TEXAS McCombs

MASTER OF SCIENCE IN BUSINESS ANALYTICS

Introduction to supervised learning, the bias-variance tradeoff, and linear regression

Outline

Introduction to supervised learning

Bias-variance tradeoff (a key idea of this course!)

Linear regression

Introduction to Predictive Models

Simply put, the goal is to predict a target variable Y with input variables X!

In Data Mining terminology this is know as **supervised learning** (also called *Predictive Analytics*).

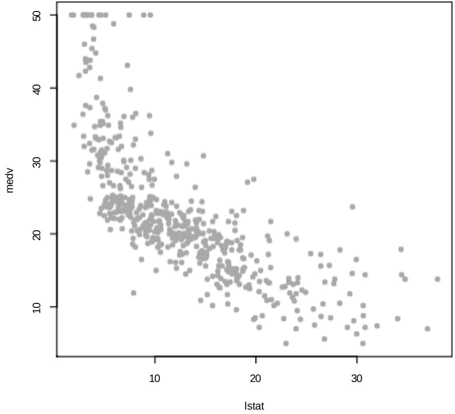
In general, a useful way to think about it is that Y and X are related in the following way:

$$Y_i = f(X_i) + !_i$$

The main purpose of this part of the course is to *learn or estimate* $f(\cdot)$ from data

Example: Boston Housing

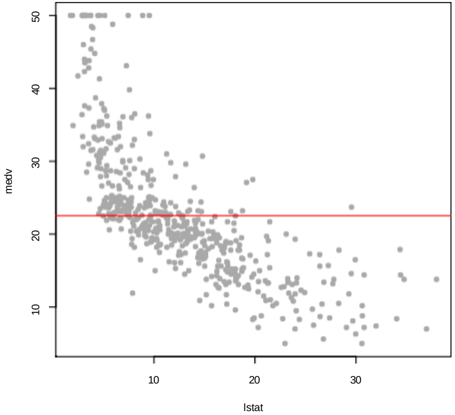
We might be interested in predicting the median house value as a function of some measure of social economic level... here's some data:



What should $f(\cdot)$ be?

Example: Boston Housing

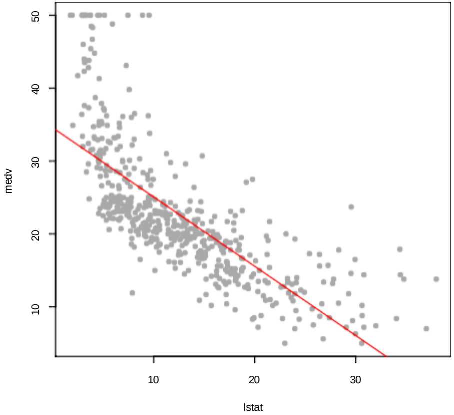
How about this...



If *lstat* = 30 what is the prediction for *medv*?

Example: Boston Housing

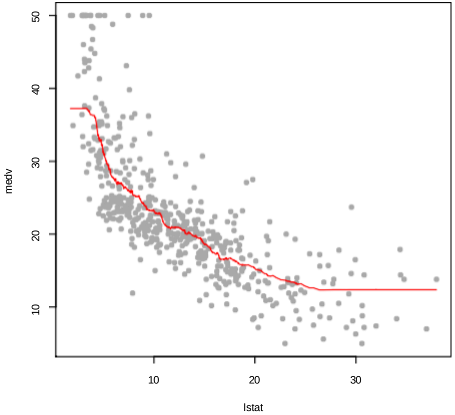
or this...



If *lstat* = 30 what is the prediction for *medv*?

Example: Boston Housing

or even this?



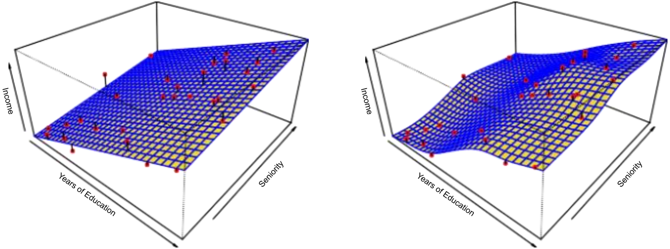
If *lstat* = 30 what is the prediction for *medv*?

How do we estimate $f(\cdot)$?

- Using *training data*:

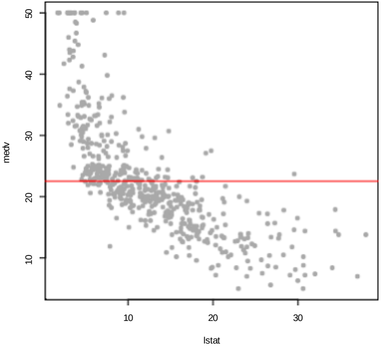
$$\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$$

- We use a statistical method to *estimate* the function $f(\cdot)$
- Two general methodological strategies:
  1. parametric models (restricted assumptions about $f(\cdot)$)
  2. non-parametric models (flexibility in defining $f(\cdot)$)

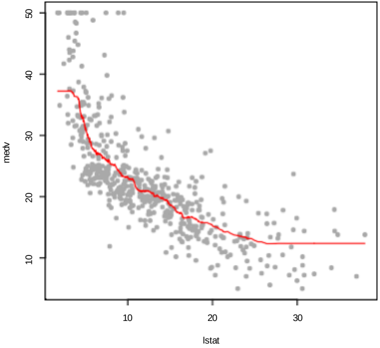


Back to Boston Housing

Parametric Model
(*Y* = μ + *!*)



Non-Parametric Model
(k-nearest neighbors)

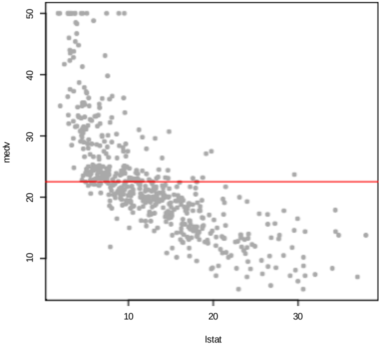


Back to Boston Housing

Simplest parametric model: $Y_i = \mu + !_i$

Using the training data, we estimate $f(\cdot)$ as

$$f(\cdot) = \hat{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$



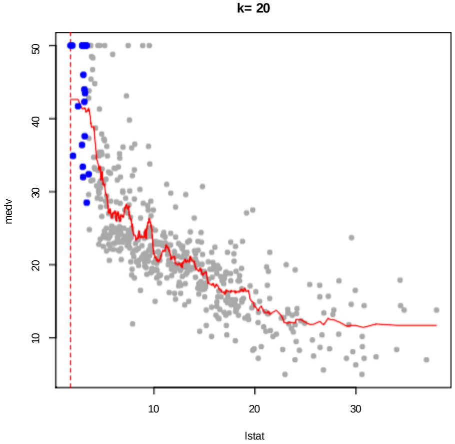
Back to Boston Housing

The above strategy averages all points in the training set... maybe points that are “closer” to the place I am trying to predict should be more relevant...

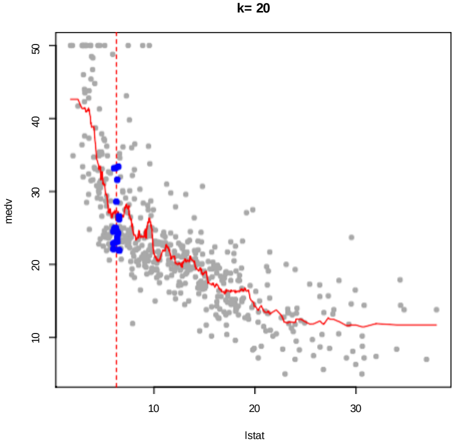
How about averaging the closest 20 neighbors?
What do I mean by closest? We will choose the 20 points that are closest to the X value we are trying to predict.

This is what is called the k-nearest neighbors algorithm

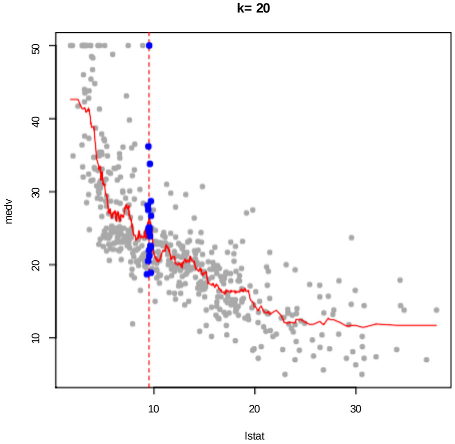
Back to Boston Housing



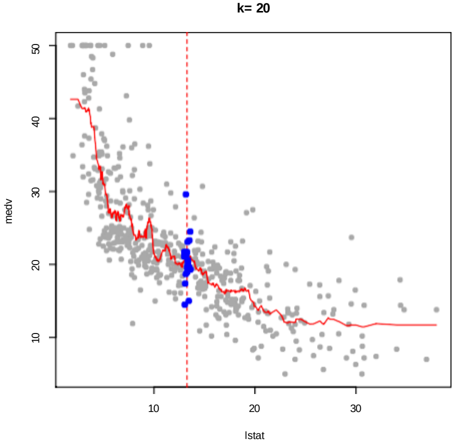
Back to Boston Housing



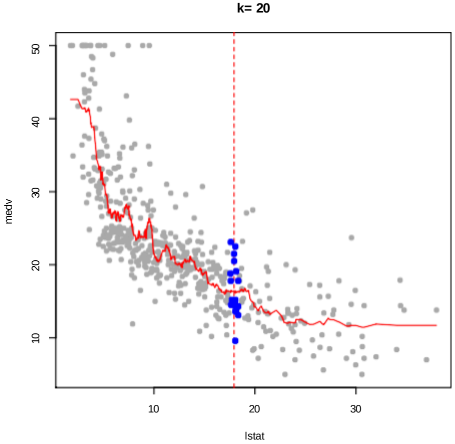
Back to Boston Housing



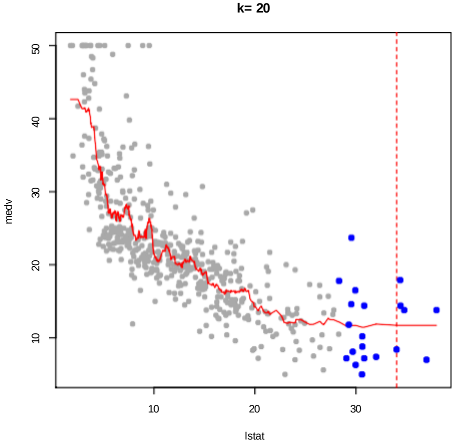
Back to Boston Housing



Back to Boston Housing

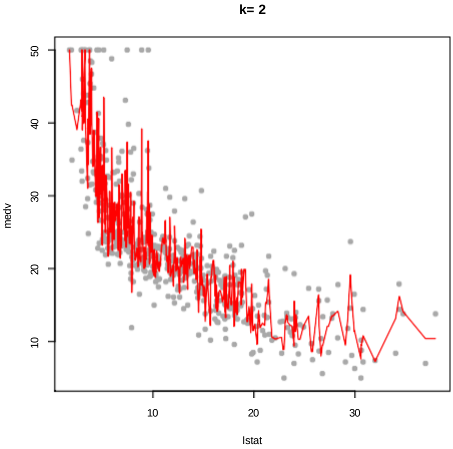


Back to Boston Housing



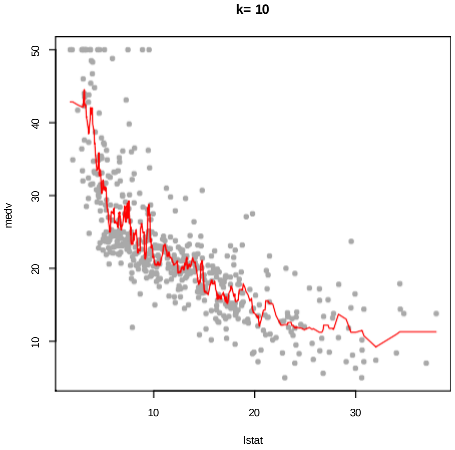
Back to Boston Housing

Okay, that seems sensible but why not use 2 neighbors or 200 neighbors?



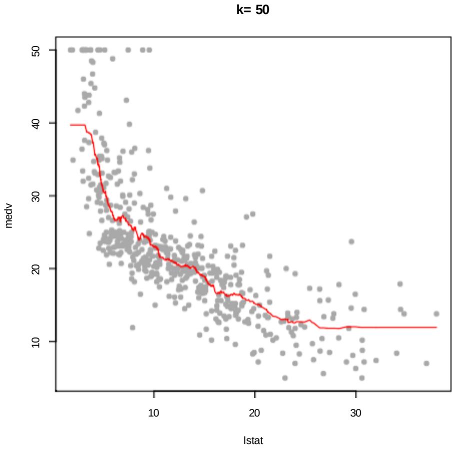
Back to Boston Housing

Okay, that seems sensible but why not use 5 neighbors or 200 neighbors?



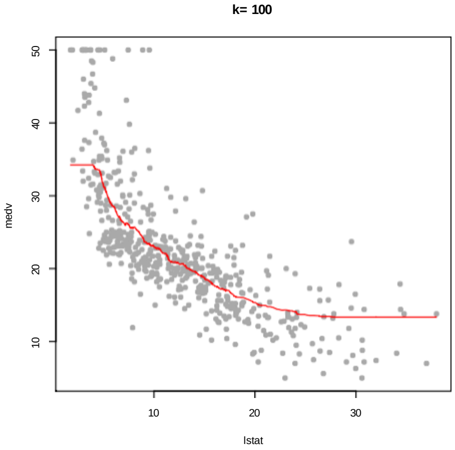
Back to Boston Housing

Okay, that seems sensible but why not use 5 neighbors or 200 neighbors?



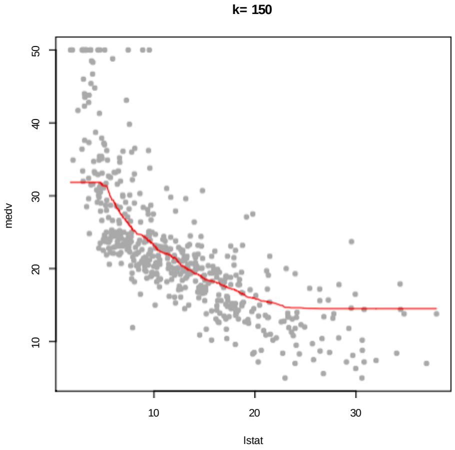
Back to Boston Housing

Okay, that seems sensible but why not use 5 neighbors or 200 neighbors?



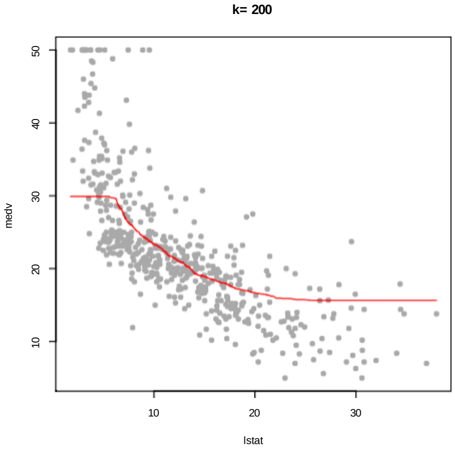
Back to Boston Housing

Okay, that seems sensible but why not use 5 neighbors or 200 neighbors?



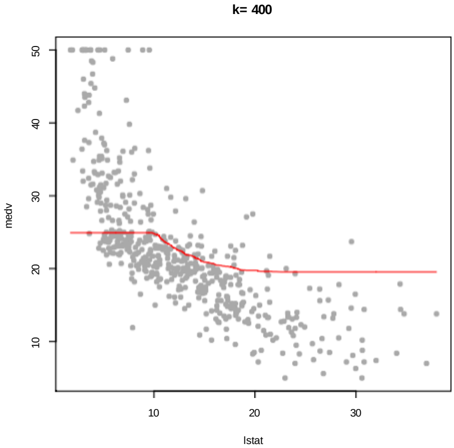
Back to Boston Housing

Okay, that seems sensible but why not use 5 neighbors or 200 neighbors?



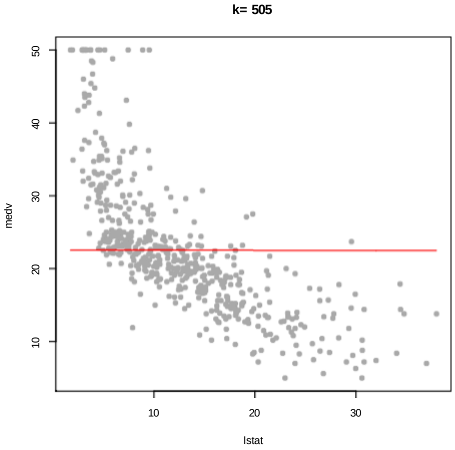
Back to Boston Housing

Okay, that seems sensible but why not use 5 neighbors or 200 neighbors?



Back to Boston Housing

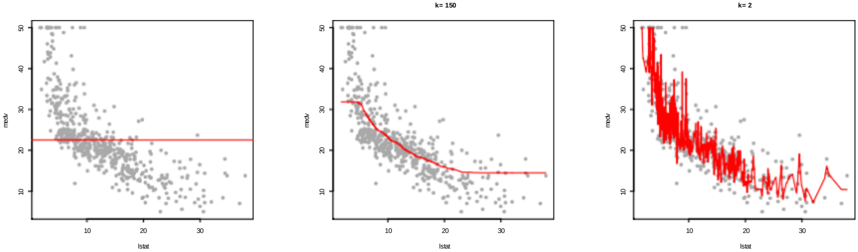
Okay, that seems sensible but why not use 5 neighbors or 200 neighbors?



Complexity, Generalization and Interpretation

- As we have seen in the examples above, there are lots of options in estimating $f(X)$.
- Some methods are very flexible some are not... *why would we ever choose a less flexible model?*
  1. Simple, more restrictive methods are usually easier to interpret
  2. More importantly, it is often the case that simpler models are more accurate in making future predictions.

Not too simple, but not too complex!



Measuring Accuracy

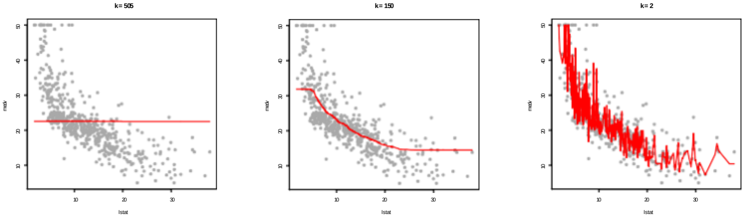
How accurate are each of these models?

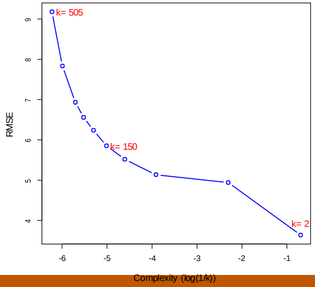
Using the training data a standard measure of accuracy is the *root mean-squared error*

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}(X_i)\right)^2}$$

This measure, on average, how large the “mistakes” (errors) made by the model are...

Measuring Accuracy (Boston housing, again)





So, I guess we should just go with the most complex model, i.e., *k* = 2, right?

Out-of Sample Predictions

But, do we really care about explaining what we have already seen?

Key Idea: what really matters is our prediction accuracy out-of-sample!!!

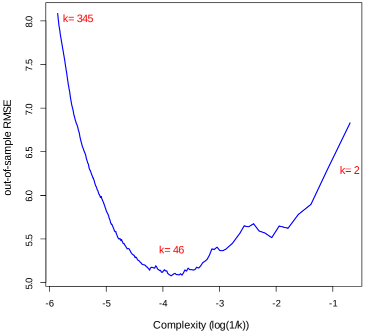
Suppose we have m additional observations (X_i^o, Y_i^o), for $i = 1, \ldots, m$, that we did not use to fit the model. Let's call this dataset the *validation set* (also known as *hold-out set* or *test set*)

Let's look at the out-of-sample RMSE:

$$RMSE^o = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(Y_i^o - f(X_i^o)\right)^2}$$

Out-of Sample Predictions

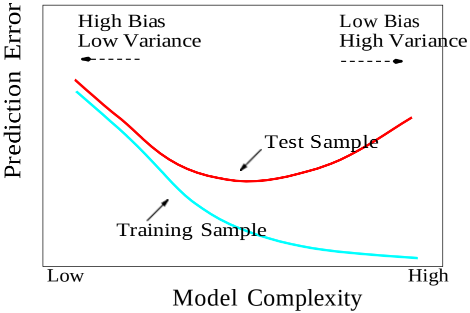
In our Boston housing example, I randomly chose a training set of size 400. I re-estimate the models using only this set and use the models to predict the remaining 106 observations (validation set)...



Now, the model where k = 46 looks like the most accurate choice!!

Not too simple but not too complex!!!

The Key Idea of our Course!



This shows the typical behavior of the in and out-of-sample prediction error as a function of the complexity of the model... too flexible models will adapt itself too closely to the the training set and will not generalize well, i.e., not be very good for the test data.

Bias-Variance Trade-Off

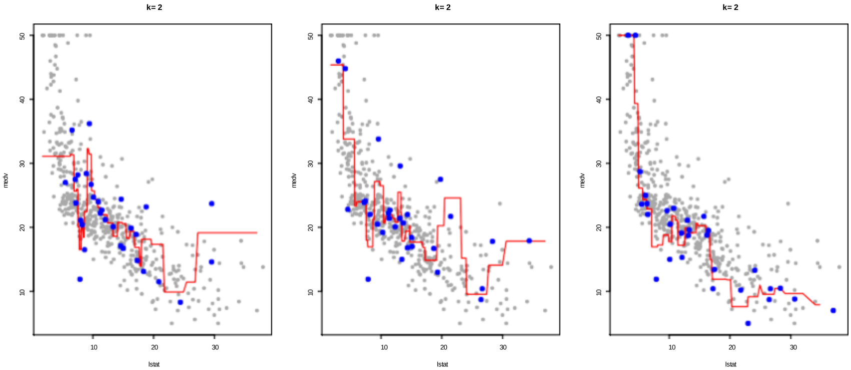
Why do complex models behave poorly in making predictions?

Let's start with an example...

- In the Boston housing example, I will randomly choose 30 observations to be in the training set 3 different times...
- for each training set I will estimate $f(\cdot)$ using the k-nearest neighbors idea... first with $k = 2$ and them with $k = 20$

Bias-Variance Trade-Off

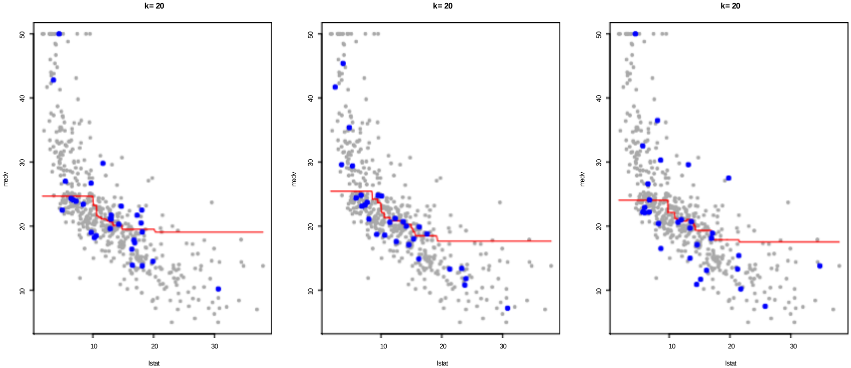
k=2 Hi variability...



(blue points are the training data used)

Bias-Variance Trade-Off

k=20 Low variability ... but BIAS!!



(blue points are the training data used)

Bias-Variance Trade-Off

What did we see here?

- When $k = 2$, it seems that the estimate of $f(\cdot)$ varies a lot between training sets...
- When $k = 20$ the estimates look a lot more stable...

Now, imagine that you are trying to predict *medv* when *lstat* = 20... compare the changes in the predictions made by the different training sets under $k = 2$ and $k = 20$... what do you see?

Bias-Variance Trade-Off

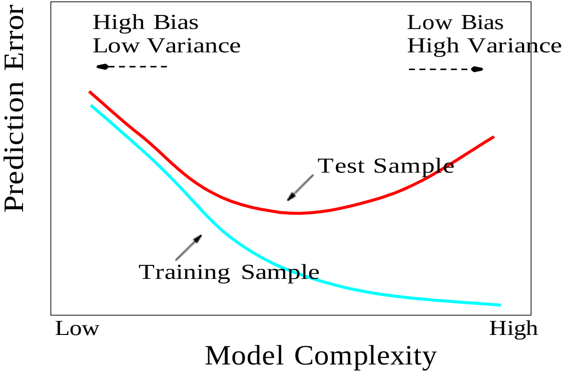
- This is an illustration of what is called the *bias-variance trade-off*.
- In general, simple models are trying to explain a complex, real problem with not a lot of flexibility so it introduces bias... on the other hand, by being simple the estimates tend to have low variance
- On the other hand, complex models are able to quickly adapt to the real situation and hence lead to small bias... however, by being to adaptable, it tends to vary a lot, i.e., high variance.

Bias-Variance Trade-Off

- In other words, we are trying to capture important patterns of the data that generalize to the future observations we are trying to predict. Models that are too simple are not able to capture relevant patterns and might have too big of a bias in predictions...
- Models that are too complex will “chase” irrelevant patterns in the training data that are not likely to exist in future data... so, it will lead to predictions that will vary a lot as things could change a lot depending on what sample we happen to see.
- Our goal is to find the sweet spot in the bias-variance trade-off!!

Bias-Variance Trade-Off

Once again, this is the key idea of the course!!



Bias-Variance Trade-Off

Let’s get back to our original representation of the problem... it helps us understand what is going on...

$$Y_f = f(X_f) + !$$

- We need flexible enough models to find $f(\cdot)$ without imposing bias...
- ... but, too flexible models will “chase” non-existing patterns in $!$ leading to unwanted variability

Bias-Variance Trade-Off

A more detailed look at this idea... assume we are trying to make a prediction for Y_f using X_f and our inferred $f(\cdot)$. We hope to make small mistake measured by squared distance... Let's explore how our mistakes will behave *on average*.

$$
\begin{aligned}
\mathrm{E}\left[(Y_f - \hat{f}(X_f))^2\right] &= \left(f(X_f) - \mathrm{E}\left[\hat{f}(X_f)\right]\right)^2 + \mathrm{Var}\left[\hat{f}(X_f)\right] + \mathrm{Var}(\epsilon) \\
&= \mathrm{Bias}\left[\hat{f}(X_f)\right]^2 + \mathrm{Var}\left[\hat{f}(X_f)\right] + \mathrm{Var}(\epsilon)
\end{aligned}
$$

hence, the *Bias-Variance Trade-Off!!*

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 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?

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:

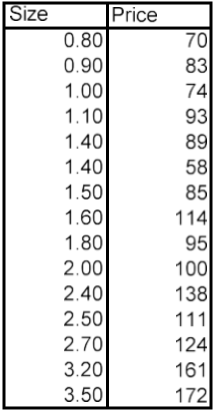
- Y = price of house (e.g. thousands of dollars)

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

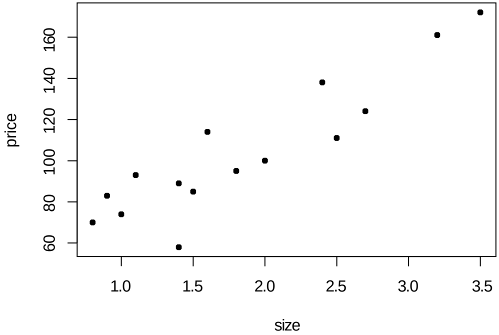
What does this data look like?



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

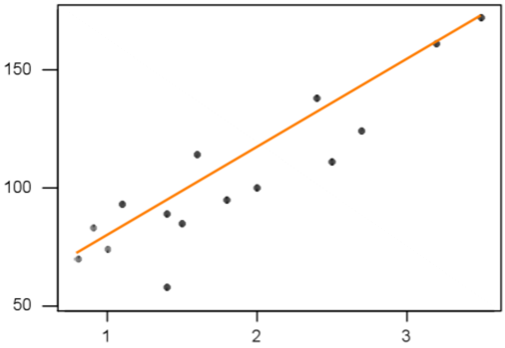
Y = response or outcome variable
X = explanatory or input variables

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:



Linear prediction

Recall that the equation of a line is:

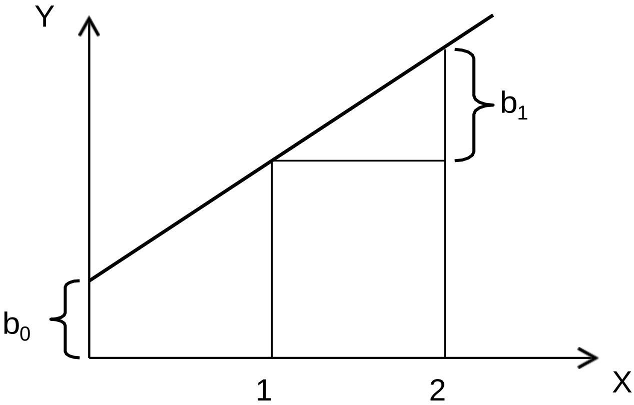
$$Y = b_0 + b_1X$$

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



$Y = b_0 + b_1X$

Linear prediction

Q: How to find the "best line"?

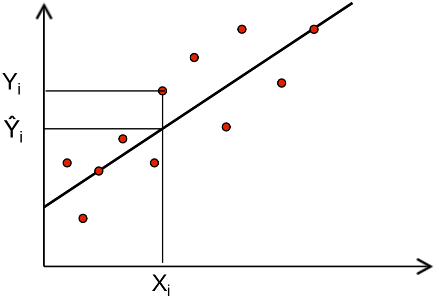
We desire a strategy for estimating the slope and intercept parameters in the model
$\hat{Y} = b_0 + b_1X$

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

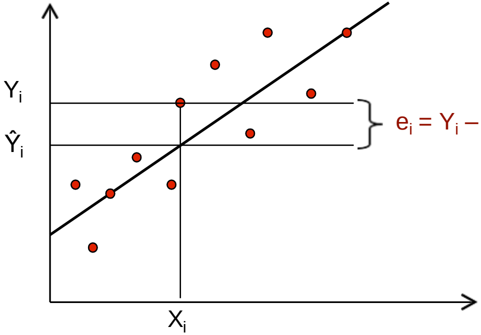
What is the “fitted value”?



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?



$e_i = Y_i – \hat{Y}_i$ = Residual *i*

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.

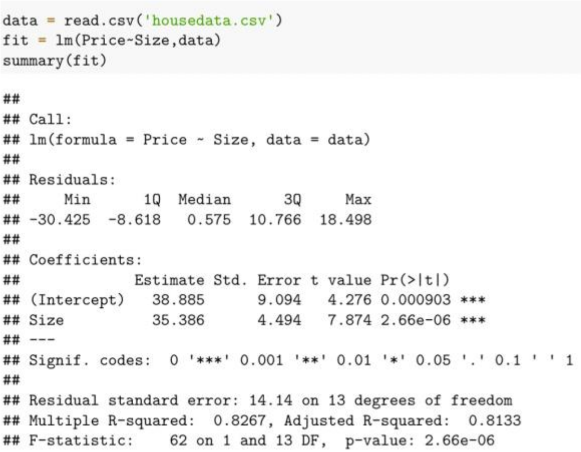
The line fitting process:

- 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



```
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
```

Example 2: Offensive performance in baseball

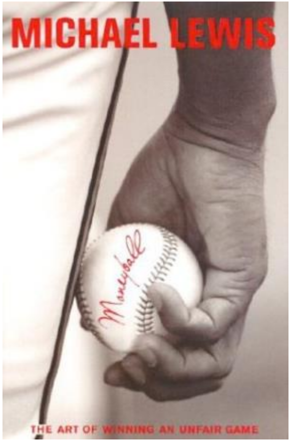
Problems:

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

Solutions:

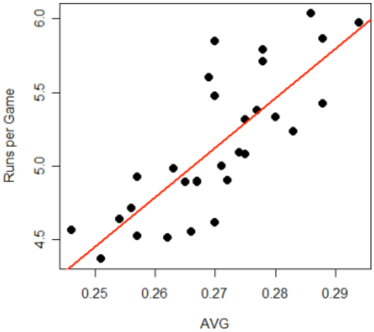
- Compare *prediction rules* that forecast runs as a function of either AVG (batting average), SLG (slugging percentage – total bases divided by at bats) or OBP (on base percentage)

Example 2: Offensive performance in baseball



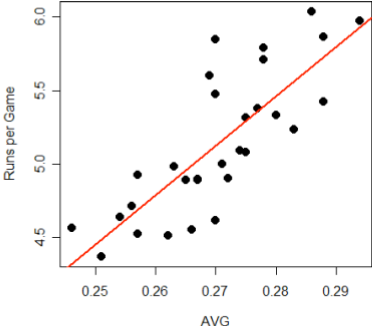
Baseball data – using AVG

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

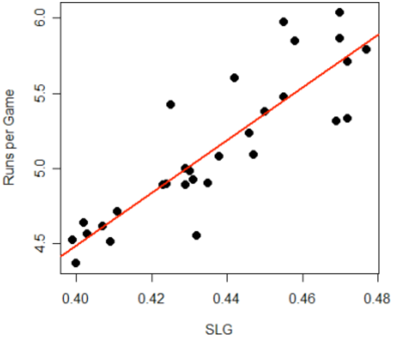
Baseball data – using AVG



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

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

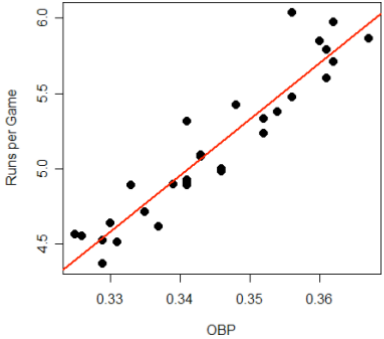
Baseball Data – using SLG



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

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

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:

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

where,

- $\overline{X}$ and $\overline{Y}$ 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

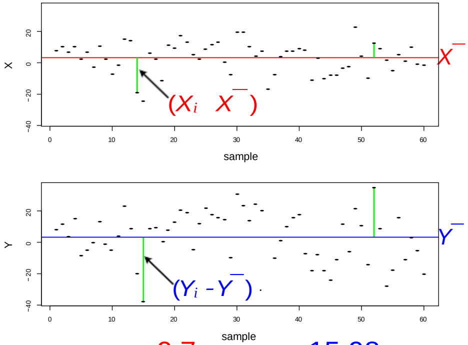
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

- Sample Standard Deviation:

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

Visual: Standard deviation

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

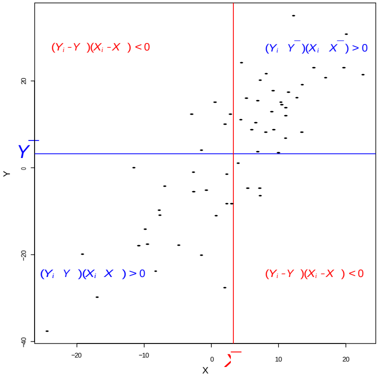


$s_x = 9.7$ $s_y = 15.98$

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:

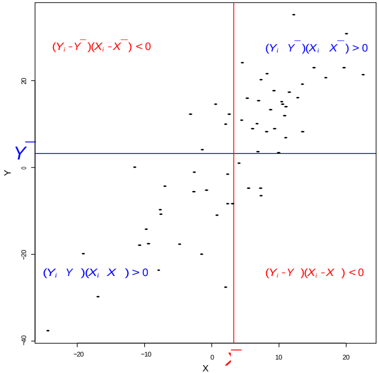
$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{3\,\overline{s_x^2 s_y^2}} = \frac{\operatorname{cov}(X, Y)}{s_x s_y}$$

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

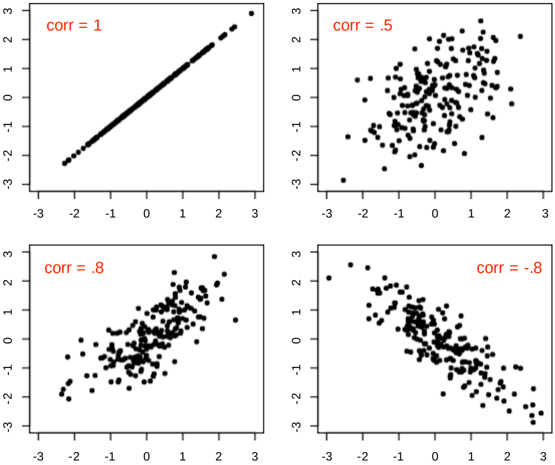
This gives the direction (negative or positive) and strength (0 → 1)
of the linear relationship between X and Y.

Correlation

$$\text{corr}(Y,X) = \frac{\text{cov}(X,Y)}{3\,\overline{s_x^2 s_y^2}} = \frac{\text{cov}(X,Y)}{s_x s_y} = \frac{125.9}{15.98 \times 9.7} = 0.812$$



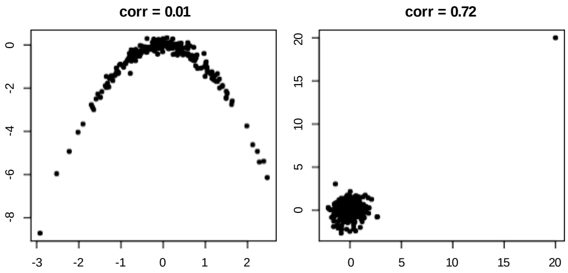
Correlation



Correlation

Only measures linear relationships:

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:

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

$$= \frac{\text{cov}(X, Y)}{\text{var}(X)}$$

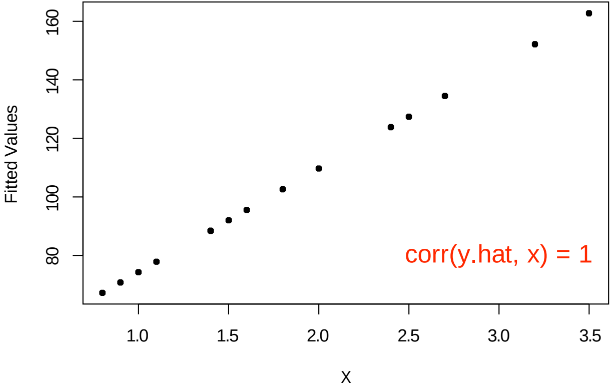
So, the right slope is the **correlation coefficient** times a **scaling factor** that ensures the proper units for b_1

More on least squares

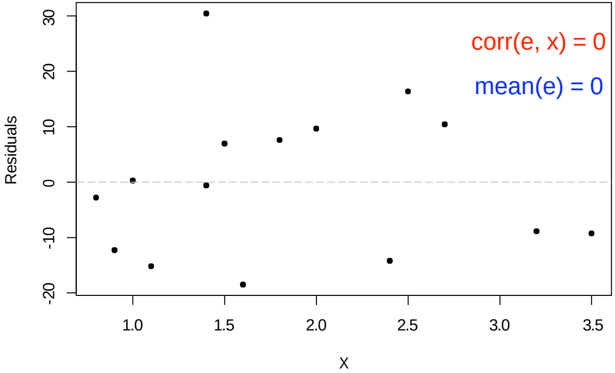
From now on, terms “fitted values” ($\hat{Y}_i$) and “residuals” (e_i) refer to those obtained from the least squares line.

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

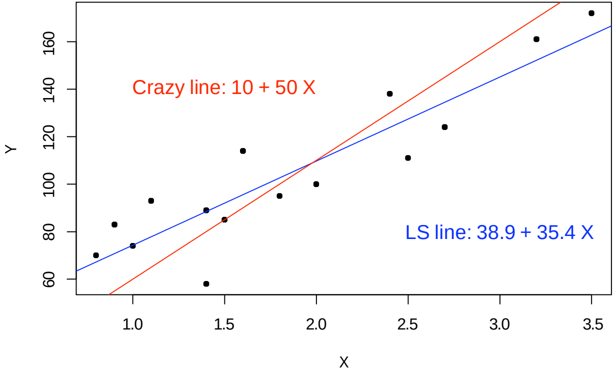


The residuals and X



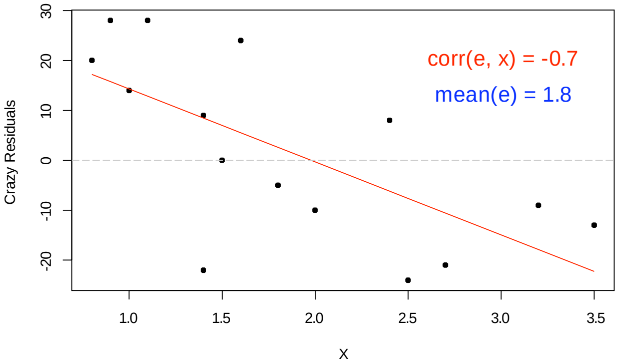
Why?

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

Decomposing the variance

Q: How well does the least squares line explain variation in Y?

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

Since $\hat{Y}$ and e are uncorrelated, i.e. $\mathrm{corr}(\hat{Y}, e) = 0$,

$$\mathrm{var}(Y) = \mathrm{var}(\hat{Y} + e) = \mathrm{var}(\hat{Y}) + \mathrm{var}(e)$$

$$\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{n-1} + \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1}$$

Given that $\bar{e} = 0$, and $\bar{\hat{Y}} = \bar{Y}$ (why?) we get to:

$$\boxed{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2}$$

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.

SSR = SST ⇒ 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:

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

- $0 < R^2 < 1$.
- 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 the modeling goal

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

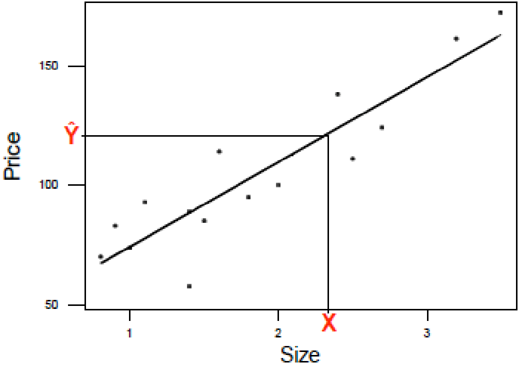
The least squares line is a prediction rule:

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

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 the modeling goal

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!

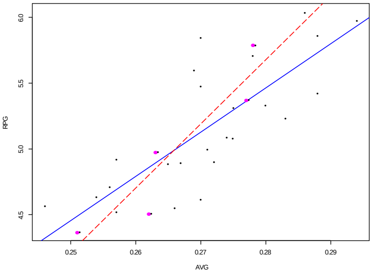
We will build a probability model (e.g., Normal distribution).

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.

Prediction and the modeling goal

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.

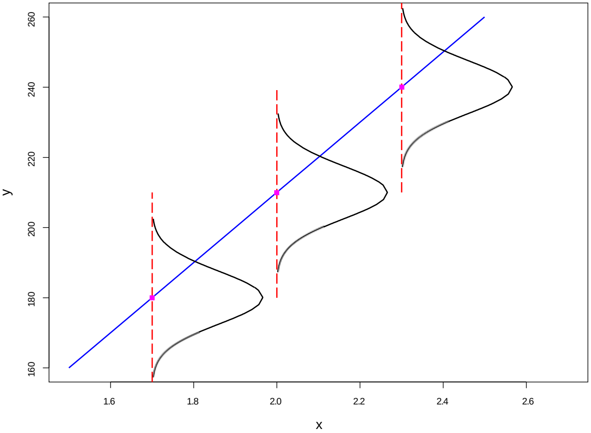
In order to do this we must invest in a probability model.

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.

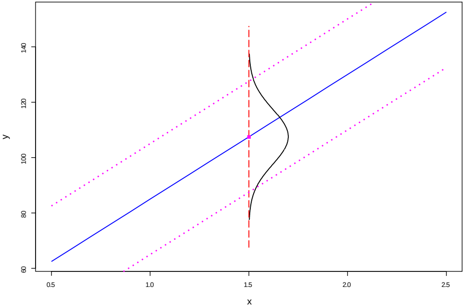
What do you know about Y from the model?

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

Thus our prediction for price is $Y \mid (X = 1.5) \sim N(107.5, 10^2)$
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:

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

Why do we choose this probability model?

Put differently, why do we have $\varepsilon \sim \mathrm{N}(0, \sigma^2)$?

- $E[\varepsilon] = 0 \Leftrightarrow E[Y \mid X] = \beta_0 + \beta_1 X$
  ($E[Y \mid X]$ is "conditional expectation of Y given X").
- Many things are close to Normal (central limit theorem).
- It works! This is a very robust model for the world.

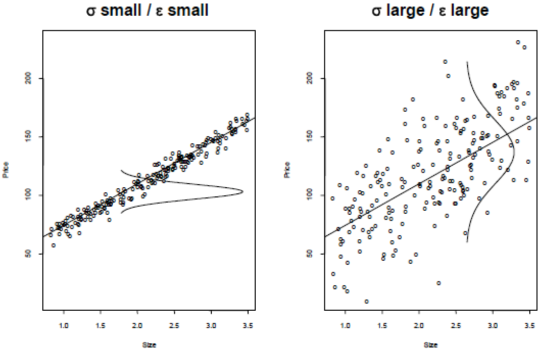
We can think of $\beta_0 + \beta_1 X$ as the "true" regression line.

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:



Digging into the moments of the SLR model

More on the conditional distribution:

$$Y|X \sim \mathrm{N}(E[Y|X], \mathrm{var}(Y|X)).$$

- The conditional mean is $E[Y|X] = E[\beta_0 + \beta_1 X + \varepsilon] = \beta_0 + \beta_1 X$.
- The conditional variance is $\mathrm{var}(Y|X) = \mathrm{var}(\beta_0 + \beta_1 X + \varepsilon) = \mathrm{var}(\varepsilon) = \sigma^2$.
- $\sigma^2 < \mathrm{var}(Y)$ if X and Y are related.

Summary of simple linear regression

Assume that all observations are drawn from our regression model and that errors on those observations are independent.

The model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where ε is independent and identically distributed $\mathrm{N}(0, \sigma^2)$.

- independence means that knowing ε_i doesn’t affect your views about ε_j
- identically distributed means that we are using the same Normal for every ε_i

Key characteristics of linear regression model

- Mean of Y is linear in X.
- Error terms (deviations from line) are Normally distributed (very few deviations are more than 2 standard devations away from the regression mean).
- Error terms have constant variance.

Estimation for the SLR model

SLR assumes every observation in the dataset was generated by the model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

This is a model for the conditional distribution of Y given X.

We use Least Squares *to estimate* β_0 and β_1:

$$\hat{\beta}_1 = b_1 = r_{xy} \times \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = b_0 = \overline{Y} - b_1 \overline{X}$$

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.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \$$$

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:

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

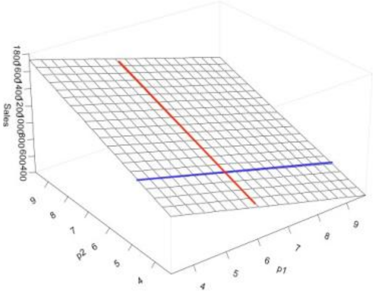
Holding all other variables constant, β_j is the average change in Y per unit change in X_j.

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).

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



Least squares again!

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

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!

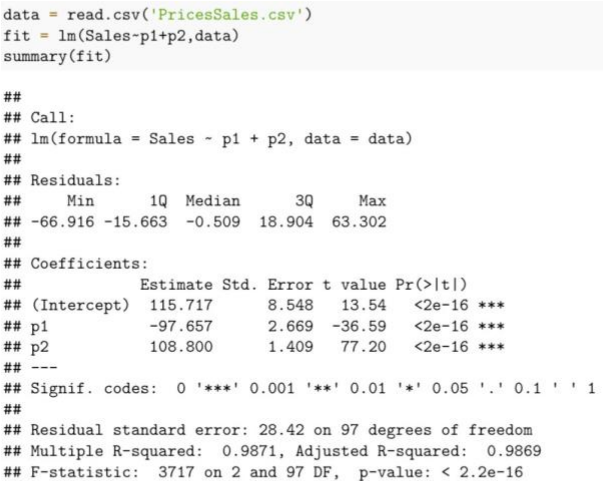
The data...

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

Fitting the MLR model



```r
data = read.csv('PricesSales.csv')
fit = lm(Sales~p1+p2,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ p1 + p2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.916 -15.663  -0.509  18.904  63.302
##
## 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
```

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 decide to charge \$8. How much will I sell?

Our model is:

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

with $! \sim N(0, \sigma^2)$

Our estimates are $b_0 = 115$, $b_1 = -97$, $b_2 = 109$ and $s = 28$ which leads to

$$\text{Sales} = 115 + -97 * P1 + 109 * P2 + !$$

with $! \sim N(0, 28^2)$

Plug-in Prediction in MLR

By plugging-in the numbers,

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

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

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

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

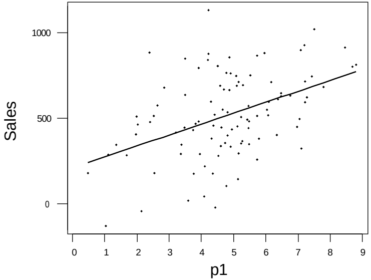
But what about the right-hand-side?

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

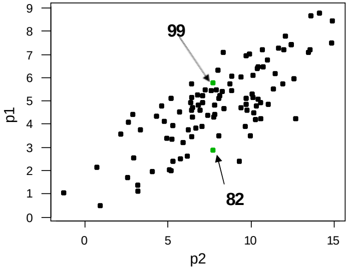
$$Sales = 116 - 97.7P1 + 109P2$$

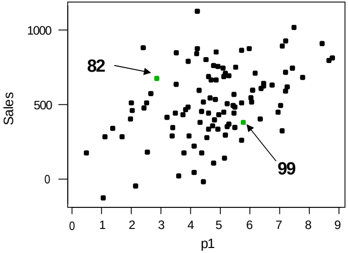
Does this look better? How did it happen? Remember: −97.7 is the affect on sales of a change in *P*1 **with *P*2 held fixed!**

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 *P*1, holding *P*2 constant, corresponds to a drop in Sales!

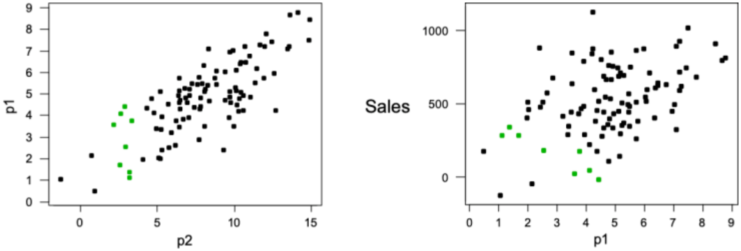




Note the strong relationship (dependence) between *P*1 and *P*2!

Understanding multiple regression

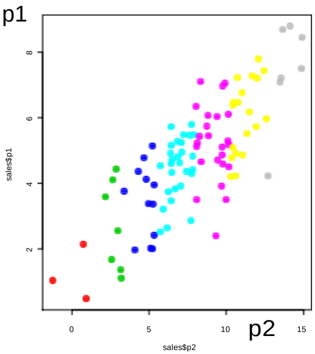
Let’s look at a subset of points where *P*1 varies and *P*2 is held approximately constant...



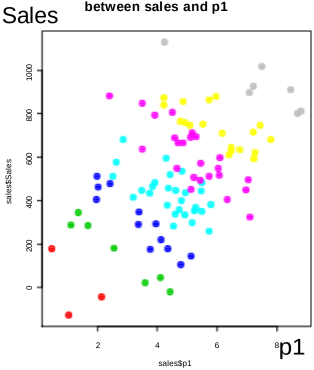
For a fixed level of *P*2, variation in *P*1 is negatively correlated with Sales!

Understanding multiple regression

Below, different colors indicate different ranges for *P*2...



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 *P*1 is associated with larger *P*2 and the overall effect leads to bigger sales

→ With *P*2 held fixed, a larger *P*1 leads to lower sales

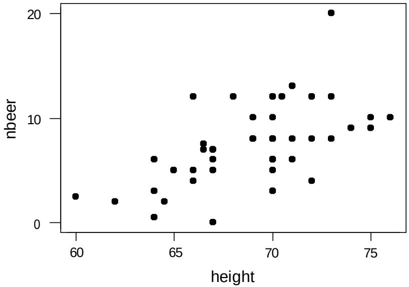
→ MLR does the trick and unveils the **correct** economic relationship between Sales and prices!

Example: Beers, height, weight, and getting drunk

Beer data (from class last year)

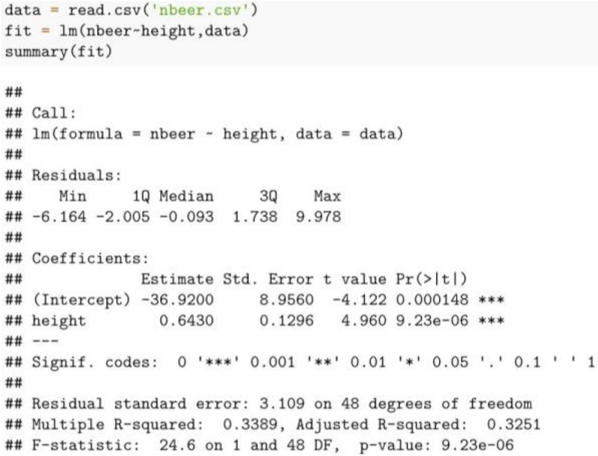
– **nbeer** – number of beers before getting drunk

– **height** and **weight**



Is number of beers related to height?

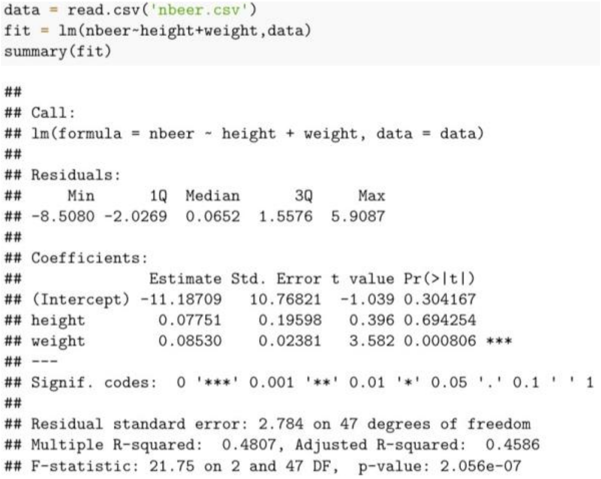
R output: Yes!



```
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
```

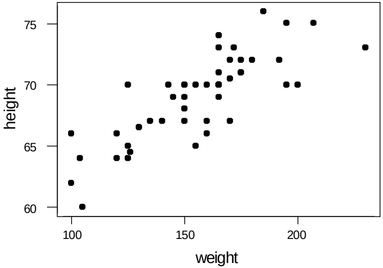
R output: What about now?



```
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



The correlations:

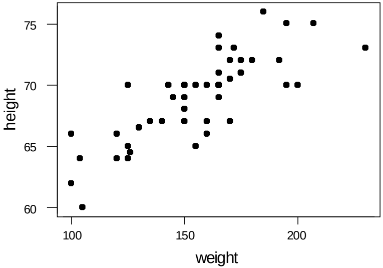
```
           nbeer   weight
weight     0.692
height     0.582    0.806
```

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



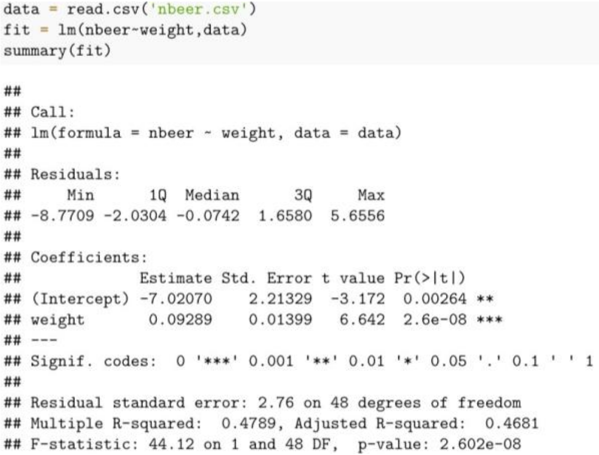
The correlations:

```
         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.

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

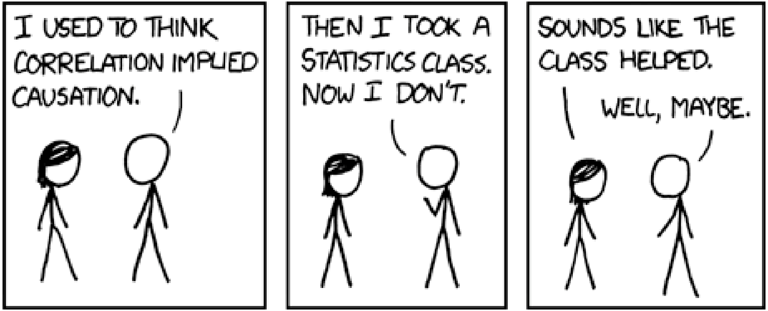
In general, when we see a relationship between *y* and *x* (or *x*'s), that relationship may be driven by variables "lurking" in the background which are related to your current *x*'s.

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


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.