THE UNIVERSITY OF TEXAS AT AUSTIN

McCOMBS BUSINESS ANALYTICS

Regression and Model Selection

Book Chapters 3 and 6.

Carlos M. Carvalho
The University of Texas McCombs School of Business

1. Simple Linear Regression
2. Multiple Linear Regression
3. Dummy Variables
4. Residual Plots and Transformations
5. Variable Selection and Regularization
6. Dimension Reduction Methods

1. 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

1st Example: Predicting House Prices

Problem:

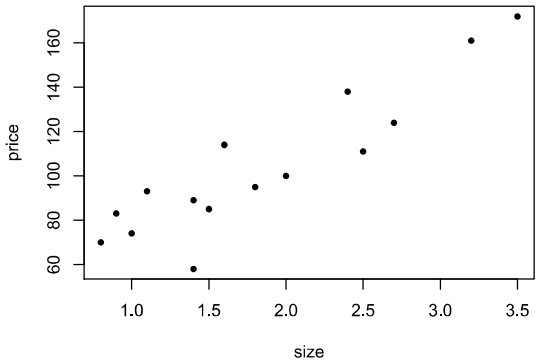
- Predict market price based on observed characteristics

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

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
$X1, X2, X3, \ldots, Xp$ = explanatory or input variables

The general relationship approximated by:

$$Y = f(X_1, X_2, \ldots, X_p) + e$$

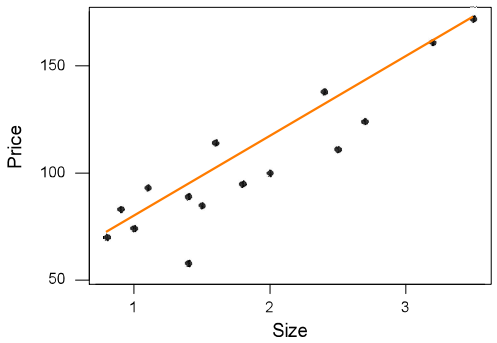
And a linear relationship is written

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p + e$$

Linear Prediction

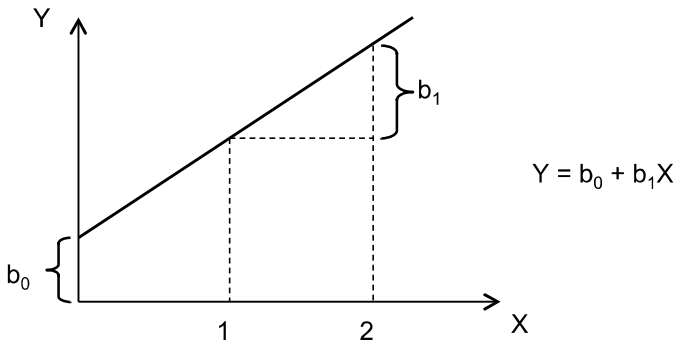
Appears to be a linear relationship between price and size:

As size goes up, price goes up.



The line shown was fit by the "eyeball" method.

Linear Prediction



$Y = b_0 + b_1 X$

Our "eyeball" line has $b_0 = 35$, $b_1 = 40$.

Linear Prediction

Can we do better than the eyeball method?

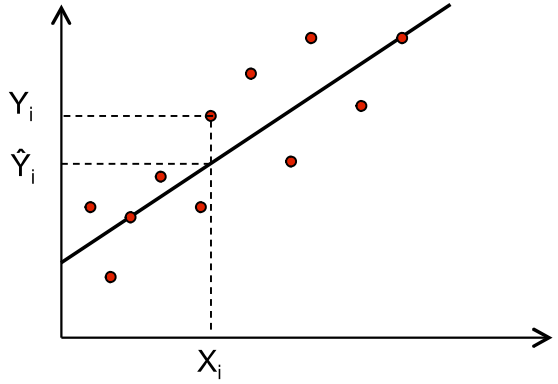
We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

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

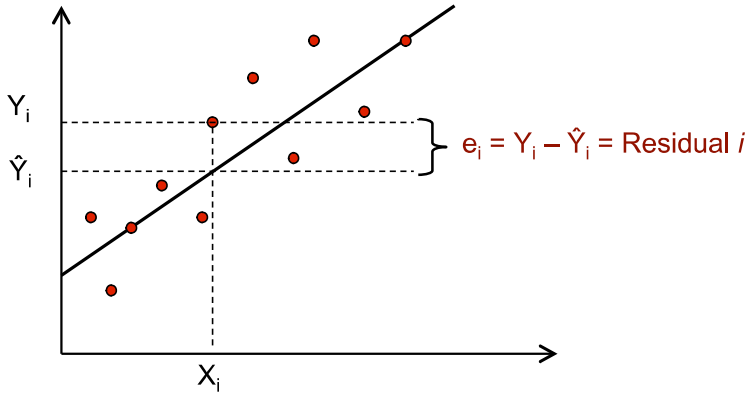
What is the "fitted value"?



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

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

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

$$= \sum_{i=1}^{n}(Y_i - [b_0 + b_1 X_i])^2$$

Least Squares – Excel Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

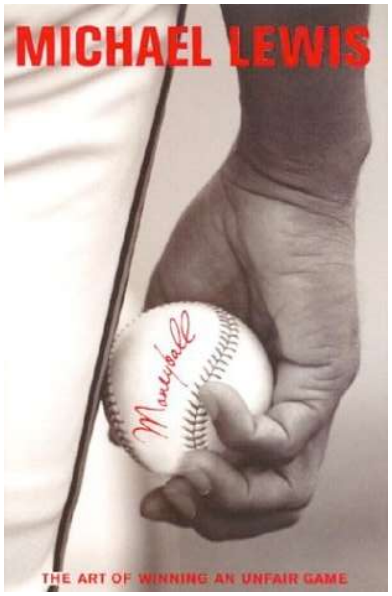
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763
Size	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846

2nd Example: Offensive Performance in Baseball

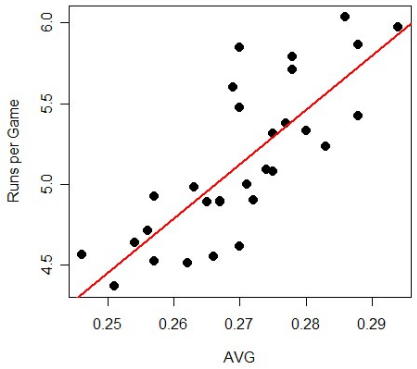
1. Problems:
   - Evaluate/compare traditional measures of offensive performance
   - Help evaluate the worth of a player
2. Solutions:
   - Compare *prediction rules* that forecast runs as a function of either AVG (batting average), SLG (slugging percentage) or OBP (on base percentage)

2nd Example: Offensive Performance in Baseball



Baseball Data – Using AVG

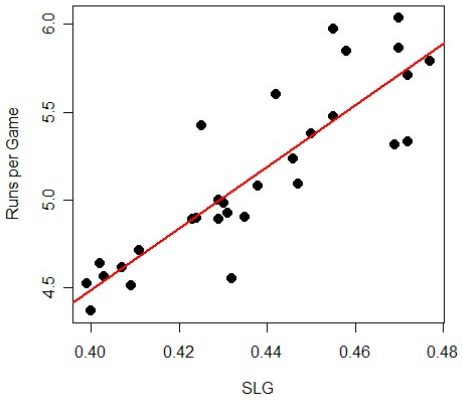
Each observation corresponds to a team in MLB. Each quantity is the average over a season.



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

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

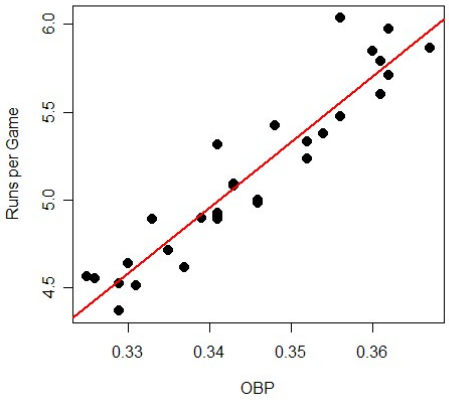
Baseball Data – Using SLG



- 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

Place your Money on OBP!!!

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

	Average Squared Error
AVG	0.083
SLG	0.055
OBP	0.026

The Least Squares Criterion

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

Sample Mean and Sample Variance

- Sample Mean: measure of centrality

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

- Sample Variance: measure of spread

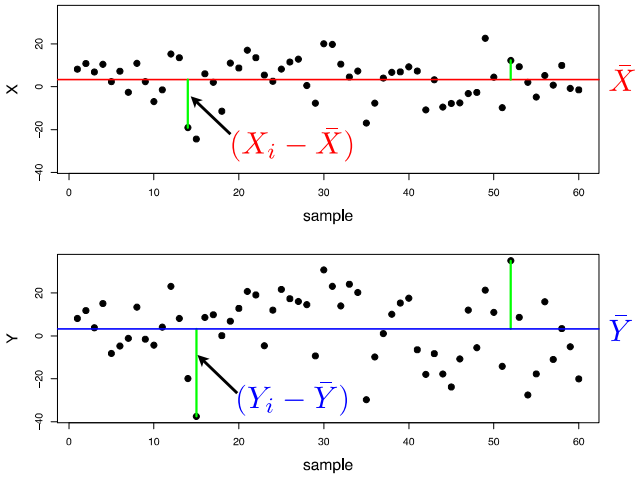
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

- Sample Standard Deviation:

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

Example

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

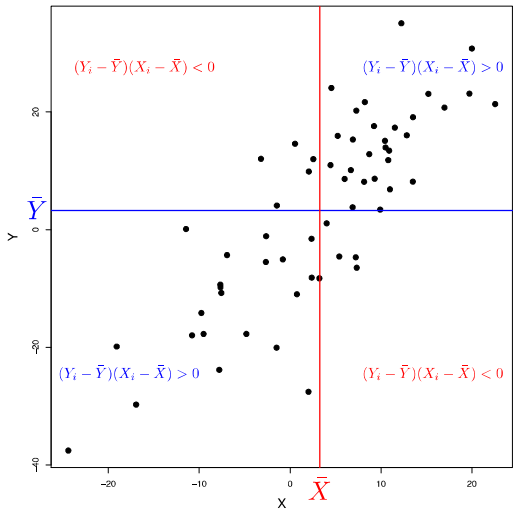


$s_x = 9.7$ $s_y = 16.0$

Covariance

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

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



- $s_y = 15.98$, $s_x = 9.7$
- $Cov(X, Y) = 125.9$

How do we interpret that?

Correlation

Correlation is the standardized covariance:

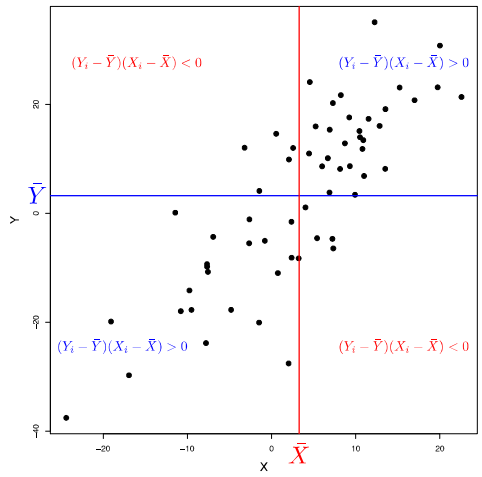
$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\text{cov}(X, Y)}{s_x s_y}$$

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

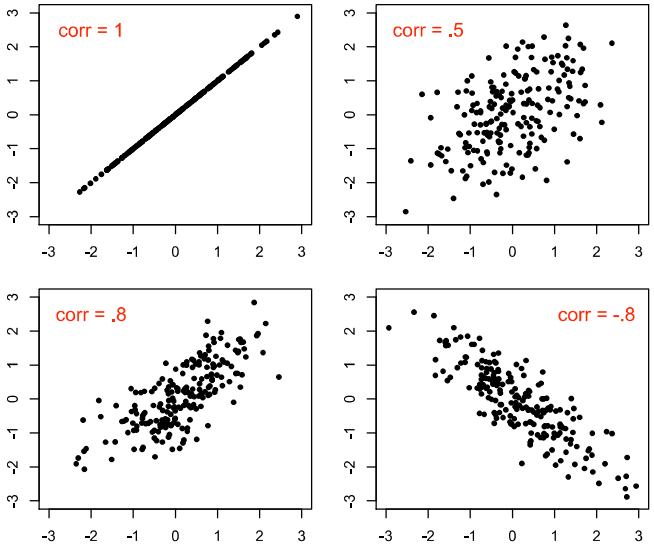
This gives the direction (- or +) and strength ($0 \rightarrow 1$) of the linear relationship between X and Y.

Correlation

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

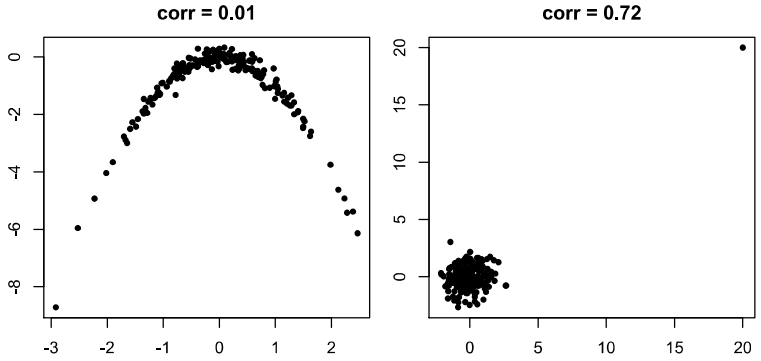


Correlation



Correlation

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



Also be careful with influential observations.

Back to Least Squares

1. 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 rotate the line through that point until getting the "right" slope

2. Slope:

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

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

Decomposing the Variance

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. $\text{corr}(\hat{Y}, e) = 0$,

$$\begin{aligned} \text{var}(Y) &= \text{var}(\hat{Y} + e) = \text{var}(\hat{Y}) + \text{var}(e) \\ \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} &= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{n-1} + \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1} \end{aligned}$$

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

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

Decomposing the Variance

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

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:

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

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

Back to the House Data

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38.86468274	9.09390389	4.275006499	0.000902712	19.23849785	58.53086763	19.23849785	58.53086763
X Variable 1	39.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846	25.67708664	45.09483846

SSR

SST

SSE

$$R^2 = \frac{SSR}{SST} = 0.82 = \frac{12395}{14991}$$

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

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!

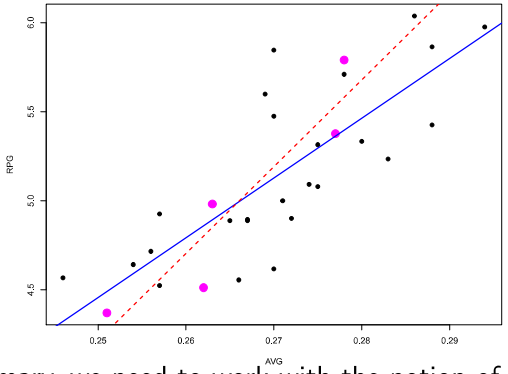
We will build a probability model (e.g., normal distribution).

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

- Suppose you only had the purple points in the graph. 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.

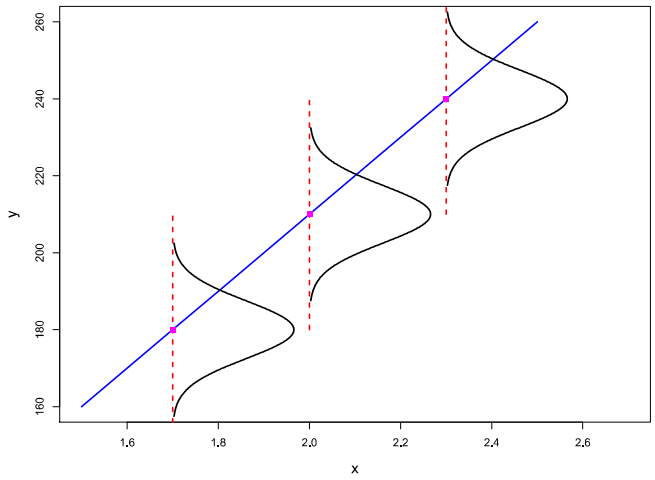
Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

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

- $\beta_0 + \beta_1 X$ 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.

The Simple Linear Regression Model

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



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.

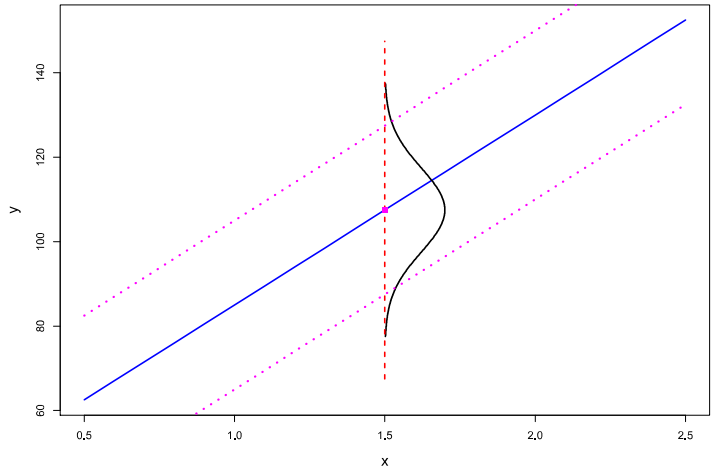
What do you know about Y from the model?

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

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 $87.5 < Y < 127.5$

Conditional Distributions

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



The conditional distribution for Y given X is Normal:

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

Estimation of Error Variance

We estimate s^2 with:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2}$$

(2 is the number of regression coefficients; i.e. 2 for β_0 and β_1).

We have $n-2$ degrees of freedom because 2 have been "used up" in the estimation of b_0 and b_1.

We usually use $s = \sqrt{SSE/(n-2)}$, in the same units as Y. It's also called the regression standard error.

Estimation of Error Variance

Where is s in the Excel output?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

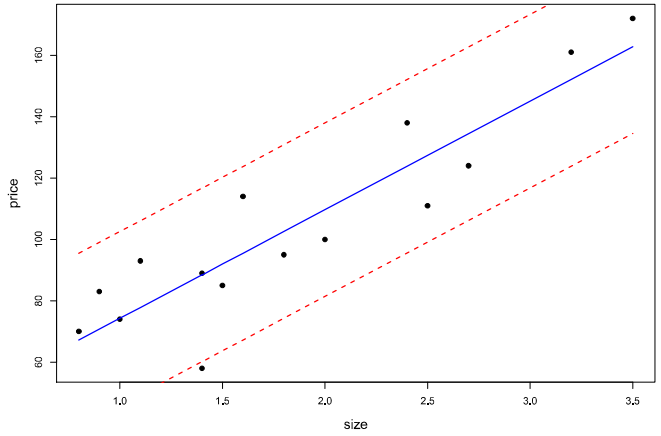
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763	19.23849785	58.53086763
X Variable 1	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846	25.67708664	45.09483846

Remember that whenever you see "standard error" read it as estimated standard deviation: σ is the standard deviation.

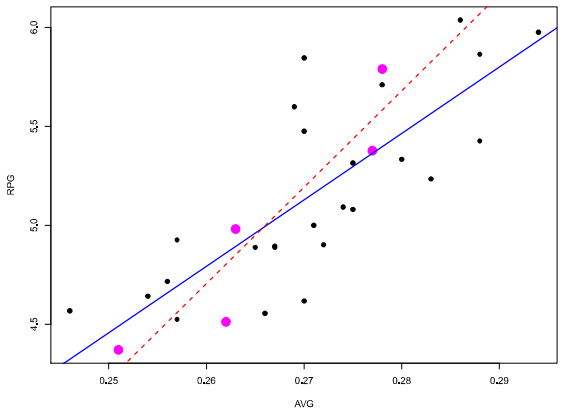
One Picture Summary of SLR

- The plot below has the house data, the fitted regression line ($b_0 + b_1 X$) and $\pm 2 * s$...
- From this picture, what can you tell me about β_0, β_1 and σ^2? How about b_0, b_1 and s^2?



Understanding Variation... Runs per Game and AVG

- blue line: all points
- red line: only purple points
- Which slope is closer to the true one? How much closer?



The Importance of Understanding Variation

When estimating a quantity, it is vital to develop a notion of the precision of the estimation; for example:

- estimate the slope of the regression line
- estimate the value of a house given its size
- estimate the expected return on a portfolio
- estimate the value of a brand name
- estimate the damages from patent infringement

Why is this important?
We are making decisions based on estimates, and these may be very sensitive to the accuracy of the estimates!

Sampling Distribution of b_1

The sampling distribution of b_1 describes how estimator $b_1 = \hat{\beta}_1$ varies over different samples with the X values fixed.

It turns out that b_1 is normally distributed (approximately): $b_1 \sim N(\beta_1, s_{b_1}^2)$.

- b_1 is unbiased: $E[b_1] = \beta_1$.
- s_{b_1} is the standard error of b_1. In general, the standard error is the standard deviation of an estimate. It determines how close b_1 is to β_1.
- This is a number directly available from the regression output.

Sampling Distribution of b_1

Can we intuit what should be in the formula for s_{b_1}?

- How should s figure in the formula?
- What about n?
- Anything else?

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

Three Factors:
sample size (n), error variance (s^2), and X-spread (s_x).

Sampling Distribution of b_0

The intercept is also normal and unbiased: $b_0 \sim N(\beta_0, s^2_{b_0})$.

$$s^2_{b_0} = \text{var}(b_0) = s^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s^2_x} \right)$$

What is the intuition here?

Understanding Variation... Runs per Game and AVG

Regression with all points

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.798496529
R Square	0.637596707
Adjusted R Square	0.624653732
Standard Error	0.298493066
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4.38915033	4.38915	49.26199	1.239E-07
Residual	28	2.494747094	0.089098		
Total	29	6.883897424			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-3.936410446	1.294049995	-3.04193	0.005063	-6.587152	-1.2856692
AVG	33.57186945	4.783211061	7.018689	1.24E-07	23.773906	43.369833

$s_{b_1} = 4.78$

Understanding Variation... Runs per Game and AVG

Regression with subsample

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.933601392
R Square	0.87161156
Adjusted R Square	0.828815413
Standard Error	0.244815842
Observations	5

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.220667405	1.220667	20.36659	0.0203329
Residual	3	0.17980439	0.059935		
Total	4	1.400471795			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.956288201	2.874375987	-2.76801	0.069684	-17.10384	1.191259
AVG	48.69444328	10.78997028	4.512936	0.020333	14.355942	83.03294

$s_{b_1} = 10.78$

Confidence Intervals

- 68% Confidence Interval: $b_1 \pm 1 \times s_{b_1}$
- 95% Confidence Interval: $b_1 \pm 2 \times s_{b_1}$
- 99% Confidence Interval: $b_1 \pm 3 \times s_{b_1}$

Same thing for b_0

- 95% Confidence Interval: $b_0 \pm 2 \times s_{b_0}$

The confidence interval provides you with a set of plausible values for the parameters

Example: Runs per Game and AVG

Regression with all points

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.798496529
R Square	0.637596707
Adjusted R Square	0.624653732
Standard Error	0.298493066
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	4.38915033	4.38915	49.26199	1.239E-07
Residual	28	2.494747094	0.089098		
Total	29	6.883897424			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-3.936410446	1.294049995	-3.04193	0.005063	-6.587152	-1.2856692
AVG	33.57186945	4.783211061	7.018689	1.24E-07	23.773906	43.369833

$$[b_1 - 2 \times s_{b_1}; b_1 + 2 \times s_{b_1}] \approx [23.77; 43.36]$$

Testing

Suppose we want to assess whether or not β_1 equals a proposed value β_1^0. This is called hypothesis testing.

Formally we test the null hypothesis:

$H_0 : \beta_1 = \beta_1^0$

vs. the alternative

$H_1 : \beta_1 \neq \beta_1^0$

Testing

That are 2 ways we can think about testing:

1. Building a test statistic... the t-stat,

$$t = \frac{b_1 - \beta_1^0}{s_{b_1}}$$

   This quantity measures how many standard deviations the estimate (b_1) from the proposed value (β_1^0).

   If the absolute value of t is greater than 2, we need to worry (why?)... we reject the hypothesis.

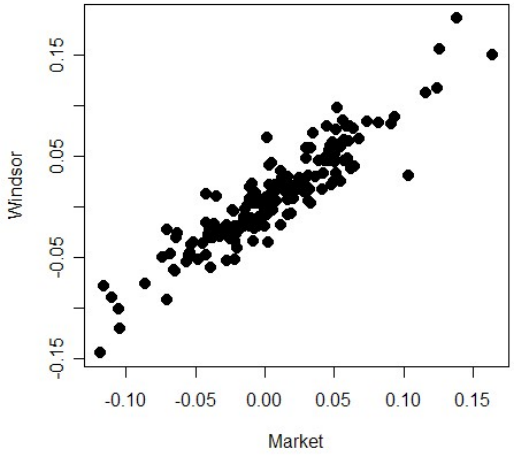
2. Looking at the confidence interval. If the proposed value is outside the confidence interval you reject the hypothesis.

   Notice that this is equivalent to the t-stat. An absolute value for t greater than 2 implies that the proposed value is outside the confidence interval... therefore reject.

   This is my preferred approach for the testing problem. You can't go wrong by using the confidence interval!

Example: Mutual Funds

Let's investigate the performance of the Windsor Fund, an aggressive large cap fund by Vanguard...



The plot shows monthly returns for Windsor vs. the S&P500

Example: Mutual Funds

Consider a CAPM regression for the Windsor mutual fund.

$$r_w = \beta_0 + \beta_1 r_{sp500} + \epsilon$$

Let's first test $\beta_1 = 0$

$H_0 : \beta_1 = 0$. Is the Windsor fund related to the market?
$H_1 : \beta_1 \neq 0$

Example: Mutual Funds

Regression Statistics	
Multiple R	0.923417768
R Square	0.852700374
Adjusted R Square	0.851872848
Standard Error	0.018720015
Observations	180

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.3611	0.361099761	1030.421266	6.0291E-76
Residual	178	0.062378	0.000350439		
Total	179	0.423478			

	Coefficients	*tandard Em*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	0.003646881	0.001409	2.587596412	0.010462425	0.000865657	0.006428	0.000866	0.006428
X Variable 1	0.935717012	0.02915	32.10017549	6.0291E-76	0.878193151	0.993241	0.878193	0.993241

b_1 $\quad s_{b_1}$ $\quad \frac{b_1}{s_{b_1}}$

- $t = 32.10...$ reject $\beta_1 = 0$!!
- the 95% confidence interval is $[0.87; 0.99]$... again, reject!!

Example: Mutual Funds

Now let's test $\beta_1 = 1$. What does that mean?

H_0 : $\beta_1 = 1$ Windsor is as risky as the market.

H_1 : $\beta_1 \neq 1$ and Windsor softens or exaggerates market moves.

We are asking whether or not Windsor moves in a different way than the market (e.g., is it more conservative?).

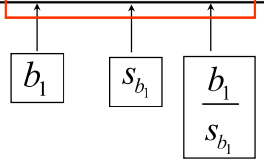
Example: Mutual Funds

Regression Statistics	
Multiple R	0.923417768
R Square	0.852700374
Adjusted R Square	0.851872848
Standard Error	0.018720015
Observations	180

ANOVA

	df	SS	MS	F	Significance F
Regression	1	0.3611	0.361099761	1030.421266	6.0291E-76
Residual	178	0.062378	0.000350439		
Total	179	0.423478			

	Coefficients	tandard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Upper 95.0%
Intercept	0.003646881	0.001409	2.587596412	0.010462425	0.000865657	0.006428	0.000866	0.006428
X Variable 1	0.935717012	0.02915	32.10017549	6.0291E-76	0.878193151	0.993241	0.878193	0.993241



- $t = \frac{b_1 - 1}{s_{b_1}} = \frac{-0.0643}{0.0291} = -2.205\ldots$ reject.
- the 95% confidence interval is $[0.87; 0.99]$... again, reject, but...

Testing – Why I like Conf. Int.

- Suppose in testing $H_0 : \beta_1 = 1$ you got a t-stat of 6 and the confidence interval was

$$[1.00001, 1.00002]$$

Do you reject $H_0 : \beta_1 = 1$? Could you justify that to you boss? Probably not! (why?)

Testing – Why I like Conf. Int.

- Now, suppose in testing $H_0 : \beta_1 = 1$ you got a t-stat of -0.02 and the confidence interval was

$$[-100, 100]$$

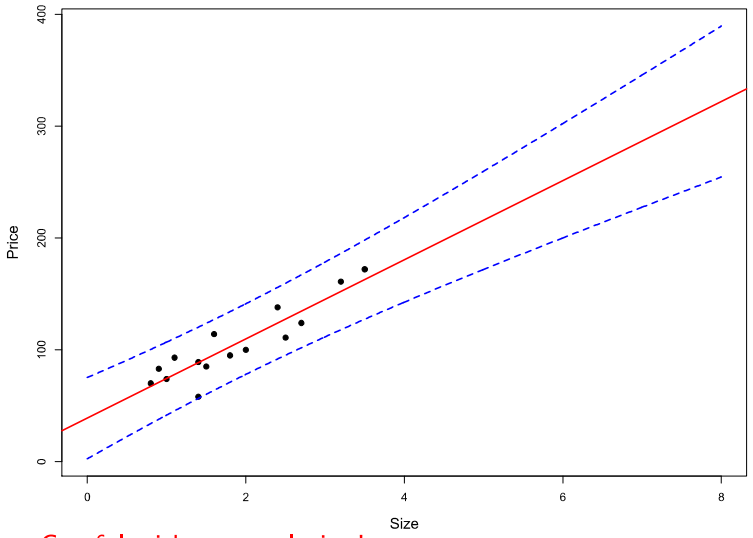
Do you accept $H_0 : \beta_1 = 1$? Could you justify that to you boss? Probably not! (why?)

The Confidence Interval is your best friend when it comes to testing!!

Testing – Summary

- Large t or small p-value mean the same thing...
- p-value < 0.05 is equivalent to a t-stat > 2 in absolute value
- Small p-value means something weird happen if the null hypothesis was true...
- Bottom line, small p-value $\rightarrow$ REJECT! Large $t \rightarrow$ REJECT!
- But remember, always look at the confidence interval!

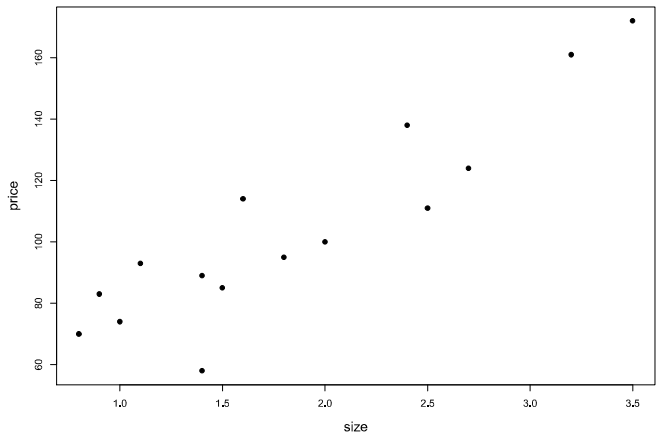
Forecasting



- Careful with extrapolation!

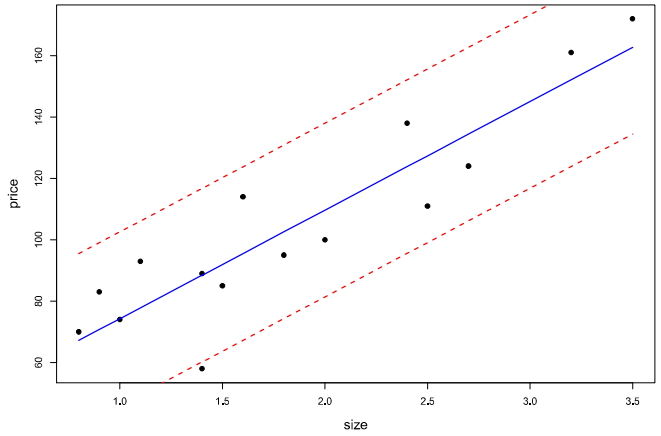
House Data – one more time!

- $R^2 = 82\%$
- Great R^2, we are happy using this model to predict house prices, right?



House Data – one more time!

- But, $s = 14$ leading to a predictive interval width of about US$60,000!! How do you feel about the model now?
- As a practical matter, s is a much more relevant quantity than R^2. Once again, *intervals* are your friend!



2. The Multiple Regression Model

Many problems involve more than one independent variable or factor which affects the dependent or response variable.

- More than size to predict house price!
- Demand for a product given prices of competing brands, advertising,house hold attributes, etc.

In SLR, the conditional mean of Y depends on X. The Multiple Linear Regression (MLR) model extends this idea to include more than one independent variable.

The MLR Model

Same as always, but with more covariates.

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

Recall the key assumptions of our linear regression model:

(i) The conditional mean of Y is linear in the X_j variables.

(ii) The error term (deviations from line)
- are normally distributed
- independent from each other
- identically distributed (i.e., they have constant variance)

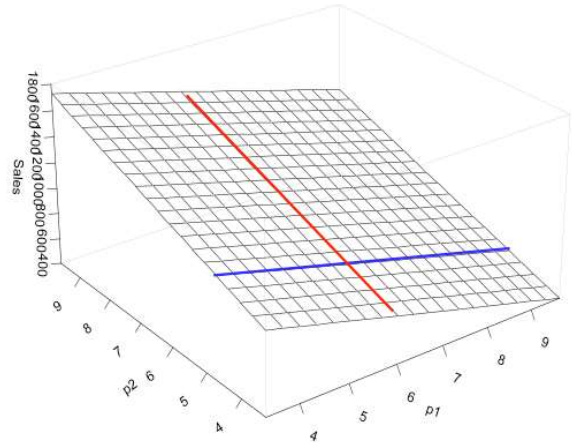
$$Y|X_1 \ldots X_p \sim N(\beta_0 + \beta_1 X_1 \ldots + \beta_p X_p, \sigma^2)$$

The MLR Model

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

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

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

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

Regression Statistics	
Multiple R	0.99
R Square	0.99
Adjusted R Square	0.99
Standard Error	28.42
Observations	100.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.00	6004047.24	3002023.62	3717.29	0.00
Residual	97.00	78335.60	807.58		
Total	99.00	6082382.84			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	115.72	8.55	13.54	0.00	98.75	132.68
p1	-97.66	2.67	-36.60	0.00	-102.95	-92.36
p2	108.80	1.41	77.20	0.00	106.00	111.60

$b_0 = \hat{\beta}_0 = 115.72,\ b_1 = \hat{\beta}_1 = -97.66,\ b_2 = \hat{\beta}_2 = 108.80,$
$s = \hat{\sigma} = 28.42$

Least Squares

Just as before, each b_i is our estimate of β_i

Fitted Values: $\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \ldots + b_p X_p$.

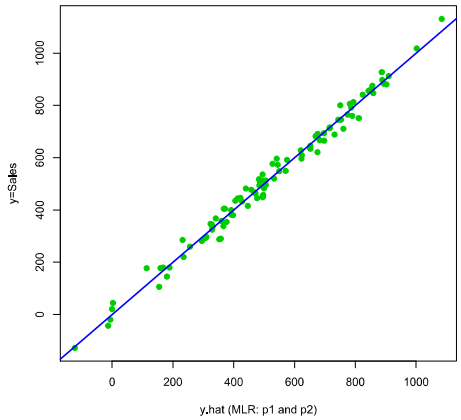
Residuals: $e_i = Y_i - \hat{Y}_i$.

Least Squares: Find $b_0, b_1, b_2, \ldots, b_p$ to minimize $\sum_{i=1}^{n} e_i^2$.

In MLR the formulas for the b_i's are too complicated so we won't talk about them...

Fitted Values in MLR

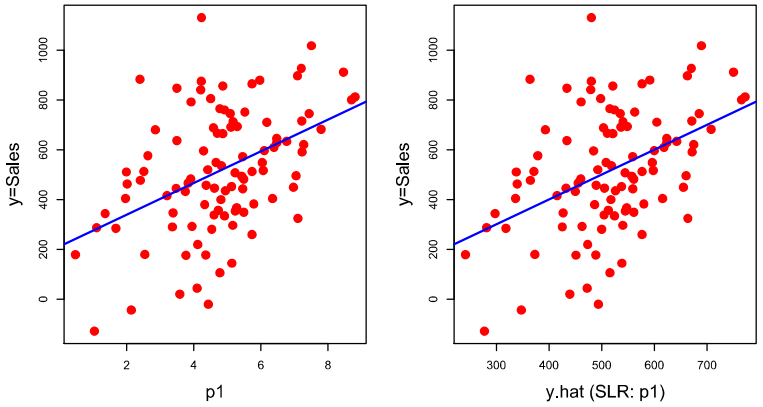
Useful way to plot the results for MLR problems is to look at Y (true values) against $\hat{Y}$ (fitted values).



If things are working, these values should form a nice straight line. Can you guess the slope of the blue line?

Fitted Values in MLR

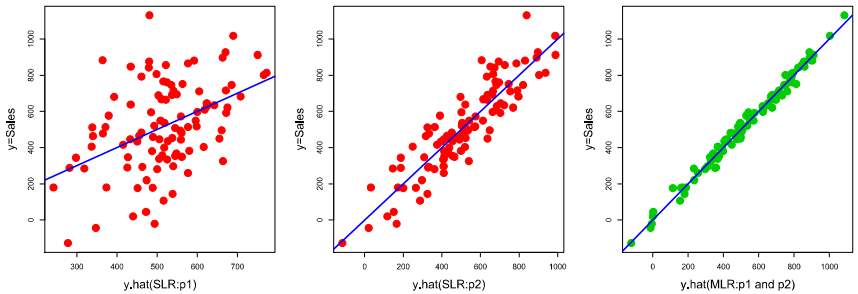
With just $P1$...



- Left plot: *Sales* vs $P1$
- Right plot: Sales vs. $\hat{y}$ (only $P1$ as a regressor)

Fitted Values in MLR

Now, with $P1$ and $P2$...



- First plot: *Sales* regressed on $P1$ alone...
- Second plot: *Sales* regressed on $P2$ alone...
- Third plot: Sales regressed on $P1$ and $P2$

R-squared

- We still have our old variance decomposition identity...

$$SST = SSR + SSE$$

- ... and R^2 is once again defined as

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

  telling us the percentage of variation in Y explained by the X's.
- In Excel, R^2 is in the same place and "Multiple R" refers to the correlation between $\hat{Y}$ and Y.

Intervals for Individual Coefficients

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

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

- We denote the sampling distribution of each estimator as

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

Intervals for Individual Coefficients

Intervals and t-statistics are exactly the same as in SLR.

- A 95% C.I. for β_j is approximately $b_j \pm 2s_{b_j}$
- The t-stat: $t_j = \dfrac{(b_j - \beta_j^0)}{s_{b_j}}$ is the number of standard errors between the LS estimate and the null value (β_j^0)
- As before, we reject the null when t-stat is greater than 2 in absolute value
- Also as before, a small p-value leads to a rejection of the null
- Rejecting when the p-value is less than 0.05 is equivalent to rejecting when the $|t_j| > 2$

Intervals for Individual Coefficients

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

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

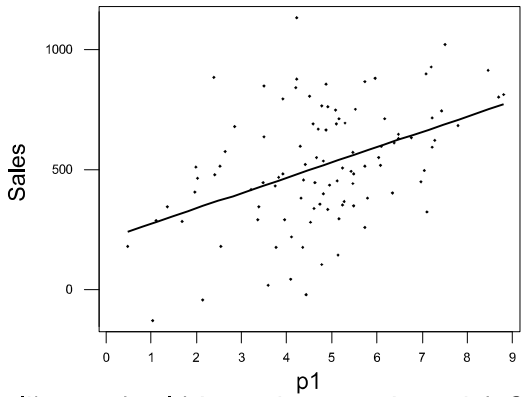
Understanding Multiple Regression

The Sales Data:

- *Sales* : units sold in excess of a baseline
- *P1*: our price in $ (in excess of a baseline price)
- *P2*: competitors price (again, over a baseline)

Understanding Multiple Regression

- If we regress Sales on our own price, we obtain a somewhat surprising conclusion... the higher the price the more we sell!!



- It looks like we should just raise our prices, right? NO, not if you have taken this statistics class!

Understanding Multiple Regression

- The regression equation for Sales on own price (P1) is:

$$Sales = 211 + 63.7P1$$

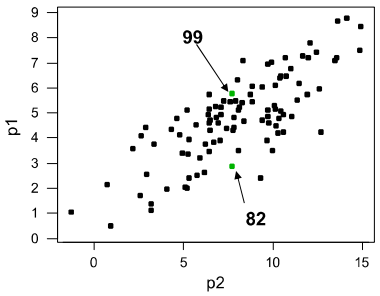
- If now we add the competitors price to the regression we get

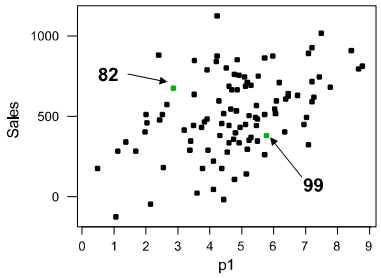
$$Sales = 116 - 97.7P1 + 109P2$$

- 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!

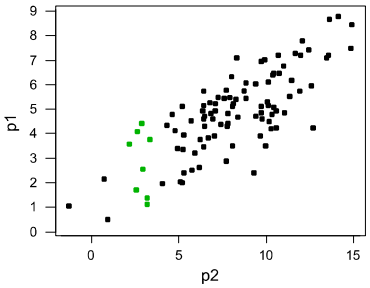


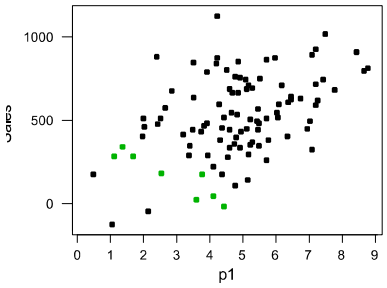


- Note the strong relationship (dependence) between $P1$ and $P2$!!

Understanding Multiple Regression

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

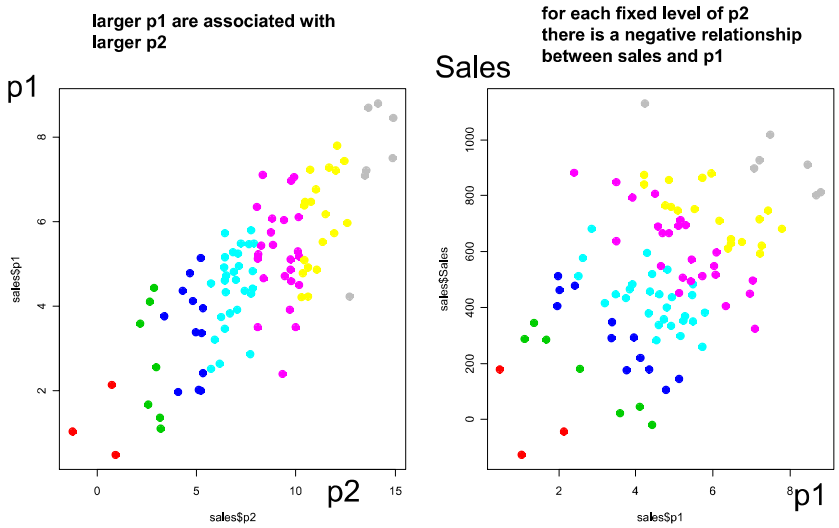




- For a fixed level of $P2$, variation in $P1$ is negatively correlated with Sales!!

Understanding Multiple Regression

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

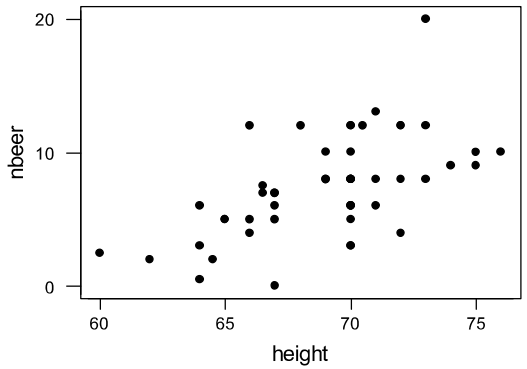


- Summary:
  1. A larger $P1$ is associated with larger $P2$ and the overall effect leads to bigger sales
  2. With $P2$ held fixed, a larger $P1$ leads to lower sales
  3. MLR does the trick and unveils the "correct" economic relationship between Sales and prices!

Understanding Multiple Regression

Beer Data (from an MBA class)

- *nbeer* – number of beers before getting drunk
- *height and weight*



Is number of beers related to height?

Understanding Multiple Regression

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

Regression Statistics	
Multiple R	0.58
R Square	0.34
Adjusted R Square	0.33
Standard Error	3.11
Observations	50.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.00	237.77	237.77	24.60	0.00
Residual	48.00	463.86	9.66		
Total	49.00	701.63			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-36.92	8.96	-4.12	0.00	-54.93	-18.91
height	0.64	0.13	4.96	0.00	0.38	0.90

Yes! Beers and height are related...

Understanding Multiple Regression

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

Regression Statistics	
Multiple R	0.69
R Square	0.48
Adjusted R Square	0.46
Standard Error	2.78
Observations	50.00

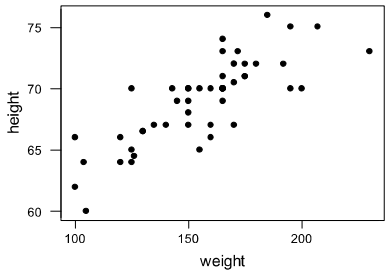
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.00	337.24	168.62	21.75	0.00
Residual	47.00	364.38	7.75		
Total	49.00	701.63			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-11.19	10.77	-1.04	0.30	-32.85	10.48
weight	0.09	0.02	3.58	0.00	0.04	0.13
height	0.08	0.20	0.40	0.69	-0.32	0.47

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

Understanding Multiple Regression



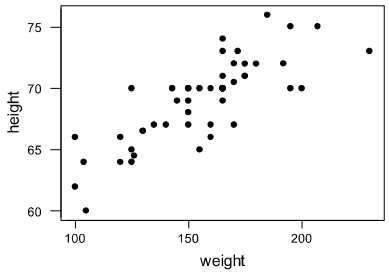
The correlations:

	nbeer	weight
weight	0.692	
height	0.582	0.806

The two x's are highly correlated !!

- If we regress "beers" only on height we see an effect. Bigger heights go with more beers.
- However, when height goes up weight tends to go up as well... in the first regression, height was a proxy for the real *cause* of drinking ability. Bigger people can drink more and weight is a more accurate measure of "bigness".

Understanding Multiple Regression



The correlations:

	nbeer	weight
weight	0.692	
height	0.582	0.806

The two x's are highly correlated !!

- In the multiple regression, when we consider only the variation in height that is not associated with variation in weight, we see no relationship between height and beers.

Understanding Multiple Regression

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

Regression Statistics	
Multiple R	0.69
R Square	0.48
Adjusted R	0.47
Standard E	2.76
Observatio	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	336.0317807	336.0318	44.11878	2.60227E-08
Residual	48	365.5932193	7.616525		
Total	49	701.625			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.021	2.213	-3.172	0.003	-11.471	-2.571
weight	0.093	0.014	6.642	0.000	0.065	0.121

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

Understanding Multiple Regression

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

This makes it hard to reliably find "causal" relationships. Any correlation (association) you find could be caused by other variables in the background... correlation is NOT causation

Any time a report says two variables are related and there's a suggestion of a "causal" relationship, ask yourself whether or not other variables might be the real reason for the effect. Multiple regression allows us to control for all important variables by including them into the regression. "Once we control for weight, height and beers are NOT related"!!

Back to Baseball – Let's try to add AVG on top of OBP

Regression Statistics	
Multiple R	0.948136
R Square	0.898961
Adjusted R Square	0.891477
Standard Error	0.160502
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	6.188355	3.094177	120.1119098	3.63577E-14
Residual	27	0.695541	0.025761		
Total	29	6.883896			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.933633	0.844353	-9.396107	5.30996E-10	-9.666102081	-6.201163
AVG	7.810397	4.014609	1.945494	0.062195793	-0.426899658	16.04769
OBP	31.77892	3.802577	8.357205	5.74232E-09	23.9766719	39.58116

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

Is AVG any good?

Back to Baseball - Now let's add SLG

Regression Statistics	
Multiple R	0.955698
R Square	0.913359
Adjusted R Square	0.906941
Standard Error	0.148627
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	6.28747	3.143735	142.31576	4.56302E-15
Residual	27	0.596426	0.02209		
Total	29	6.883896			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.014316	0.81991	-8.554984	3.60968E-09	-8.69663241	-5.332
OBP	27.59287	4.003208	6.892689	2.09112E-07	19.37896463	35.80677
SLG	6.031124	2.021542	2.983428	0.005983713	1.883262806	10.17899

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

What about now? Is SLG any good

Correlations

AVG	1		
OBP	0.77	1	
SLG	0.75	0.83	1

- When AVG is added to the model with OBP, no additional information is conveyed. AVG does nothing "on its own" to help predict Runs per Game...
- SLG however, measures something that OBP doesn't (power!) and by doing something "on its own" it is relevant to help predict Runs per Game. (Okay, but not much...)

3. Dummy Variables... Example: House Prices

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
...	...	...	...

Dummy Variables... Example: House Prices

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
...	...	...				

Dummy Variables... Example: House Prices

$$Price_i = \beta_0 + \beta_1 dn1_i + \beta_2 dn2_i + \beta_3 Size_i + \epsilon_i$$

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

Dummy Variables... Example: House Prices

$$Price_i = \beta_0 + \beta_1 dn1 + \beta_2 dn2 + \beta_3 Size + \epsilon_i$$

Regression Statistics	
Multiple R	0.828
R Square	0.685
Adjusted R Square	0.677
Standard Error	15.260
Observations	128

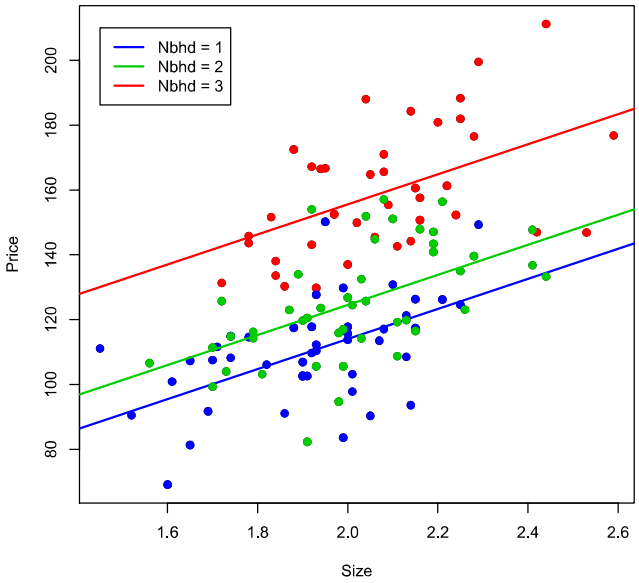
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	3	62809.1504	20936	89.9053	5.8E-31
Residual	124	28876.0639	232.87		
Total	127	91685.2143			

	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*.ower 95%*	*Jpper 95%*
Intercept	62.78	14.25	4.41	0.00	34.58	90.98
dn1	-41.54	3.53	-11.75	0.00	-48.53	-34.54
dn2	-30.97	3.37	-9.19	0.00	-37.63	-24.30
size	46.39	6.75	6.88	0.00	33.03	59.74

$$Price_i = 62.78 - 41.54 dn1 - 30.97 dn2 + 46.39 Size + \epsilon_i$$

Dummy Variables... Example: House Prices



Dummy Variables... Example: House Prices

$$Price_i = \beta_0 + \beta_1 Size + \epsilon_i$$

Regression Statistics	
Multiple R	0.553
R Square	0.306
Adjusted R Square	0.300
Standard Error	22.476
Observations	128

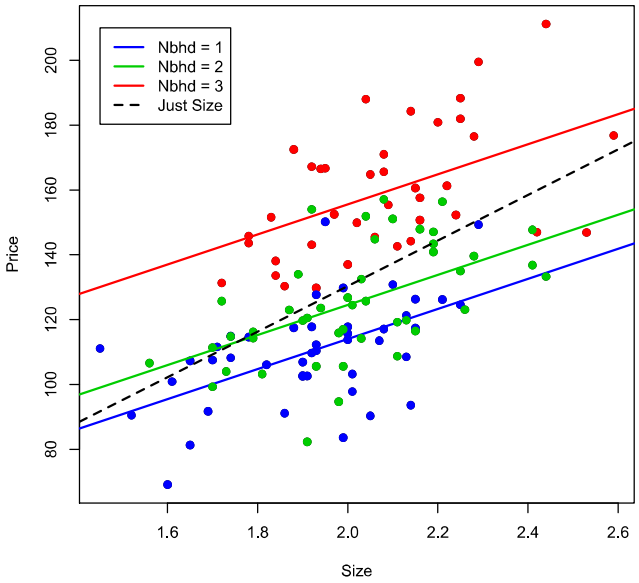
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	28036.4	28036.36	55.501	1E-11
Residual	126	63648.9	505.1496		
Total	127	91685.2			

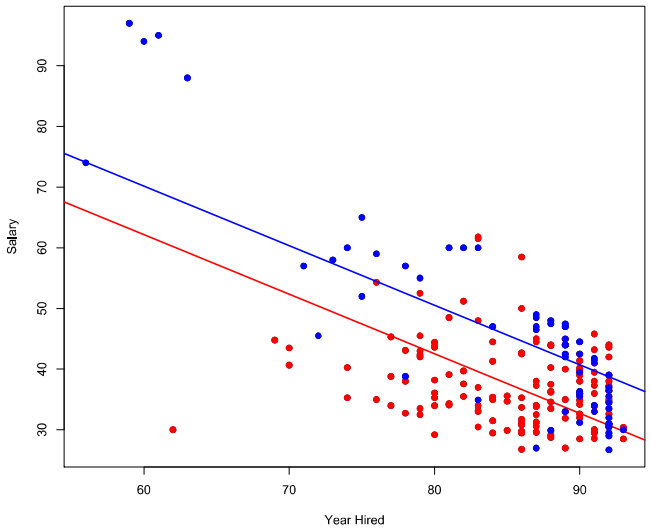
	Coefficients	*tandard Er*	*t Stat*	*P-value*	*ower 95%*	*pper 95%*
Intercept	-10.09	18.97	-0.53	0.60	-47.62	27.44
size	70.23	9.43	7.45	0.00	51.57	88.88

$$Price_i = -10.09 + 70.23 Size + \epsilon_i$$

Dummy Variables... Example: House Prices



Sex Discrimination Case



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

Sex Discrimination Case

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

Yes... Consider the following model:

$$Salary_i = \beta_0 + \beta_1 Exp_i + \beta_2 Sex_i + \beta_3 Exp_i \times Sex_i + \epsilon_i$$

For Females:

$$Salary_i = \beta_0 + \beta_1 Exp_i + \epsilon_i$$

For Males:

$$Salary_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Exp_i + \epsilon_i$$

Sex Discrimination Case

How does the data look like?

	YrHired	Gender	Salary	Sex	SexExp
1	92	Male	32.00	1	92
2	81	Female	39.10	0	0
3	83	Female	33.20	0	0
4	87	Female	30.60	0	0
5	92	Male	29.00	1	92
...	...	...			
208	62	Female	30.00	0	62

Sex Discrimination Case

$$Salary_i = \beta_0 + \beta_1 Sex_i + \beta_2 Exp + \beta_3 Exp * Sex + \epsilon_i$$

Regression Statistics	
Multiple R	0.799130351
R Square	0.638609318
Adjusted R S	0.63329475
Standard Err	6.816298288
Observations	208

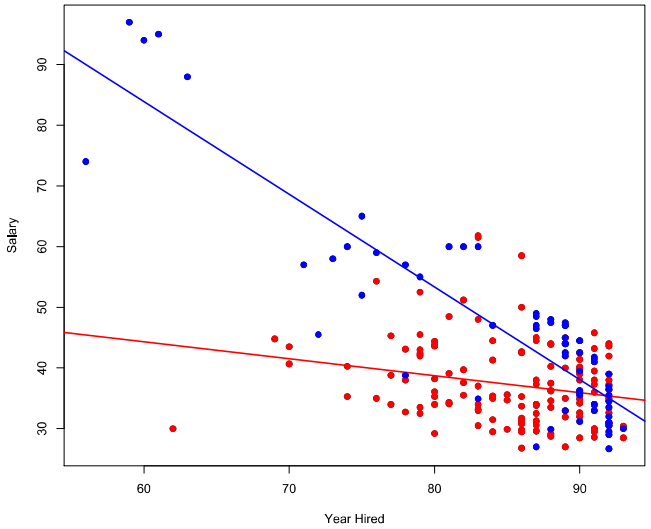
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	16748.88	5582.96	120.16	7.513E-45
Residual	204	9478.232	46.4619		
Total	207	26227.11			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	61.12479795	8.770854	6.96908	4E-11	43.831649	78.41795
Gender	114.4425931	11.7012	9.78041	9E-19	91.371794	137.5134
YrHired	-0.279963351	0.102456	-2.7325	0.0068	-0.4819713	-0.077955
GenderExp	-1.247798369	0.136676	-9.1296	7E-17	-1.5172765	-0.97832

$$Salary_i = 61 + 114 Sex_i + -0.27 Exp + -1.24 Exp * Sex + \epsilon_i$$

Sex Discrimination Case



Variable Interaction

So, 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.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} X_{2i}) + \varepsilon$$

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

This is our first non-linear model!

4. Residual Plots and Transformations

What kind of properties should the residuals have??

$$e_i \approx N(0, \sigma^2) \quad \text{iid and independent from the X's}$$

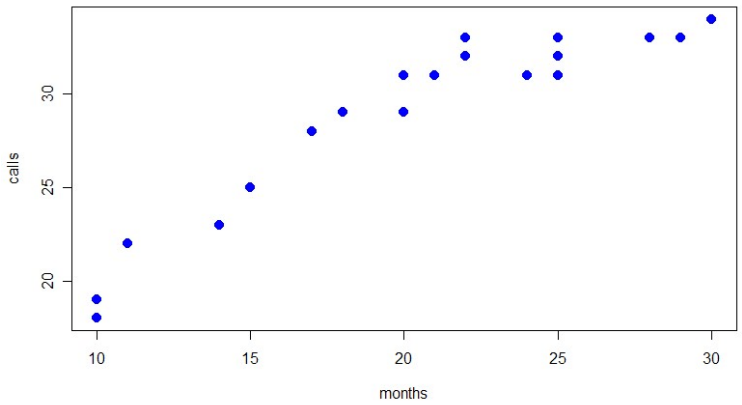
- We should see no pattern between e and each of the X's
- This can be summarized by looking at the plot between $\hat{Y}$ and e
- Remember that $\hat{Y}$ is "pure X", i.e., a linear function of the X's.

If the model is good, the regression should have pulled out of Y all of its "x ness"... what is left over (the residuals) should have nothing to do with X.

Non Linearity

Example: *Telemarketing*

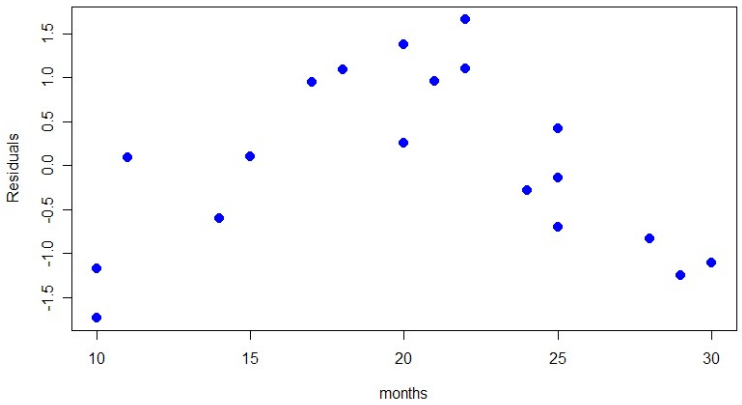
- How does length of employment affect productivity (number of calls per day)?



Non Linearity

Example: *Telemarketing*

- Residual plot highlights the non-linearity!



Non Linearity

What can we do to fix this?? We can use multiple regression and transform our X to create a no linear model...

Let's try

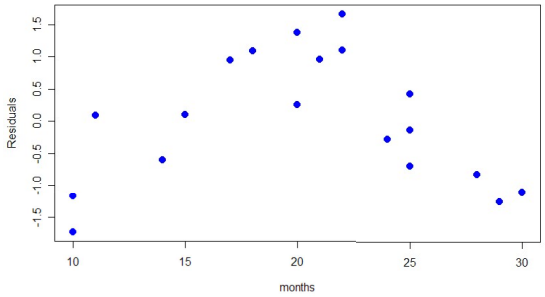
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

The data...

months	months2	calls
10	100	18
10	100	19
11	121	22
14	196	23
15	225	25
...	...	...

Telemarketing

Linear Model



Regression Statistics	
Multiple R	0.934667529
R Square	0.873603389
Adjusted R Square	0.866581356
Standard Error	1.787365193
Observations	20

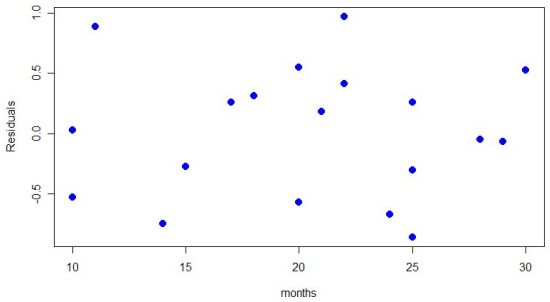
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	397.445862	397.445862	124.408882	1.62235E-09
Residual	18	57.50413798	3.194674332		
Total	19	454.95			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	13.67076987	1.426971138	9.580270766	1.7206E-08	10.67281476	16.66872498
months	0.743514848	0.066659792	11.15387296	1.62235E-09	0.603467823	0.883561873

Telemarketing

With X^2



Regression Statistics	
Multiple R	0.981014716
R Square	0.962389873
Adjusted R Square	0.957965152
Standard Error	1.003251396
Observations	20

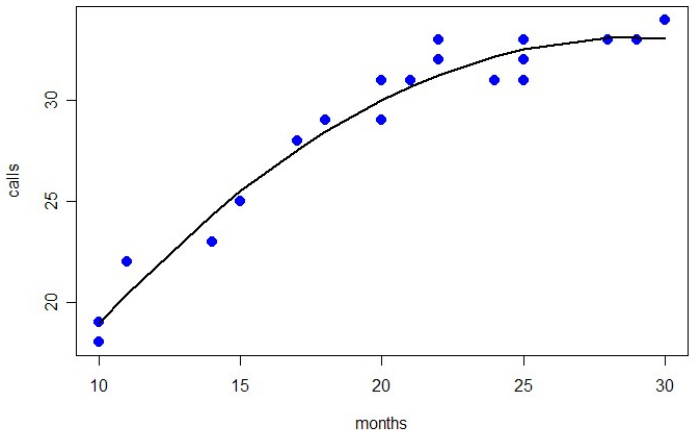
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	437.8392728	218.9196364	217.5029608	7.76409E-13
Residual	17	17.11072717	1.006513363		
Total	19	454.95			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.140471176	2.322630359	-0.060479351	0.95247918	-5.040792846	4.759850493
months	2.310202389	0.250121704	9.236313153	4.89632E-08	1.782491725	2.837913052
Months2	-0.04011825	0.00633281	-6.334983539	7.46662E-06	-0.053479312	-0.026757188

Telemarketing

$$\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2$$



Telemarketing

What is the marginal effect of X on Y?

$$\frac{\partial \mathbb{E}[Y|X]}{\partial X} = \beta_1 + 2\beta_2 X$$

- To better understand the impact of changes in X on Y you should evaluate different scenarios.
- Moving from 10 to 11 months of employment raises productivity by 1.47 calls
- Going from 25 to 26 months only raises the number of calls by 0.27.

Polynomial Regression

Even though we are limited to a linear mean, it is possible to get nonlinear regression by transforming the X variable.

In general, we can add powers of X to get polynomial regression:
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \ldots + \beta_m X^m$

You can fit any mean function if m is big enough.
Usually, $m = 2$ does the trick.

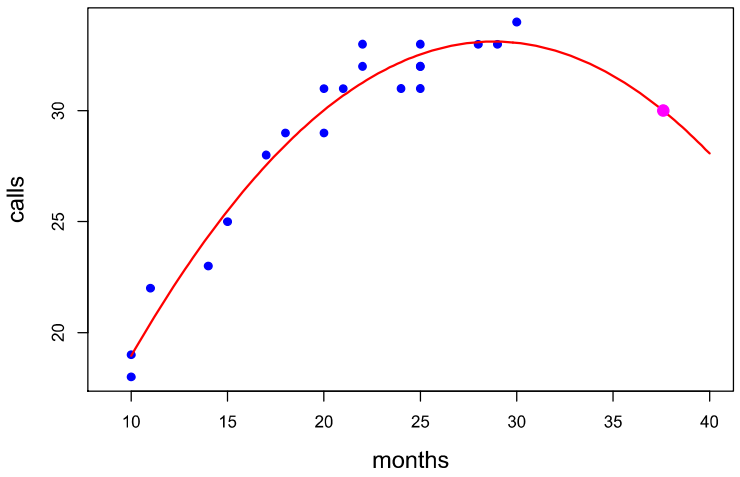
Closing Comments on Polynomials

We can always add higher powers (cubic, etc) if necessary.

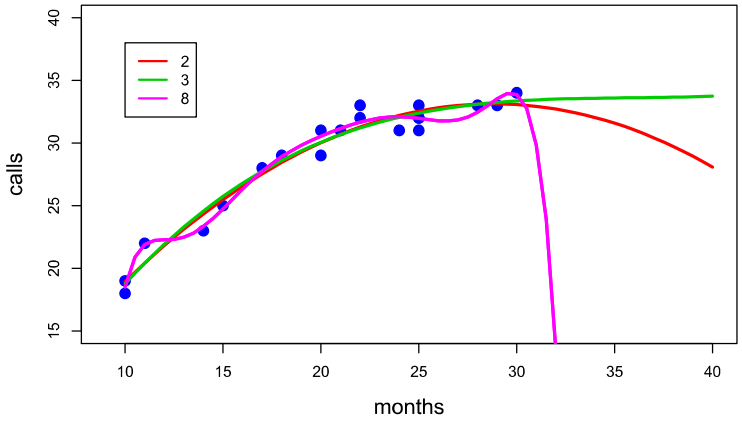
Be very careful about predicting outside the data range. The curve may do unintended things beyond the observed data.

Watch out for over-fitting... remember, simple models are "better".

Be careful when extrapolating...

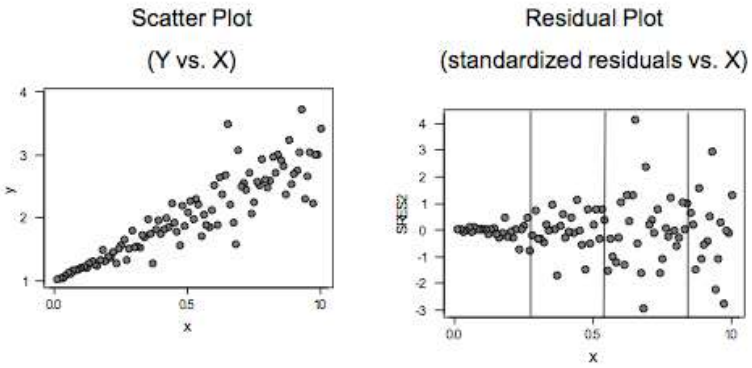


...and, be careful when adding more polynomial terms!



Non-constant Variance

Example...



This violates our assumption that all ε_i have the same σ^2.

Non-constant Variance

Consider the following relationship between Y and X:

$$Y = \gamma_0 X^{\beta_1}(1 + R)$$

where we think about R as a random *percentage error*.

- On average we assume R is 0...
- but when it turns out to be 0.1, Y goes up by 10%!
- Often we see this, the errors are multiplicative and the variation is something like $\pm 10\%$ and not ± 10.
- This leads to non-constant variance (or heteroskedasticity)

The Log-Log Model

We have data on Y and X and we still want to use a linear regression model to understand their relationship... what if we take the log (natural log) of Y?

$$\begin{aligned} \log(Y) &= \log\left[\gamma_0 X^{\beta_1}(1+R)\right] \\ \log(Y) &= \log(\gamma_0) + \beta_1 \log(X) + \log(1+R) \end{aligned}$$

Now, if we call $\beta_0 = \log(\gamma_0)$ and $\epsilon = \log(1+R)$ the above leads to

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon$$

a linear regression of $\log(Y)$ on $\log(X)$!

Elasticity and the log-log Model

In a log-log model, the slope β_1 is sometimes called elasticity.

In english, a 1% increase in X gives a beta % increase in Y.

$$\beta_1 \approx \frac{d\%Y}{d\%X} \quad \text{(Why?)}$$

Price Elasticity

In economics, the slope coefficient β_1 in the regression $\log(sales) = \beta_0 + \beta_1 \log(price) + \varepsilon$ is called price elasticity.

This is the % change in *sales* per 1% change in *price*.

The model implies that $E[sales] = A * price^{\beta_1}$
where $A = exp(\beta_0)$

Price Elasticity of OJ

A chain of gas station convenience stores was interested in the dependency between price of and Sales for orange juice...
They decided to run an experiment and change prices randomly at different locations. With the data in hands, let's first run an regression of Sales on Price:

$$Sales = \beta_0 + \beta_1 Price + \epsilon$$

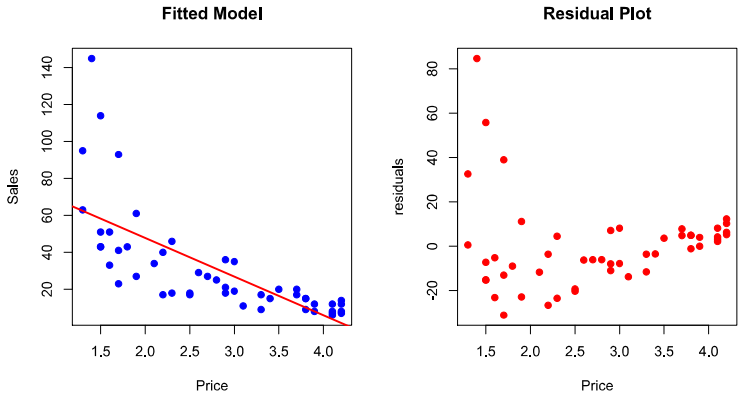
Regression Statistics	
Multiple R	0.719
R Square	0.517
Adjusted R Square	0.507
Standard Error	20.112
Observations	50.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.000	20803.071	20803.071	51.428	0.000
Residual	48.000	19416.449	404.509		
Total	49.000	40219.520			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	89.642	8.610	10.411	0.000	72.330	106.955
Price	-20.935	2.919	-7.171	0.000	-26.804	-15.065

Price Elasticity of OJ



No good!!

Price Elasticity of OJ

But... would you really think this relationship would be linear? Moving a price from \$1 to \$2 is the same as changing it form \$10 to \$11?? We should probably be thinking about the price elasticity of OJ...

$$\log(Sales) = \gamma_0 + \gamma_1 \log(Price) + \epsilon$$

Regression Statistics	
Multiple R	0.869
R Square	0.755
Adjusted R Square	0.750
Standard Error	0.386
Observations	50.000

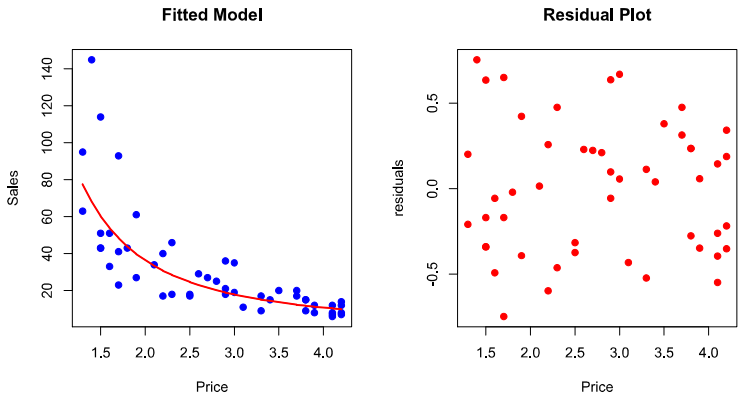
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.000	22.055	22.055	148.187	0.000
Residual	48.000	7.144	0.149		
Total	49.000	29.199			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.812	0.148	32.504	0.000	4.514	5.109
LogPrice	-1.752	0.144	-12.173	0.000	-2.042	-1.463

How do we interpret $\hat{\gamma}_1 = -1.75$?
(When prices go up 1%, sales go down by 1.75%)

Price Elasticity of OJ



Much better!!

Making Predictions

What if the gas station store wants to predict their sales of OJ if they decide to price it at \$1.8?

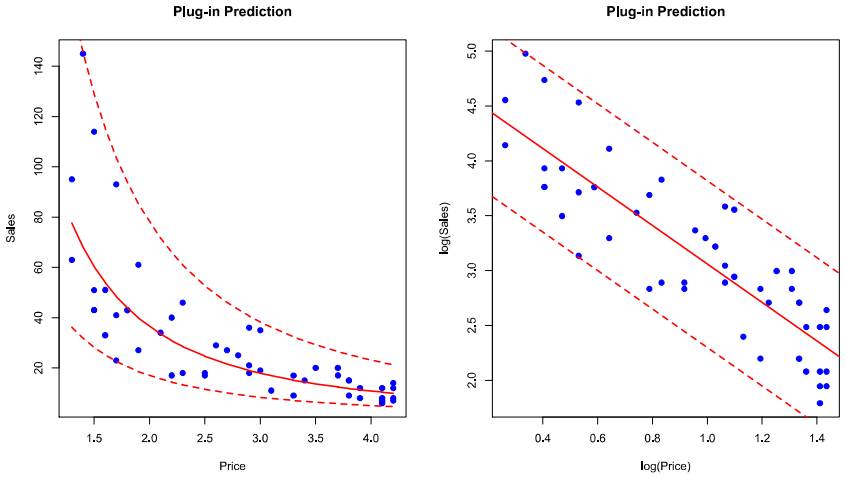
The predicted $\log(Sales) = 4.812 + (-1.752) \times \log(1.8) = 3.78$

So, the predicted $Sales = exp(3.78) = 43.82$.
How about the plug-in prediction interval?

In the log scale, our predicted interval in
$[\widehat{\log(Sales)} - 2s; \widehat{\log(Sales)} + 2s] =$
$[3.78 - 2(0.38); 3.78 + 2(0.38)] = [3.02; 4.54]$.

In terms of actual *Sales* the interval is
$[\exp(3.02), \exp(4.54)] = [20.5; 93.7]$

Making Predictions



- In the log scale (right) we have $[\hat{Y} - 2s; \hat{Y} + 2s]$
- In the original scale (left) we have $[\exp(\hat{Y}) * \exp(-2s); \exp(\hat{Y})\exp(2s)]$

Some additional comments...

- Another useful transformation to deal with non-constant variance is to take only the $\log(Y)$ and keep X the same. Clearly the "elasticity" interpretation no longer holds.
- Always be careful in interpreting the models after a transformation
- Also, be careful in using the transformed model to make predictions

Summary of Transformations

Coming up with a good regression model is usually an iterative procedure. Use plots of residuals *vs* X or $\hat{Y}$ to determine the next step.

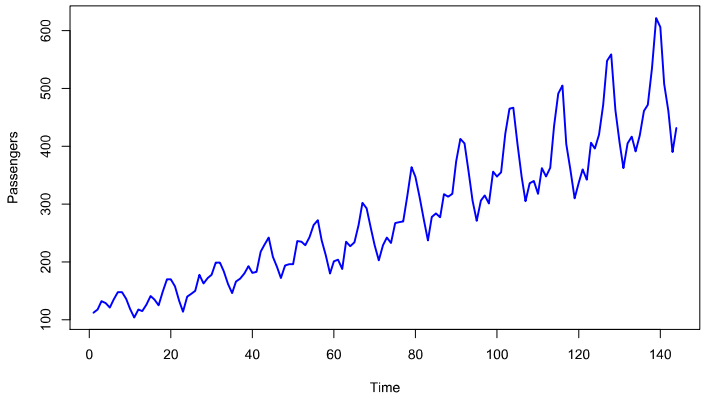
Log transform is your best friend when dealing with non-constant variance ($\log(X)$, $\log(Y)$, or both).

Add polynomial terms (e.g. X^2) to get nonlinear regression.

The bottom line: you should combine what the plots and the regression output are telling you with your common sense and knowledge about the problem. Keep playing around with it until you get something that makes sense and has nothing obviously wrong with it.

Airline Data

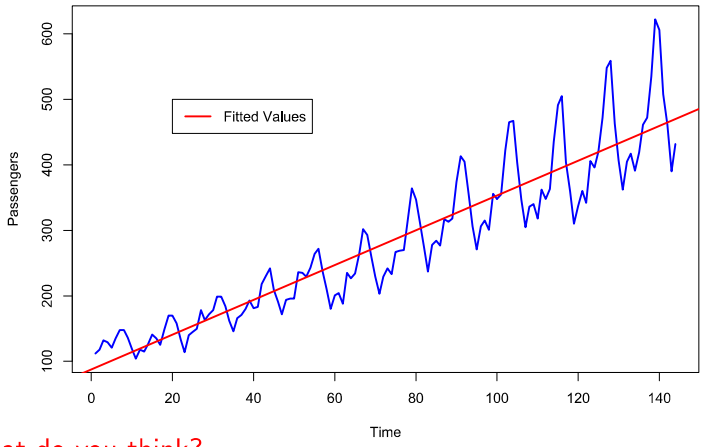
Monthly passengers in the U.S. airline industry (in 1,000 of passengers) from 1949 to 1960... we need to predict the number of passengers in the next couple of months.



Any ideas?

Airline Data

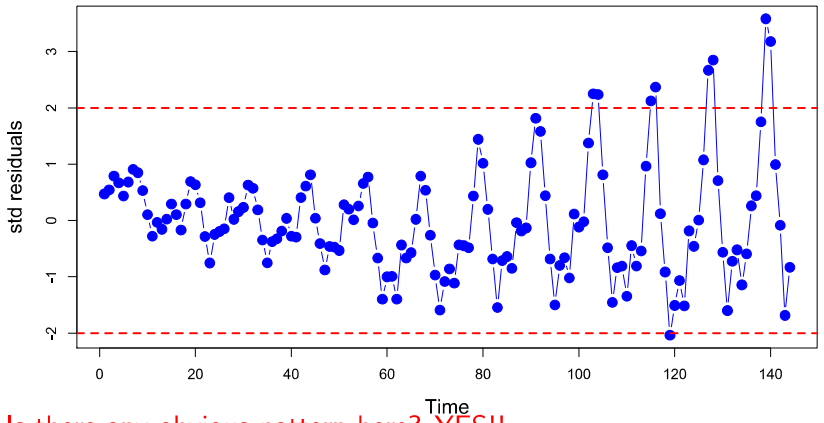
How about a "trend model"? $Y_t = \beta_0 + \beta_1 t + \epsilon_t$



What do you think?

Airline Data

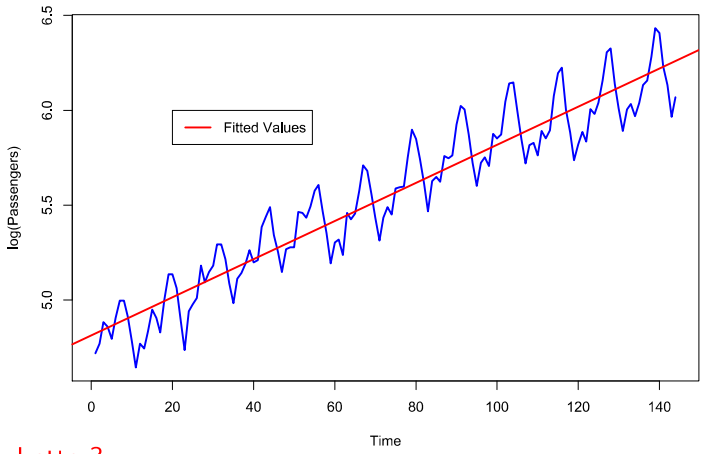
Let's look at the residuals...



Is there any obvious pattern here? YES!!

Airline Data

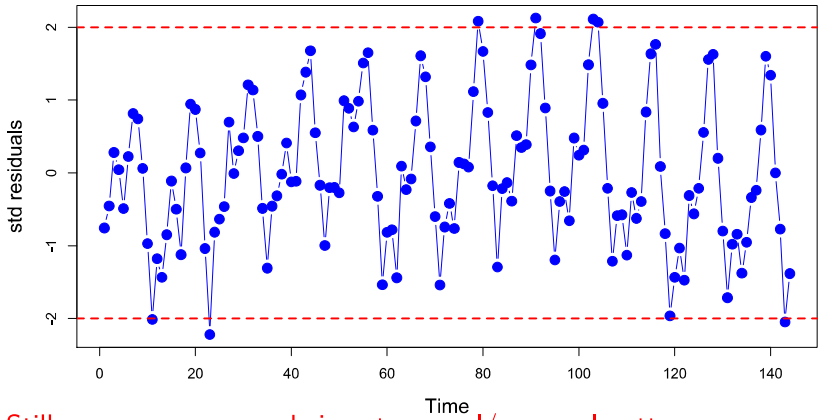
The variance of the residuals seems to be growing in time... Let's try taking the log. $\log(Y_t) = \beta_0 + \beta_1 t + \epsilon_t$



Any better?

Airline Data

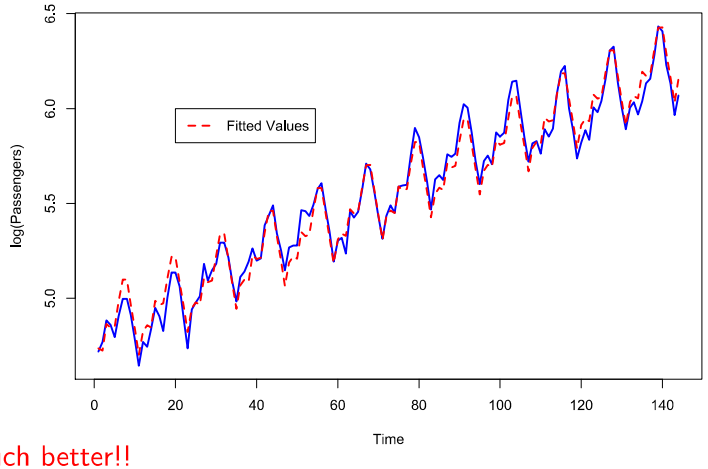
Residuals...



Still we can see some obvious temporal/seasonal pattern....

Airline Data

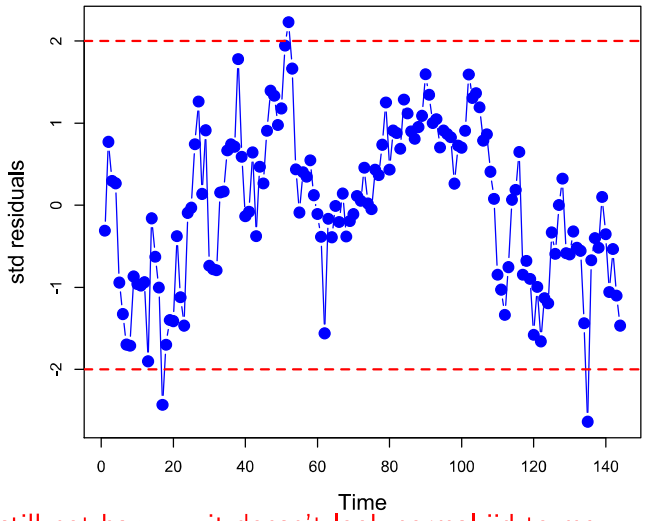
Okay, let's add dummy variables for months (only 11 dummies)...
$\log(Y_t) = \beta_0 + \beta_1 t + \beta_2 Jan + ...\beta_{12} Dec + \epsilon_t$



Much better!!

Airline Data

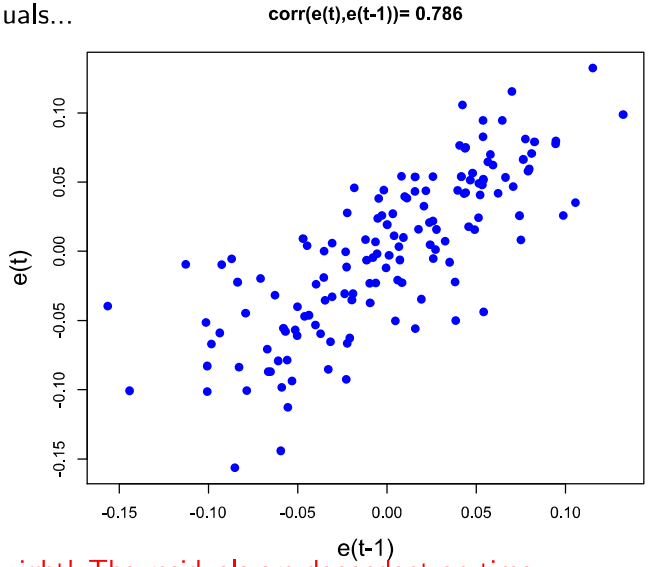
Residuals...



I am still not happy... it doesn't look normal iid to me...

Airline Data

Residuals...

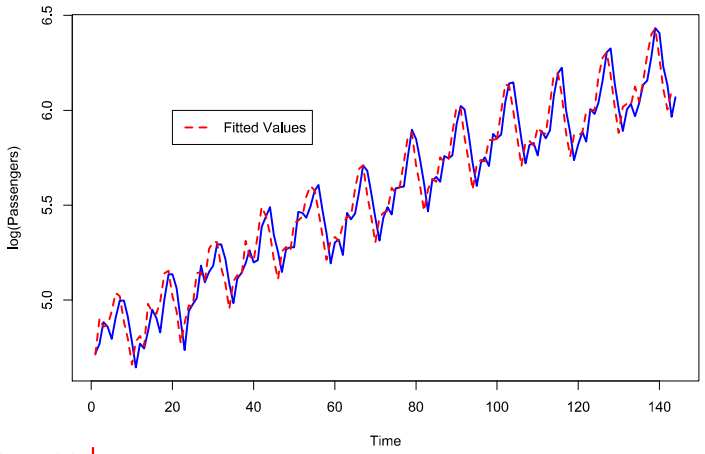


I was right! The residuals are dependent on time...

Airline Data

We have one more tool... let's add one legged term.

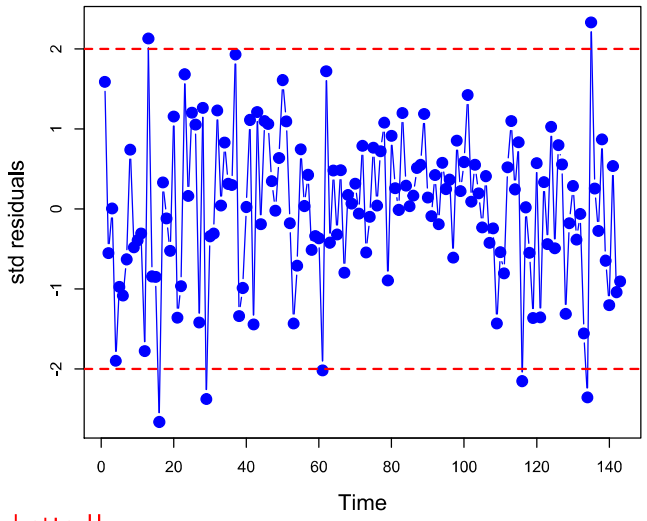
$\log(Y_t) = \beta_0 + \beta_1 t + \beta_2 Jan + ...\beta_{12} Dec + \beta_{13} \log(Y_{t-1}) + \epsilon_t$



Okay, good...

Airline Data

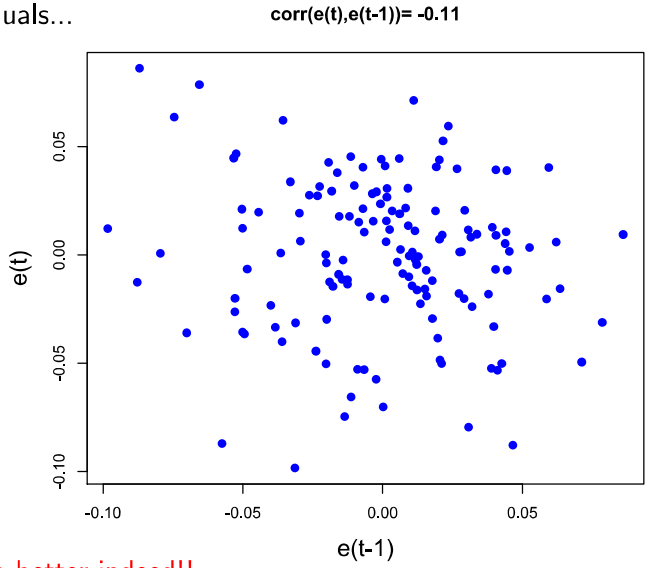
Residuals...



Much better!!

Airline Data

Residuals...



Much better indeed!!

Model Building Process

When building a regression model remember that simplicity is your friend... smaller models are easier to interpret and have fewer unknown parameters to be estimated.

Keep in mind that every additional parameter represents a cost!!

The first step of every model building exercise is the selection of the the universe of variables to be potentially used. This task is entirely solved through you experience and context specific knowledge...

- Think carefully about the problem
- Consult subject matter research and experts
- Avoid the mistake of selecting too many variables

Model Building Process

With a universe of variables in hand, the goal now is to select the model. Why not include all the variables in?

Big models tend to over-fit and find features that are specific to the data in hand... ie, not generalizable relationships.
The results are bad predictions and bad science!

In addition, bigger models have more parameters and potentially more uncertainty about everything we are trying to learn...

We need a strategy to build a model in ways that accounts for the trade-off between fitting the data and the uncertainty associated with the model

5. Variable Selection and Regularization

When working with linear regression models where the number of X variables is large, we need to think about strategies to select what variables to use...

We will focus on 3 ideas:

- Subset Selection
- Shrinkage
- Dimension Reduction

Subset Selection

The idea here is very simple: fit as many models as you can and compare their performance based on some criteria!

Issues:

- How many possible models? Total number of models $= 2^p$
  Is this large?
- What criteria to use?
  Just as before, if prediction is what we have in mind, out-of-sample predictive ability should be the criteria

Information Criteria

Another way to evaluate a model is to use Information Criteria metrics which attempt to quantify how well our model would have predicted the data (regardless of what you've estimated for the β_j's).

A good alternative is the BIC: Bayes Information Criterion, which is based on a "Bayesian" philosophy of statistics.

$$BIC = n \log(s^2) + p \log(n)$$

You want to choose the model that leads to minimum BIC.

Information Criteria

One nice thing about the BIC is that you can interpret it in terms of model probabilities. Given a list of possible models $\{M_1, M_2, \ldots, M_R\}$, the probability that model i is correct is

$$P(M_i) \approx \frac{e^{-\frac{1}{2}BIC(M_i)}}{\sum_{r=1}^{R} e^{-\frac{1}{2}BIC(M_r)}} = \frac{e^{-\frac{1}{2}[BIC(M_i)-BIC_{min}]}}{\sum_{r=1}^{R} e^{-\frac{1}{2}[BIC(M_r)-BIC_{min}]}}$$

(Subtract $BIC_{min} = \min\{BIC(M_1) \ldots BIC(M_R)\}$ for numerical stability.)

Similar, alternative criteria include AIC, C_p, adjusted R^2...
bottom line: these are only useful if we lack the ability to compare models based on their out-of-sample predictive ability!!!

Search Strategies: Stepwise Regression

One computational approach to build a regression model step-by-step is "stepwise regression" There are 3 options:

- Forward: adds one variable at the time until no remaining variable makes a significant contribution (or meet a certain criteria... could be out of sample prediction)
- Backwards: starts will all possible variables and removes one at the time until further deletions would do more harm them good
- Stepwise: just like the forward procedure but allows for deletions at each step

Shrinkage Methods

An alternative way to deal with selection is to work with all p predictors at once while placing a constraint on the size of the estimated coefficients

This idea is a regularization technique that reduces the variability of the estimates and tend to lead to better predictions.

The hope is that by having the constraint in place, the estimation procedure will be able to focus on "the important β's"

Ridge Regression

Ridge Regression is a modification of the least squares criteria that minimizes (as a function of β's)

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for some value of $\lambda > 0$

- The "blue" part of the equation is the traditional objective function of LS
- The "red" part is the shrinkage penalty, ie, something that makes costly to have big values for β

Ridge Regression

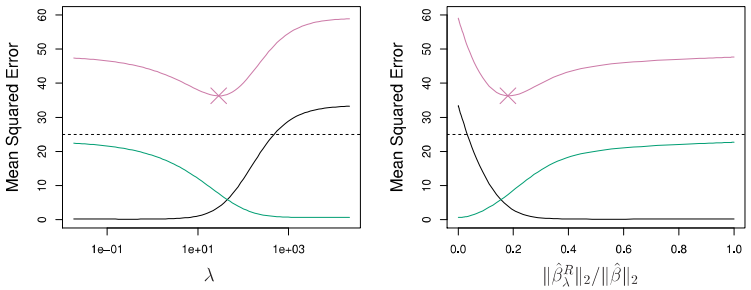
$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

- if $\lambda = 0$ we are back to least squares
- when $\lambda \to \infty$, it is "too expensive" to allow for any β to be different than 0...
- So, for different values of λ we get a different solution to the problem

Ridge Regression

- What ridge regression is doing is exploring the bias-variance trade-off! The larger the λ the more bias (towards zero) is being introduced in the solution, ie, the less flexible the model becomes... at the same time, the solution has less variance
- As always, the trick to find the "right" value of λ that makes the model not too simple but not too complex!
- Whenever possible, we will choose λ by comparing the out-of-sample performance (usually via cross-validation)

Ridge Regression

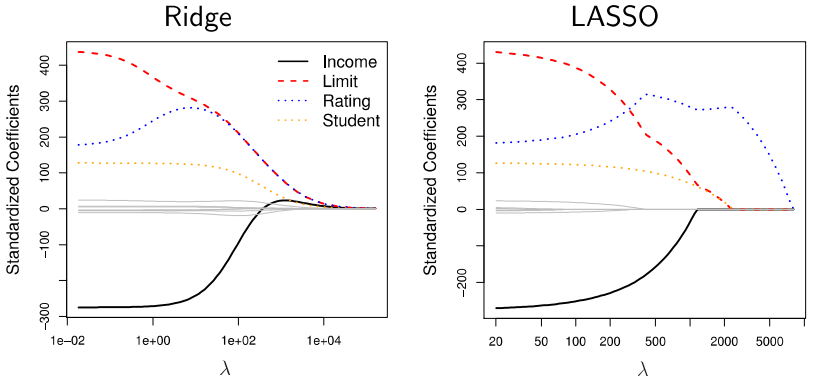


$bias^2$ (black), var(green), test MSE (purple)
Comments:

- Ridge is computationally very attractive as the "computing cost" is almost the same of least squares (contrast that with subset selection!)
- It's a good practice to always center and scale the X's before running ridge

LASSO

The LASSO is a shrinkage method that performs automatic selection. It is similar to ridge but it will provide solutions that are sparse, ie, some β's exactly equal to 0! This facilitates interpretation of the results...



LASSO

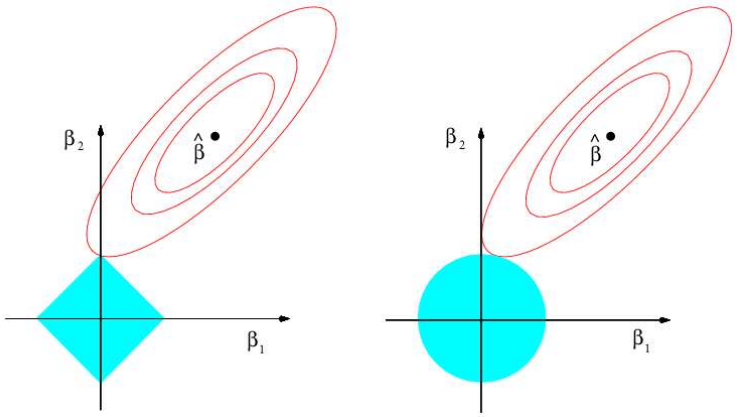
The LASSO solves the following problem:

$$\arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Once again, λ controls how flexible the model gets to be
- Still a very efficient computational strategy
- Whenever possible, we will choose λ by comparing the out-of-sample performance (usually via cross-validation)

Ridge vs. LASSO

Why does the LASSO outputs zeros?



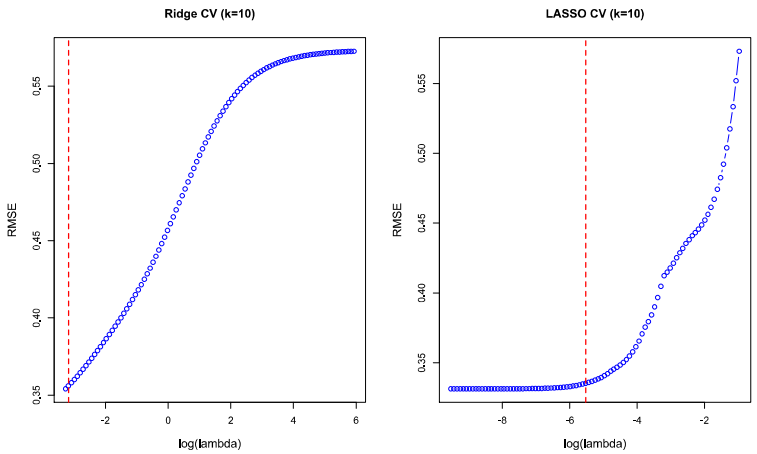
Ridge vs. LASSO

Which one is better?

- It depends...
- In general LASSO will perform better than Ridge when a relative small number of predictors have a strong effect in Y while Ridge will do better when Y is a function of many of the X's and the coefficients are of moderate size
- LASSO can be easier to interpret (the zeros help!)
- But, if prediction is what we care about the only way to decide which method is better is comparing their out-of-sample performance

Choosing λ: California Housing Data

The idea is to solve the ridge or LASSO objective function over a grid of possible values for λ...



6. Dimension Reduction Methods

Sometimes, the number (p) of X variables available is too large for us to work with the methods presented above.

Perhaps, we could first *summarize* the information in the predictors into a smaller set of variables ($m << p$) and then try to predict Y.

In general, these summaries are often linear combinations of the original variables.

Principal Components Regression

A very popular way to summarize multivariate data is Principal Components Analysis (PCA).

PCA is a **dimensionality reduction** technique that tries to represent p variables with a $k < p$ "new" variables.

These "new" variables are create by linear combinations of the original variables and the hope is that a small number of them are able to effectively represent what is going on in the original data.

Principal Components Analysis

Assume we have a dataset where p variables are observed. Let X_i be the i^{th} observation of the p-dimensional vector X. PCA writes:

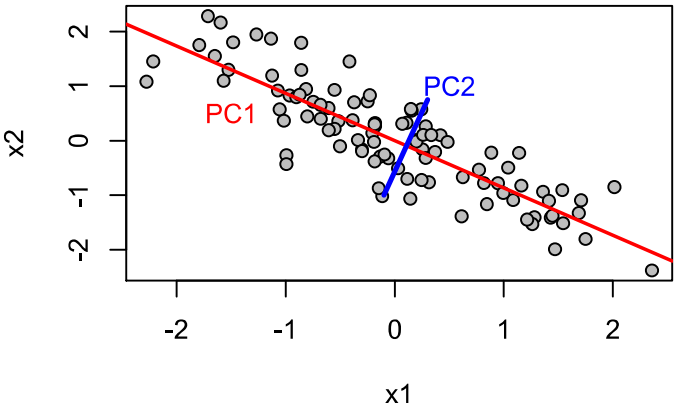
$$x_{ij} = b_{1j}z_{i1} + b_{2j}z_{i2} + \cdots + b_{kj}z_{ik} + e_{ij}$$

where z_{ij} is the i^{th} observation of the j^{th} principal component.

You can think about these z variables as the "essential variables" responsible for all the action in X.

Principal Components Analysis

Here's a picture...

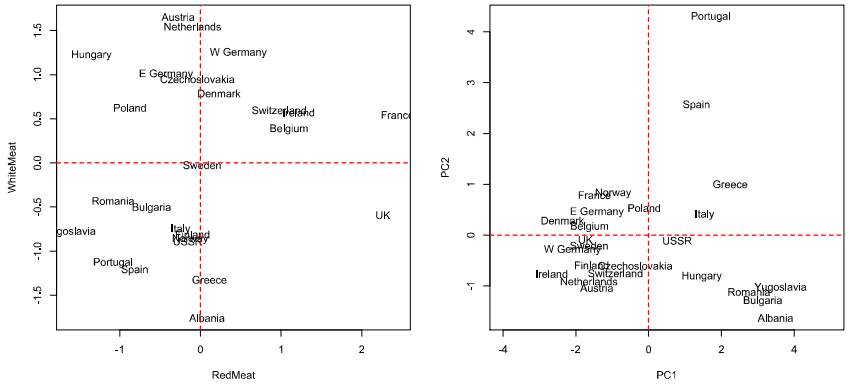


These two variables x_1 and x_2 are very correlated. $PC1$ tells you almost everything that is going on in this dataset!

PCA will look for linear combinations of the original variables that account for most of their variability!

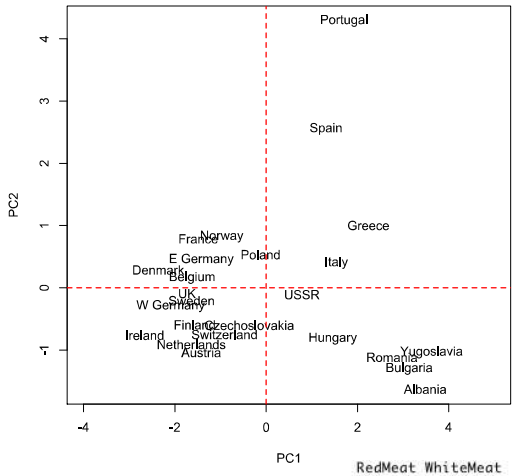
Principal Components Analysis

Let's look at a simple example... Data: Protein consumption by person by country for 7 variables: red meat, white meat, eggs, milk, fish, cereals, starch, nuts, vegetables.



Looks to me that PC1 measures how rich you are and PC2 something to do with the Mediterranean diet!

Principal Components Analysis

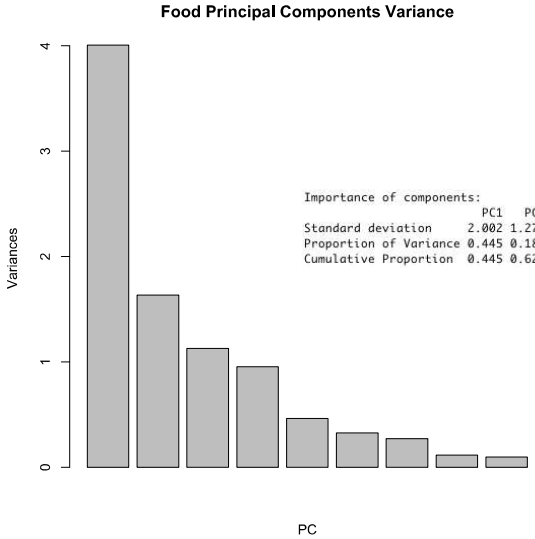


	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr.Veg
PC1	-0.30	-0.31	-0.43	-0.38	-0.14	0.44	-0.30	0.42	0.11
PC2	-0.06	-0.24	-0.04	-0.18	0.65	-0.23	0.35	0.14	0.54
PC3	-0.30	0.62	0.18	-0.39	-0.32	0.10	0.24	-0.05	0.41

These are the weights defining the principal components...

Principal Components Analysis

3 variables might be enough to represent this data... Most of the variability (75%) is explained with PC1, PC2 and PC3.




```
Importance of components:
                          PC1   PC2   PC3   PC4    PC5    PC6    PC7    PC8   PC9
Standard deviation     2.002 1.279 1.062 0.977 0.6811 0.5702 0.5212 0.3410 0.315
Proportion of Variance 0.445 0.182 0.125 0.106 0.0515 0.0361 0.0302 0.0129 0.011
Cumulative Proportion  0.445 0.627 0.752 0.858 0.9098 0.9459 0.9761 0.9890 1.000
```

Principal Components Analysis: Comments

- PCA is a great way to summarize data
- It "clusters" both variables and observations simultaneously!
- The choice of k can be evaluated as a function of the interpretation of the results or via the fit (% of the variation explained)
- The units of each PC is not interpretable in an absolute sense. However, relative to each other it is... see example above.
- Always a good idea to center the data before running PCA.

Principal Components Regression (PCR)

Let's go back to and think of predicting Y with a potentially large number of X variables...

PCA is sometimes used as a way to reduce the dimensionality of X... if only a small number of PC's are enough to represent X, I don't need to use all the X's, right? Remember, smaller models tend to do better in predictive terms!

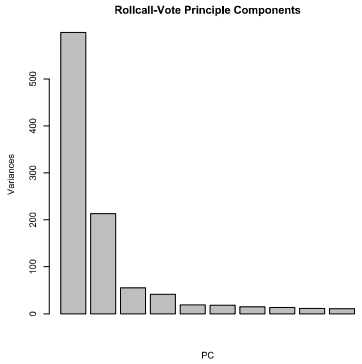
This is called Principal Component Regression. First represent X via k principal components (Z) and then run a regression of Y onto Z. PCR assumes that the *directions in which shows the most variation (the PCs), are the directions associated with Y*.

The choice of k can be done by comparing the out-of-sample predictive performance.

Principal Components Regression (PCR)

Example: Roll Call Votes in Congress... all votes in the 111^{th} Congress (2009-2011); $p = 1647$, $n = 445$.
Goal: Predict party how "liberal" a district is a function of the votes by their representative

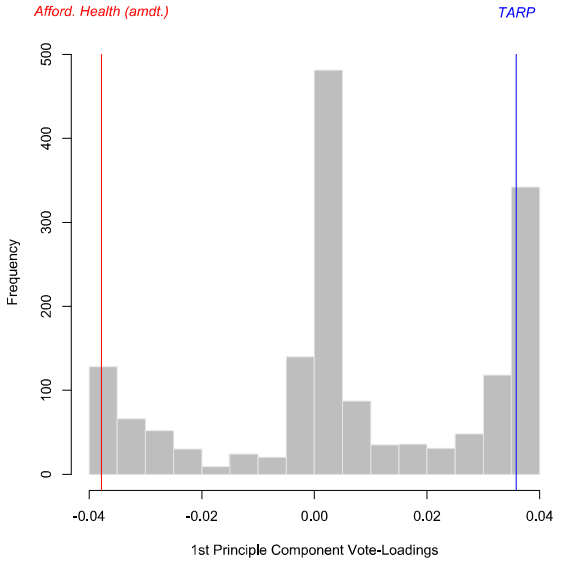
Let's first take the principal component decomposition of the votes...



It looks like 2 PC capture much of what is going on...

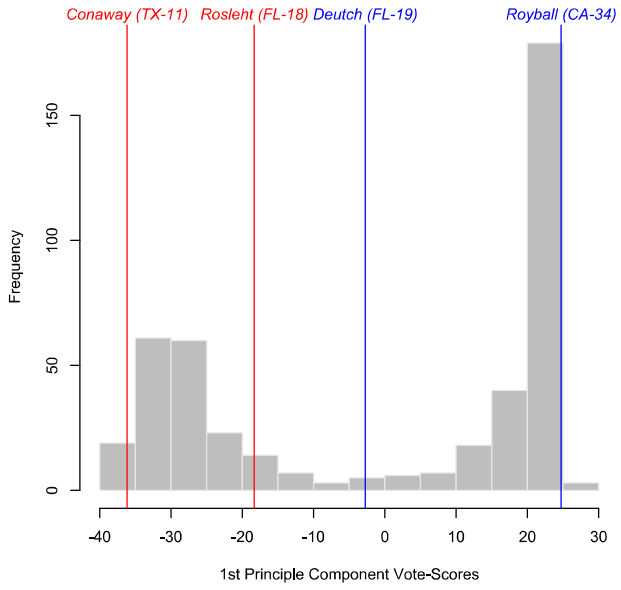
Principal Components Regression (PCR)

Histogram of loadings on PC1... What bills are important in defining PC1?



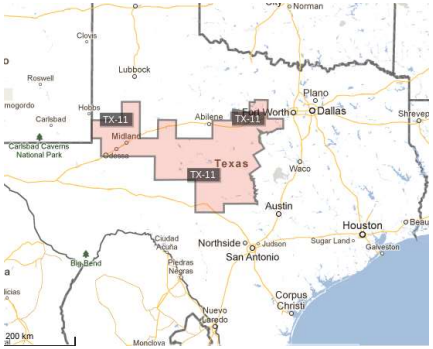
Principal Components Regression (PCR)

Histogram of PC1... "Ideology Score"



Principal Components Regression (PCR)

TX-11

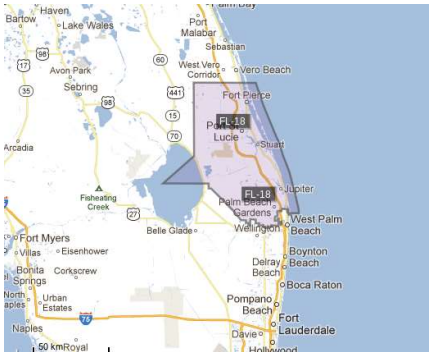


CA-34

The two extremes...

Principal Components Regression (PCR)

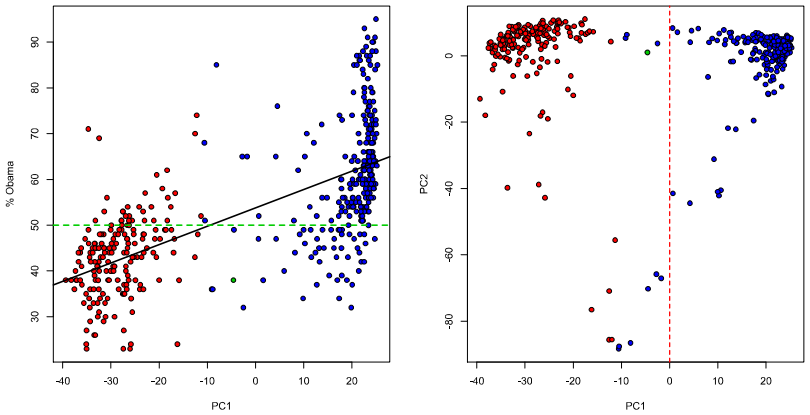
FL-18



FL-19

The swing state!

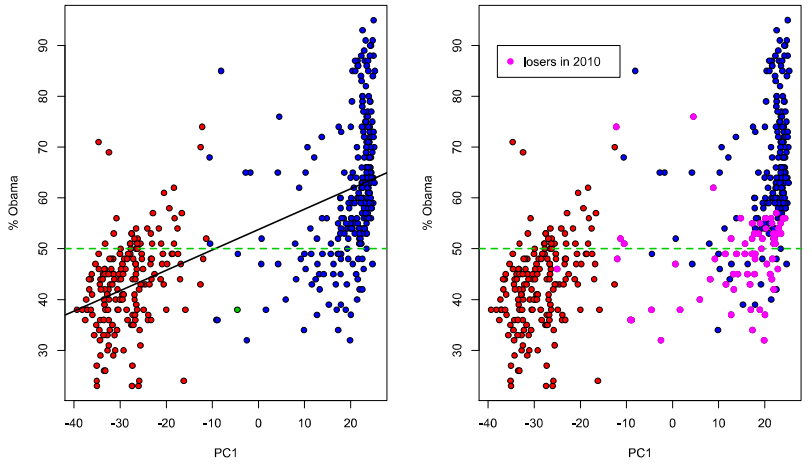
Principal Components Regression (PCR)



All we need is PC1 to predict party affiliation!

How can this picture help you understand what happened in the 2010 election?

Principal Components Regression (PCR)



Partial Least Squares (PLS)

PLS works very similarly to PCR as it will create "new" variables (Z) by taking linear combinations of the original variables (X).

The different is that PLS *attempts to find the directions of variation in X that help explain BOTH X and Y.*

It is a *supervised learning* alternative to PCR...

Partial Least Squares (PLS)

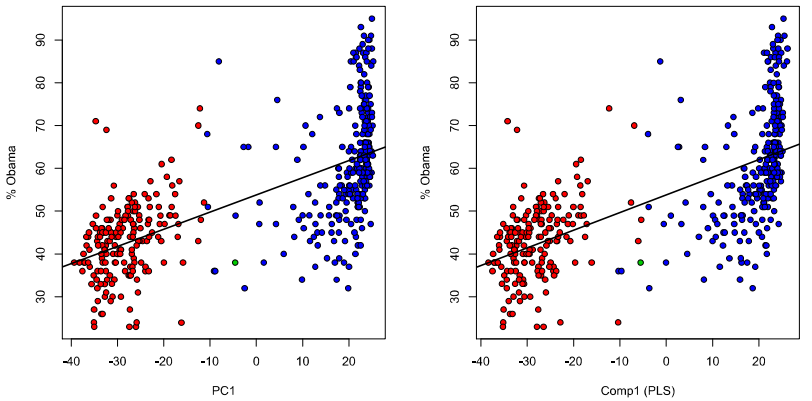
PLS works as follows:

1. The weights of the first linear combination (Z_1) is defined by the regression of Y onto each of the X's... i.e., large weights are going to placed on the X variables most related to Y in a univariate sense
2. Regress each X variable onto Z_1 and compute the residuals
3. Repeat step (1) using the residuals from (2) in place of X
4. iterate

As always, the choice of where to stop, i.e., how many Z variables to use should be done by comparing the out-of-sample predictive performance.

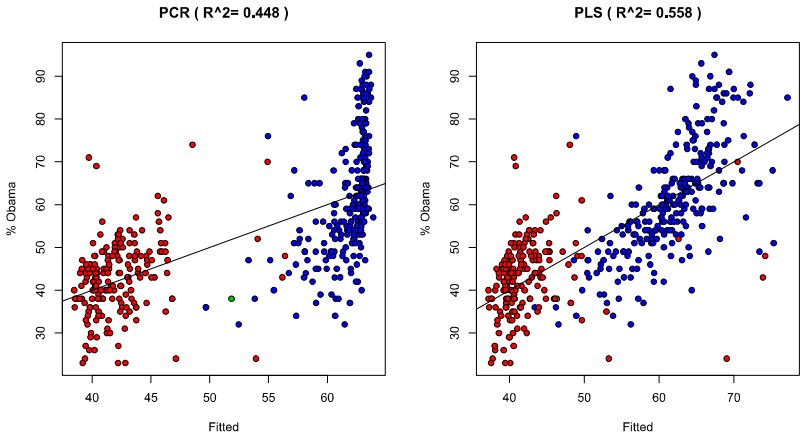
Partial Least Squares (PLS)

Roll Call Data again... it looks like the first component from PLS is the same as the first principal component!



Partial Least Squares (PLS)

But, using two components PLS does better than PCR!!



Partial Least Squares (PLS)

Not easy to understand the difference between the second component in each method (how is that for a homework!)... the bottom line is that by using the information from Y in summarizing the X variables, PLS find a second component that has the ability to explain part of Y.