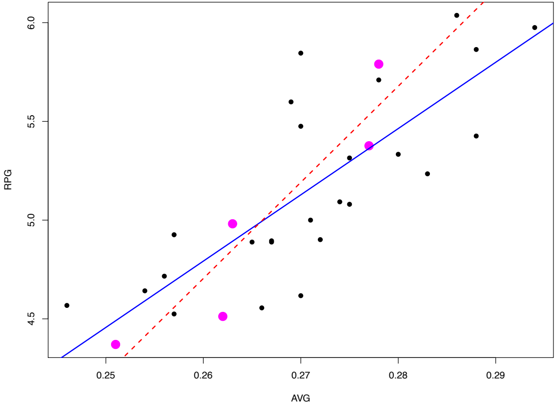
We will build a probability model (e.g., Normal distribution).

Then we can say something like "with 95% probability the error will be no less than -$28,000 or larger than $28,000".

We must also acknowledge that the "fitted" line may be fooled by particular realizations of the residuals.

We are always looking at samples of data!

We are always looking at samples! The dashed line fits the purple points. The solid line fits all the points. Which line is better? Why?



In summary, we need to work with the notion of a "true line" and a probability distribution that describes deviation around the line.

The simple linear regression model

The power of statistical inference comes from the ability to make precise statements about the accuracy of the forecasts.

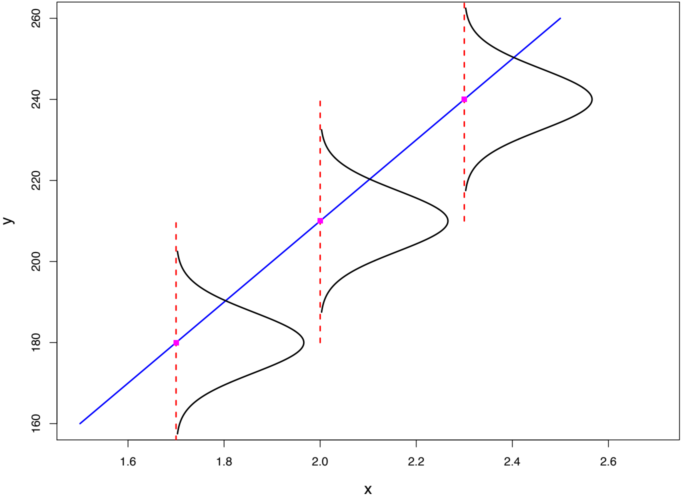
In order to do this we must invest in a probability model.

Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

$$\varepsilon \sim \mathrm{N}(0, \sigma^2)$$

- $\beta_0 + \beta_1 X$ represents the "true line"; The part of Y that depends on X.
- The error term ε is independent "idosyncratic noise"; The part of Y not associated with X.

Visually, what is going on here?



The simple linear regression model – example

You are told (without looking at the data) that

$$\beta_0 = 40;\ \beta_1 = 45;\ \sigma = 10$$

and you are asked to predict price of a 1500 square foot house.

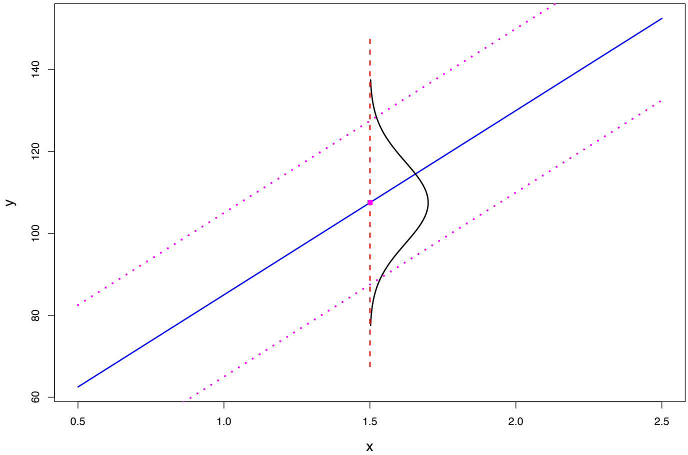
What do you know about Y from the model?

$$\begin{aligned} Y &= 40 + 45(1.5) + \varepsilon \\ &= 107.5 + \varepsilon \end{aligned}$$

Thus our prediction for price is $Y|(X = 1.5) \sim \mathrm{N}(107.5, 10^2)$
and a 95% *Prediction Interval* for Y is $87.5 < Y < 127.5$

In picture form, our model tells us about uncertainty

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



The conditional distribution for Y given X is Normal:

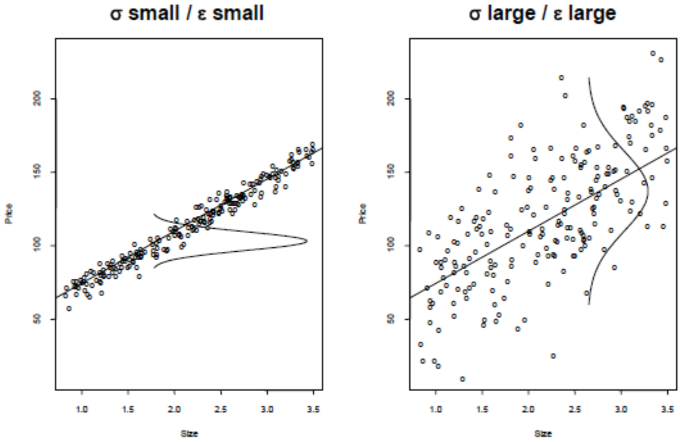
$$Y|X = x \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2).$$

The importance of σ

The conditional distribution for Y given X is Normal:

$$Y|X \sim \mathrm{N}(\beta_0 + \beta_1 X, \sigma^2).$$

σ controls dispersion:



Multiple linear regression

The multiple linear 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, 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 assumptions of our linear regression model:

→ The conditional mean of Y is linear in the X_j variables.

→ The errors (deviations from line)

- are normally distributed
- independent from each other
- identically distributed (i.e., they have constant variance)

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

The MLR model

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

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

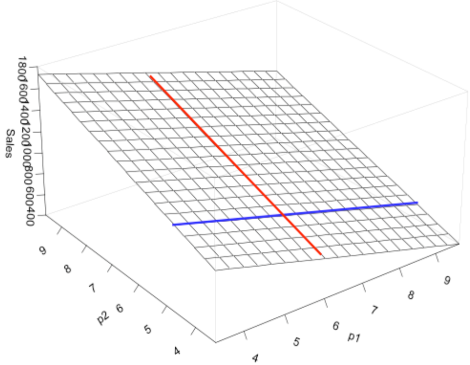
Holding all other variables constant, β_j is the average change in Y per unit change in X_j.

The MLR model

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

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

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



Least squares again!

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

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

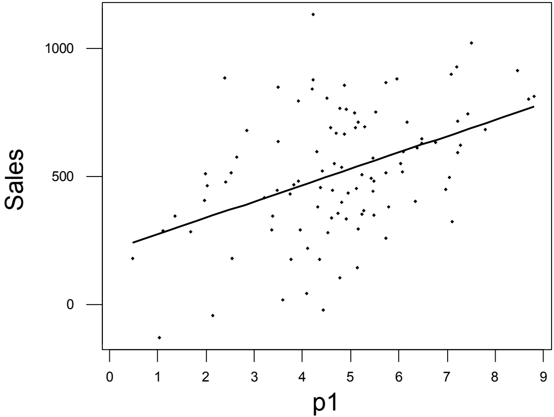
The importance of the right-hand-side

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)

The importance of 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!**



$\rightarrow$ 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

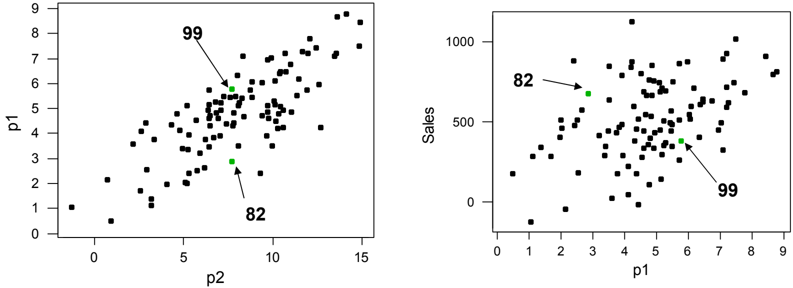
$$Sales = 116 - 97.7P1 + 109P2$$

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.

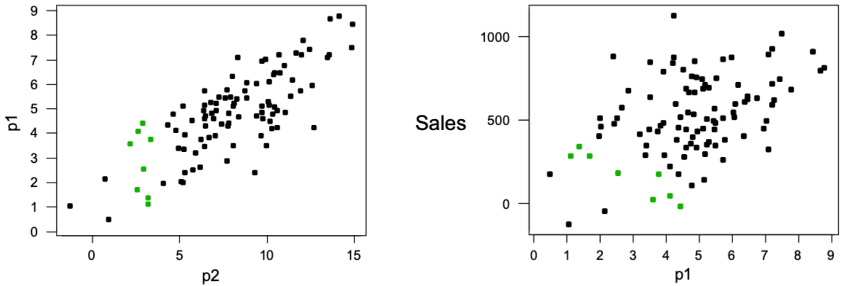
We see that an increase in $P1$, holding $P2$ constant, corresponds to a drop in Sales!



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

Understanding multiple regression

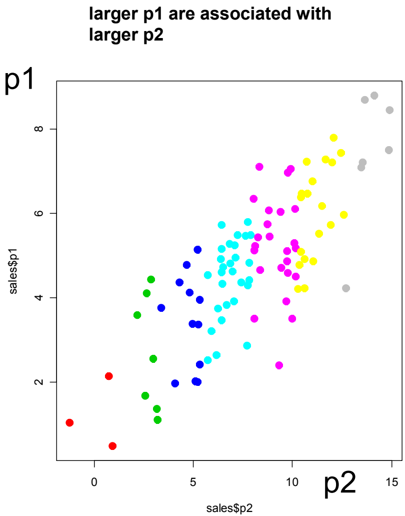
Let's look at a subset of points where $P1$ varies and $P2$ is held approximately constant...

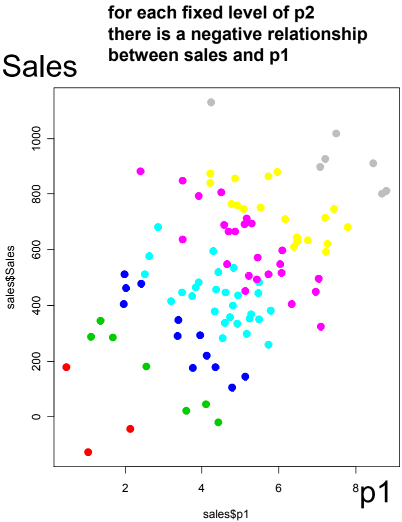


For a fixed level of $P2$, variation in $P1$ is negatively correlated with Sales!

Understanding multiple regression

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





Understanding multiple regression

Summary:

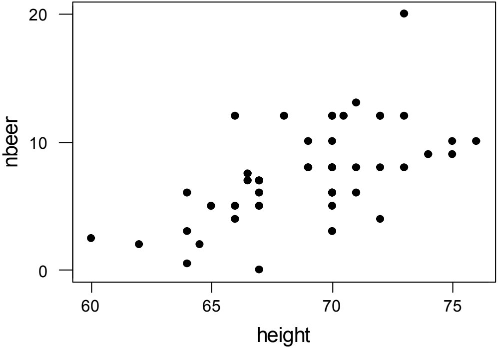
$\rightarrow$ A larger $P1$ is associated with larger $P2$ and the overall effect leads to bigger sales

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

$\rightarrow$ MLR does the trick and unveils the **correct** economic relationship between Sales and prices!

Example: Beers, height, weight, and getting drunk

Beer data (from Forbidden Courses 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)
```

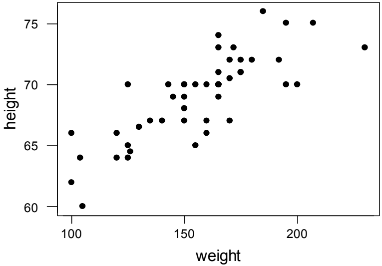
```
##
## Call:
## lm(formula = nbeer ~ height, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.164 -2.005 -0.093  1.738  9.978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9200     8.9560  -4.122 0.000148 ***
## height        0.6430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



The correlations:

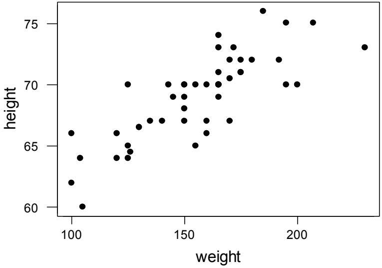
	nbeer	weight
weight	0.692	
height	0.582	0.806

The two x's are highly correlated !!

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

Summary slide

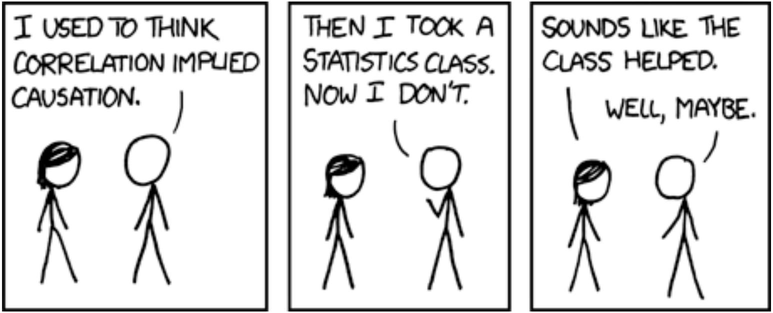
In general, when we see a relationship between y and x (or x's), that relationship may be driven by variables "lurking" in the background which are related to your current x's.

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



Dummy variables

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 gender variable 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...?"*

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 gender variable 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?

- UATX education vs. not
- foreign or domestic born citizen
- quarterback vs. wide receiver

Detecting sex discrimination

We can use regression to answer these questions, but first 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:

This can be done implicitly in R by, `Sex = factor(Gender)`. This tells R that Sex is a variable separated into its unique levels, in this case just 2!

Detecting sex discrimination

Now you can present the following model in court:

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \epsilon_i$$

How do you interpret β_1? What is your predicted salary for males and females?

Detecting sex discrimination

Now you can present the following model in court:

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \epsilon_i$$

How do you interpret β_1? What is your predicted salary for males and females?

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

β_1 is the male-female difference!

Detecting sex discrimination

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \epsilon_i$$

```
data = read.table("SalaryData.txt",header=T)
Sex = (data$Gender=="Male")
data$Sex = Sex
fit = lm(Salary~Sex,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Salary ~ Sex, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.805  -6.434  -1.860   4.115  51.495
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2099     0.8945  41.597  < 2e-16 ***
## SexTRUE       8.2955     1.5645   5.302 2.94e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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 gender 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...

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Exp}_i + \epsilon_i$$

What does that mean? Write out the model for each gender separately:

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

Detecting sex discrimination

Here is our data with this additional variable:

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

Detecting sex discrimination

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Exp} + \epsilon_i$$

```r
fit = lm(Salary~Sex+Exp,data)
summary(fit)
```

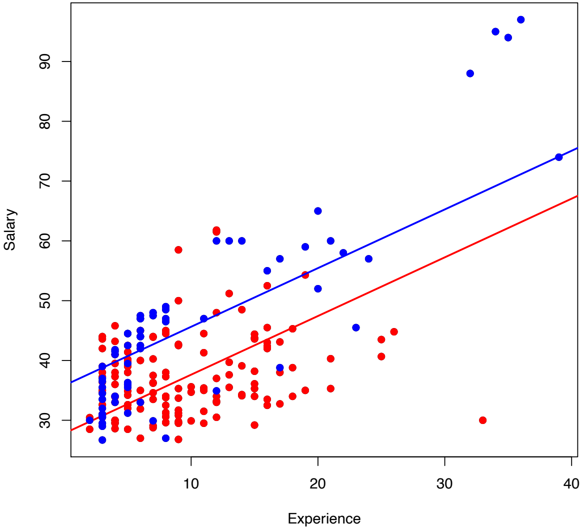
```
##
## Call:
## lm(formula = Salary ~ Sex + Exp, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.1899  -5.7484  -0.6046   4.8129  25.8548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.81190    1.02789  27.057  < 2e-16 ***
## SexTRUE      8.01189    1.19309   6.715 1.81e-10 ***
## Exp          0.98115    0.08028  12.221  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\rightarrow$ $\text{Salary}_i = 27 + 8\text{Sex}_i + 0.98\text{Exp}_i + \epsilon_i$

Is this good or bad news for the defense?

Detecting sex discrimination

$$\text{Salary}_i = \begin{cases} 27 + 0.98\text{Exp}_i + \epsilon_i & \text{females} \\ 35 + 0.98\text{Exp}_i + \epsilon_i & \text{males} \end{cases}$$



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!**

House prices revisted

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
... ... ... ...
```

House prices revisted

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

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

House prices revisted

$$\text{Price}_i = \beta_0 + \beta_1 \text{dn1}_i + \beta_2 \text{dn2}_i + \beta_3 \text{SqFt}_i + \epsilon_i$$

$$E[\text{Price}|\text{dn1} = 1, \text{SqFt}] \quad = \quad \beta_0 + \beta_1 + \beta_3 \text{SqFt} \quad (\text{Nbhd 1})$$

House prices revisted

$$\text{Price}_i = \beta_0 + \beta_1 \text{dn1}_i + \beta_2 \text{dn2}_i + \beta_3 \text{SqFt}_i + \epsilon_i$$

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

House prices revisted

$$\text{Price}_i = \beta_0 + \beta_1 \text{dn1}_i + \beta_2 \text{dn2}_i + \beta_3 \text{SqFt}_i + \epsilon_i$$

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

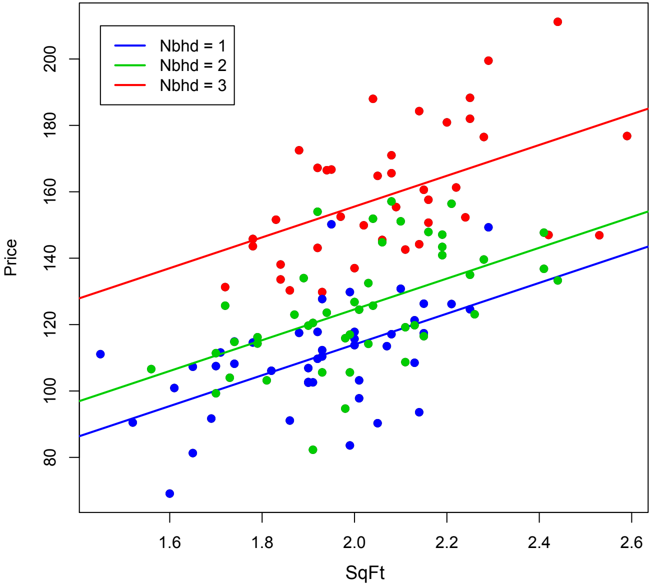
Model output

```
fit = lm(Price~dn1+dn2+SqFt,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ dn1 + dn2 + SqFt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.107 -10.924  -0.305   9.643  38.506
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.776     14.248   4.406 2.25e-05 ***
## dn1          -41.535      3.534 -11.754  < 2e-16 ***
## dn2          -30.967      3.369  -9.192 1.13e-15 ***
## SqFt          46.386      6.746   6.876 2.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.26 on 124 degrees of freedom
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6774
## F-statistic: 89.91 on 3 and 124 DF,  p-value: < 2.2e-16
```

$$\text{Price} = 62.78 - 41.54 * \text{dn1} - 30.97 * \text{dn2} + 46.39 * \text{SqFt} + \epsilon$$

What do these models look like?



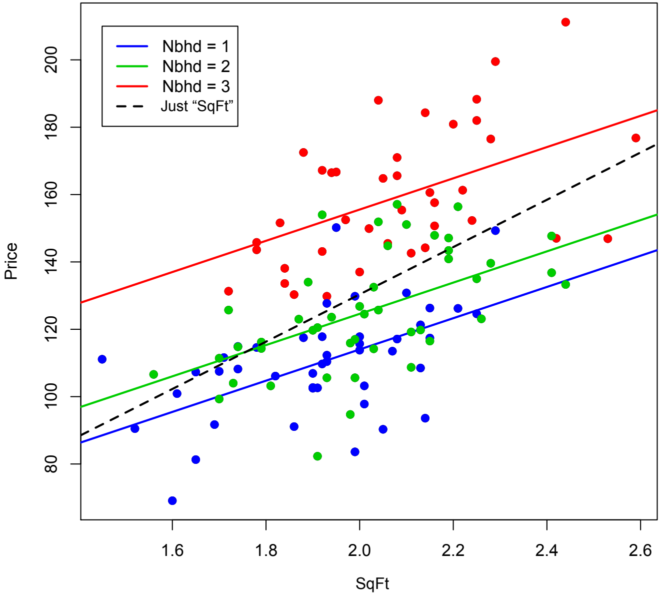
Model output only with "SqFt"

```r
fit = lm(Price~SqFt,data)
summary(fit)
```

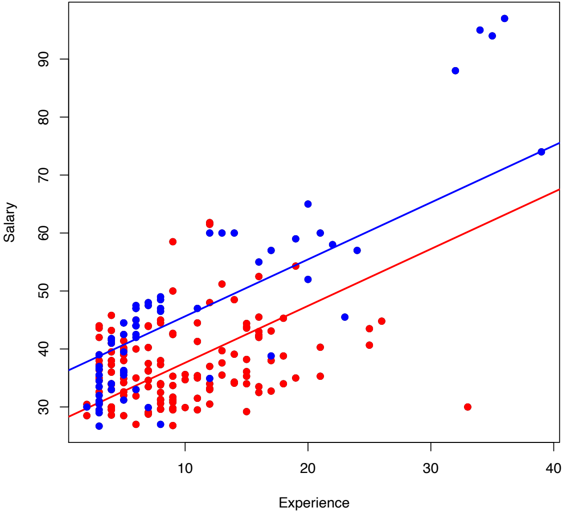
```
##
## Call:
## lm(formula = Price ~ SqFt, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.59 -16.64  -1.61  15.12  54.83
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.091     18.966  -0.532    0.596
## SqFt          70.226      9.426   7.450  1.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.48 on 126 degrees of freedom
## Multiple R-squared:  0.3058, Adjusted R-squared:  0.3003
## F-statistic:  55.5 on 1 and 126 DF,  p-value: 1.302e-11
```

$$\text{Price} = -10.09 + 70.23 * \text{SqFt} + \epsilon$$

What do these models look like?



Making the model more flexible ...



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

Making the model more flexible ...

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

Yes... Consider the following model:

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Exp}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Exp}_i \times \text{Sex}_i + \epsilon_i$$

For Females:

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Exp}_i + \epsilon_i$$

For Males:

$$\text{Salary}_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{Exp}_i + \epsilon_i$$

How are these models different from each other?

Sex discrimination case

We are just creating a new variable!

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

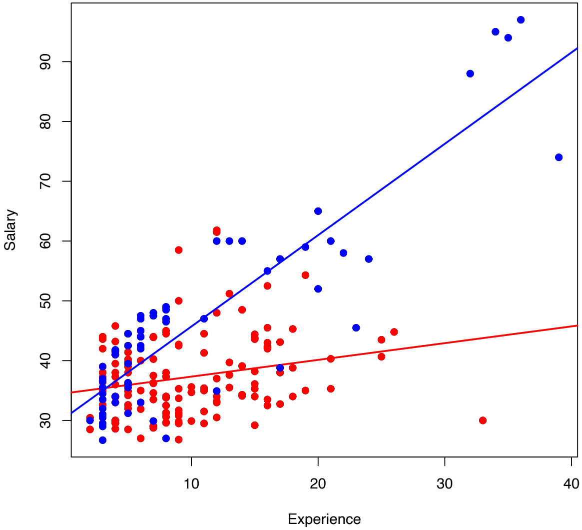
Sex discrimination case

```
fit = lm(Salary~Sex+Exp+ExpSex,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Salary ~ Sex + Exp + ExpSex, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.0685  -4.6506  -0.7679   4.4034  23.9122
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.5283     1.1380  30.342  < 2e-16 ***
## SexTRUE      -4.0983     1.6658  -2.460  0.01472 *
## Exp           0.2800     0.1025   2.733  0.00684 **
## ExpSex        1.2478     0.1367   9.130  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.816 on 204 degrees of freedom
## Multiple R-squared:  0.6386, Adjusted R-squared:  0.6333
## F-statistic: 120.2 on 3 and 204 DF,  p-value: < 2.2e-16
```

$$\text{Salary} = 34 - 4 * \text{Sex} + 0.28 * \text{Exp} + 1.24 * \text{Exp}{*}\text{Sex} + \epsilon$$

Sex discrimination case



Is this good or bad news for the plaintiff?

Variable interaction

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.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \epsilon$$

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

More topics in regression

- How do you model Y when it is **binary**? Logistic regression ...
- How to select the **best** model? Hand pick covariates, forward or backward stepwise selection, LASSO, ...
- How to get **causal estimates** from regression? Regression on binary variable, Regression on treatment and covariates (confounding), "Regression discontinuity"...
- My causal estimates are usually differences in group averages, what about **individualized causal effects**? Heterogenous treatment effects, ensemble learning and BART ...