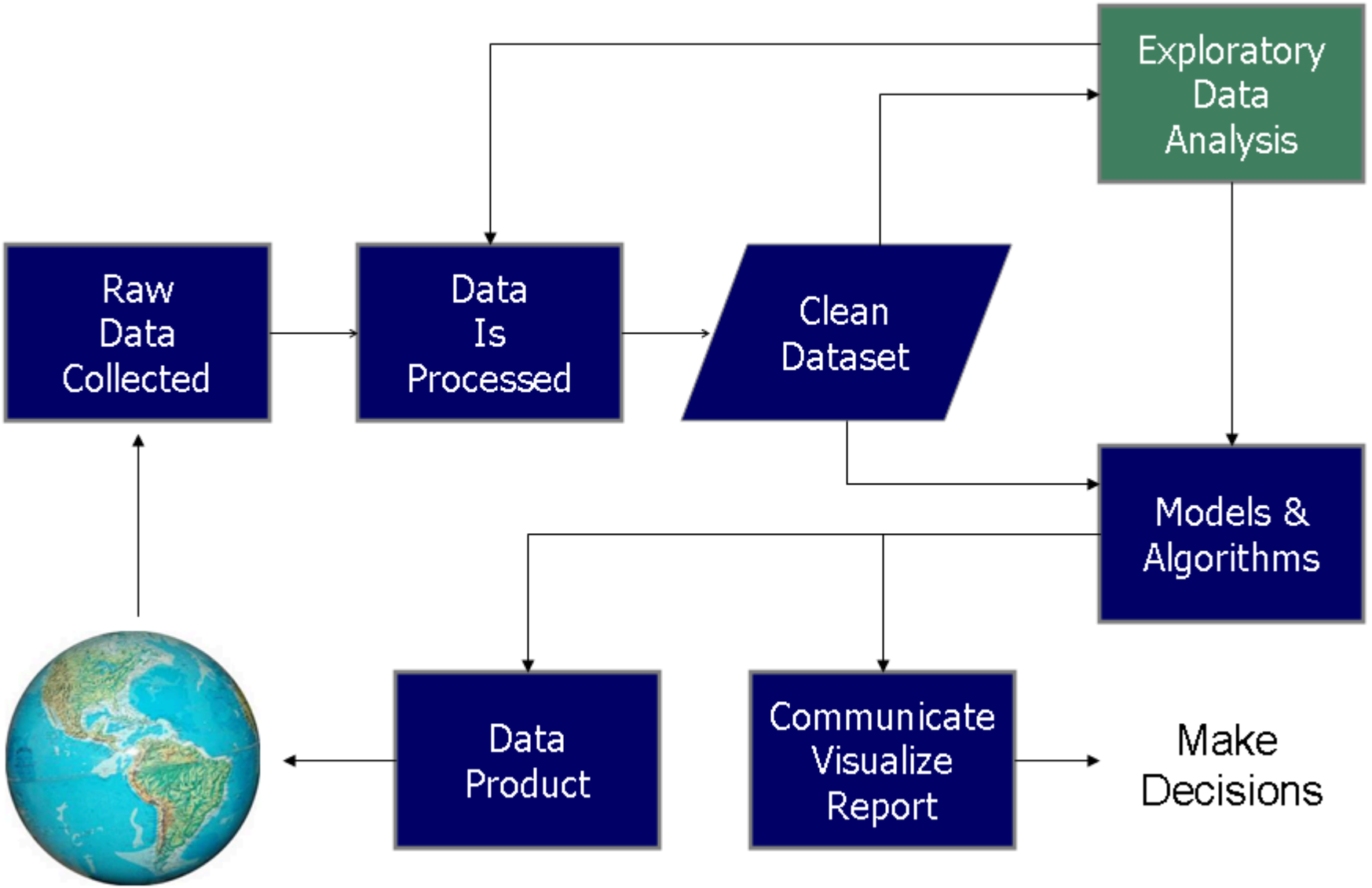
Exploratory Data Analysis
&
Linear Regression

Overview

- Exploratory Data Analysis (EDA)
  - Scatterplots, Histograms, Boxplots
- Simple Linear regression
- Multiple Linear Regression
- Assessing Accuracy and Comparing Models
  - RSS, RSE, R^2, F-Test
- Interpretation
  - Model Output

Data Science Process



Exploratory Data Analysis

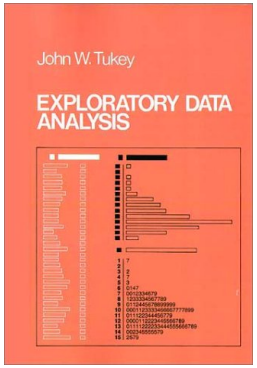
John Tukey, father of modern exploratory data analysis and data visualization

"The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."

Exploratory Data Analysis

A first *feel* for the data

John Tukey's "Exploratory Data Analysis", 1977



Objectives of EDA

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which inference is based
- Support selection of appropriate tools and techniques
- Provide basis for further data collection

Data!

But first…data munging…

Data Munging

- Dirtiness – does the data make sense?
- Missing data (imputation)
- Outliers / Anomalies
- Data type conversion
- Transforming
- Encoding, decoding, recoding
- Renaming variables
- Merging

Exploratory Data Analysis

Types of Variables

- Qualitative (Categorical)
- Quantitative (Numeric)

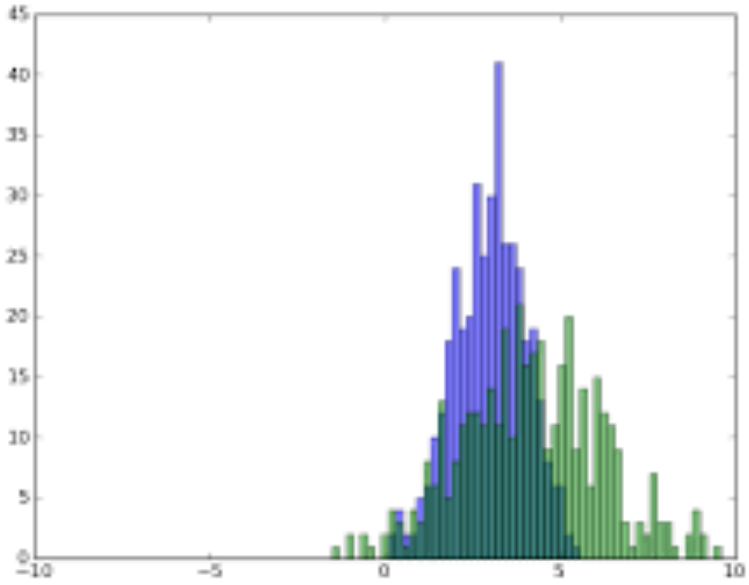
Number of Variables

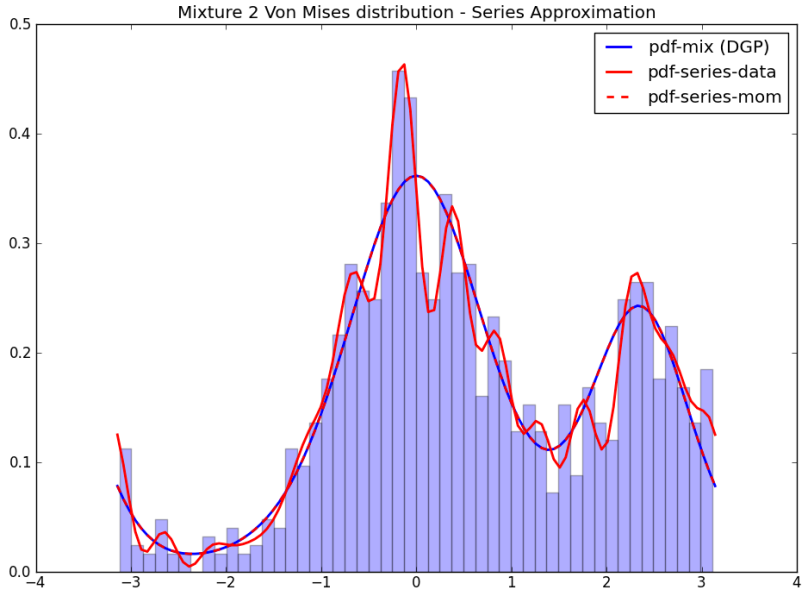
- Univariate (one)
- Bivariate (two)
- Multivariate (many)

Univariate, Numeric

Histogram (or KDE)

- Shows center, variability, skewness
- Outliers
- Be careful of binning

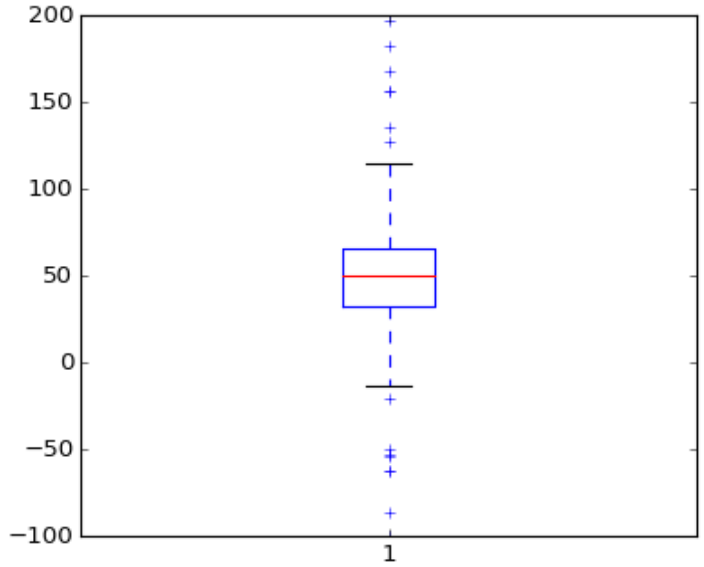




Univariate, Numeric

Boxplots

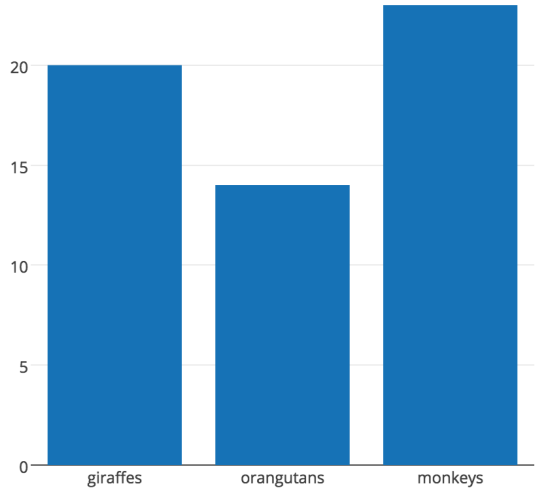
- Median
- IQR
- Range
- Outliers
- (-) Doesn't show distributional shape



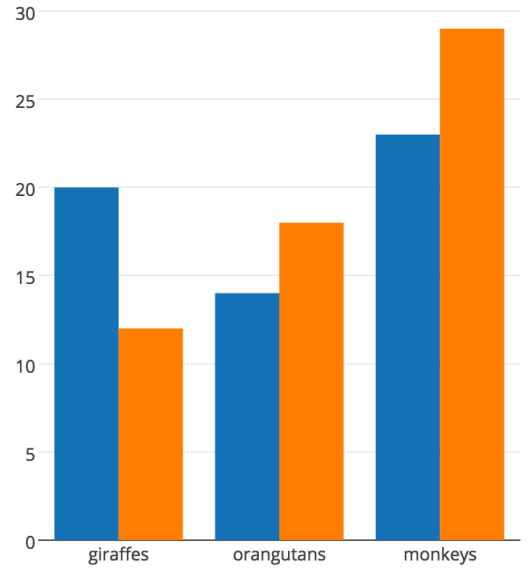
Univariate, Categorical

Barcharts

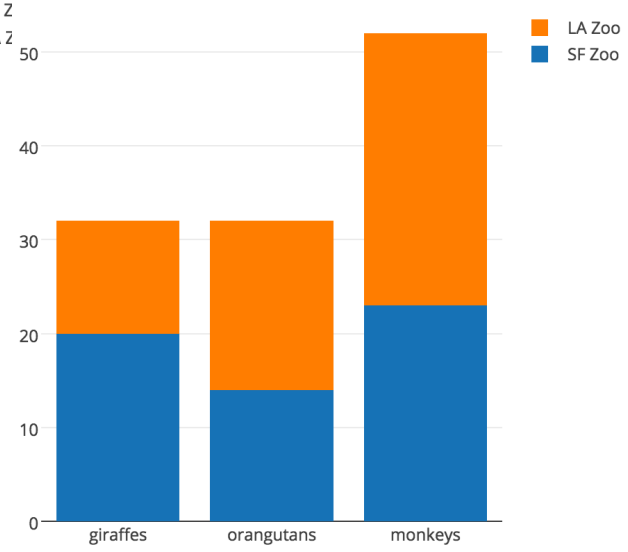
Univariate



Side-by-side

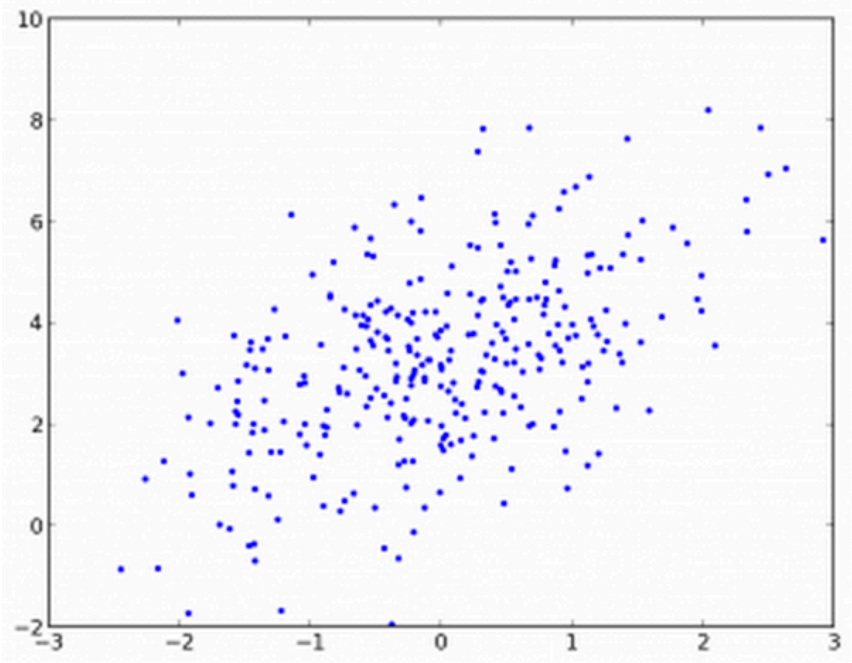


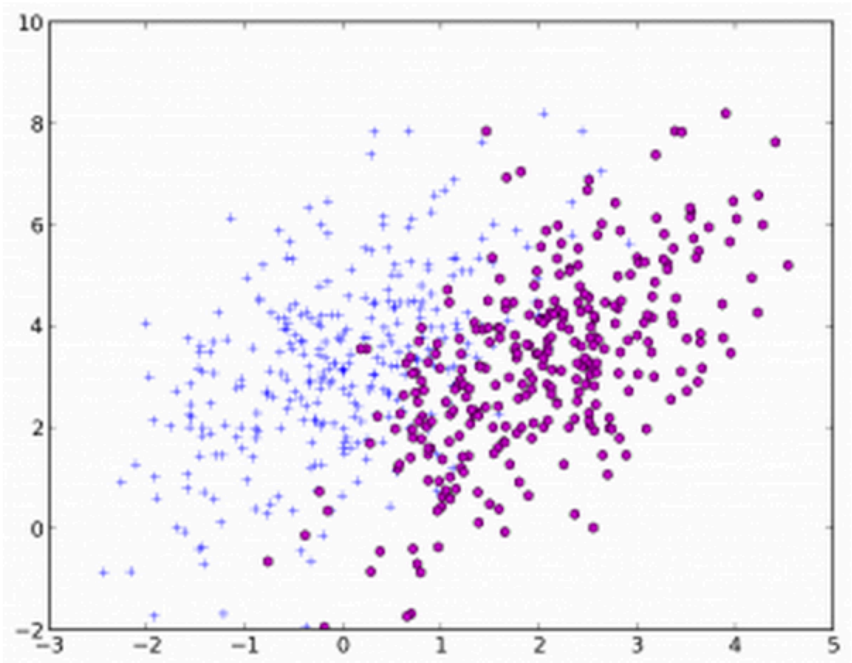
Stacked



Bivariate, Numeric vs. Numeric

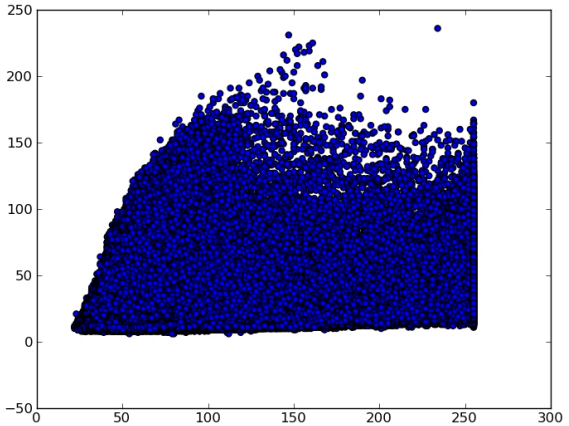
Scatterplots

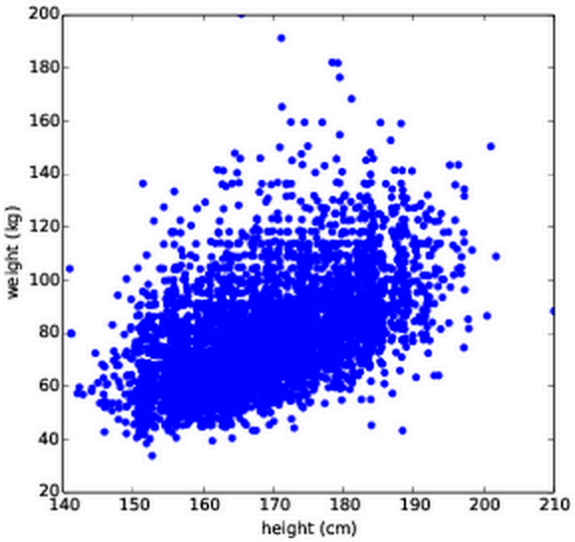




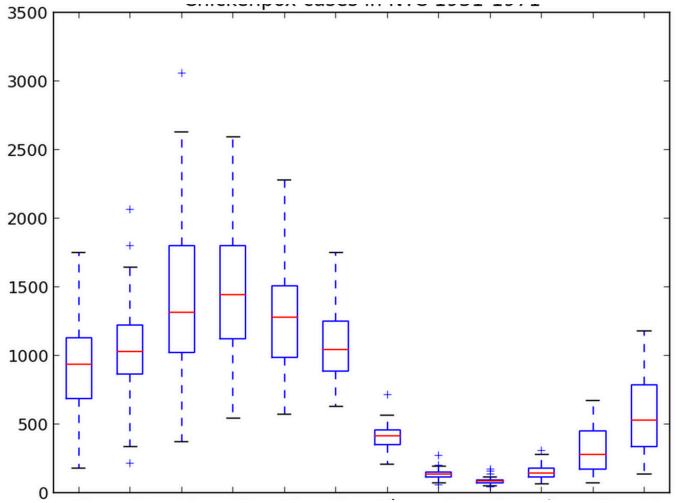
Bivariate, Numeric vs. Numeric

Scatterplots ?!?!

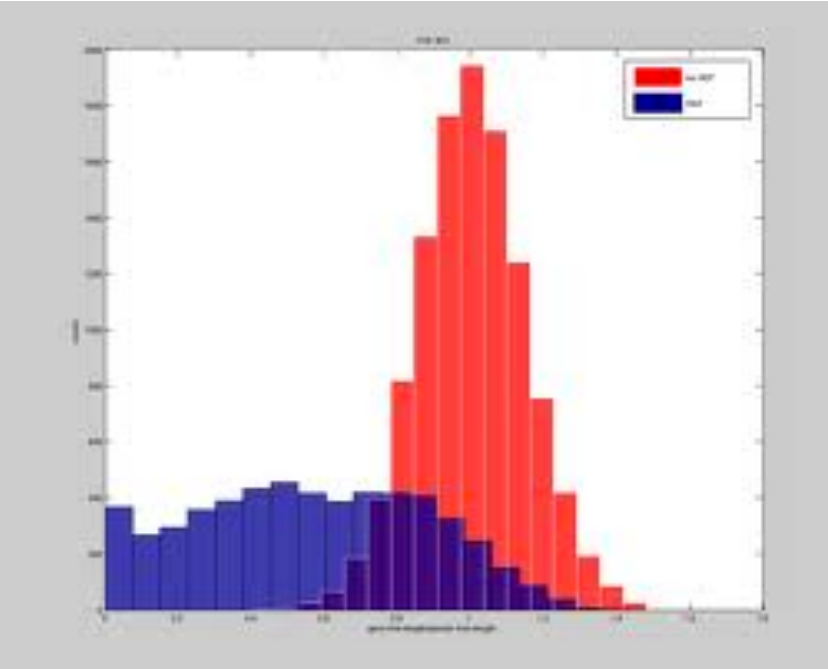


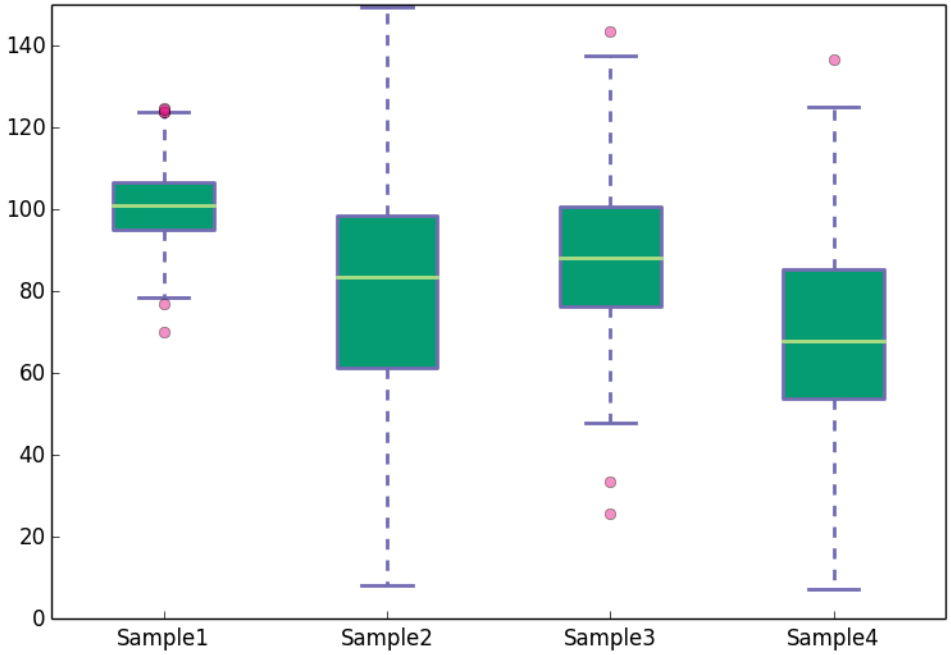


Sometimes useful to bin one of the quantitative variables



Bivariate, Numeric vs. Categorical



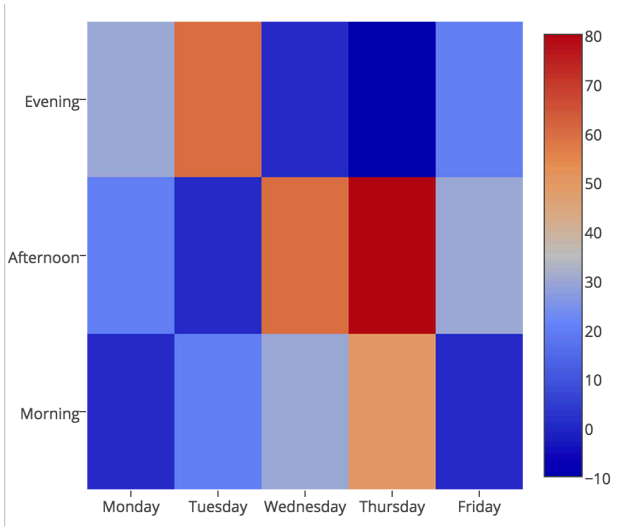


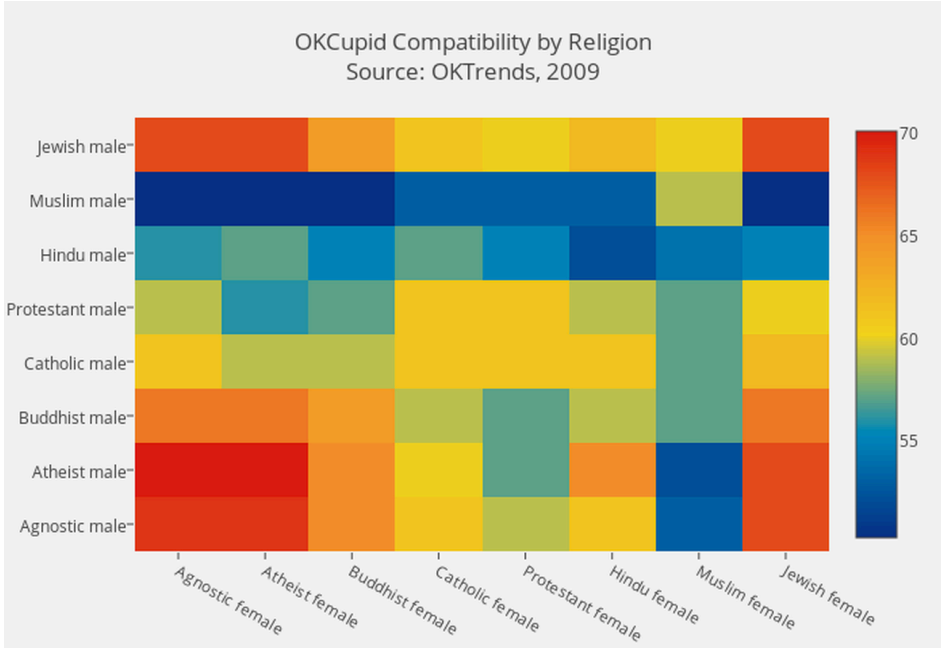
Bivariate, Categorical vs. Categorical

```
pd.crosstab(cdystonia.sex, cdystonia.site)
```

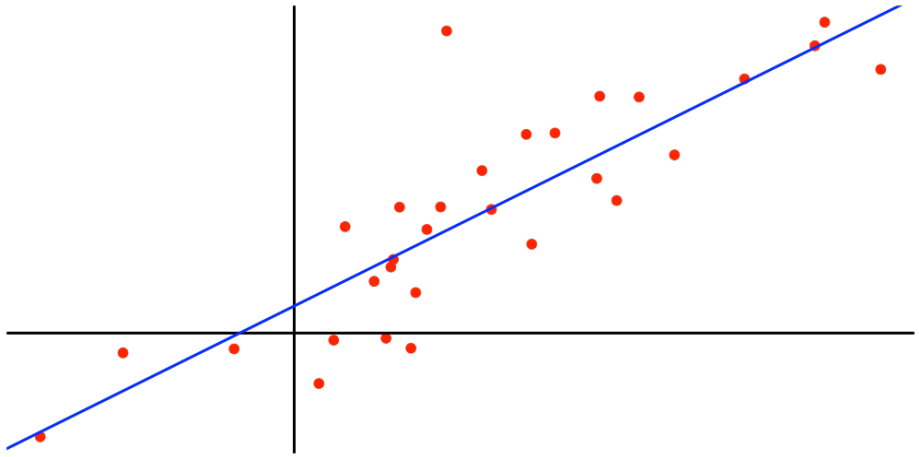
site	1	2	3	4	5	6	7	8	9
sex									
F	52	53	42	30	22	54	66	48	28
M	18	29	30	18	11	33	6	58	33

Simple cross-tabs can be handy!





Simple Linear Regression



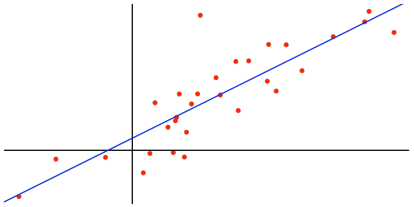
X	Y
Stock Quote	Future Stock Price
% of Diabetes	Mortality Rate
Historic Web Logs	Page Views
Airplane Flight Status	Arrival Time
Anything!	Anything!

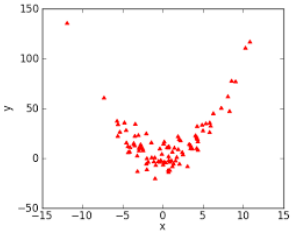
Though we use **X** to predict **Y**

(It'd be rather awkward to use Mortality Rate to predict % of Diabetes)

Simple Linear Regression

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





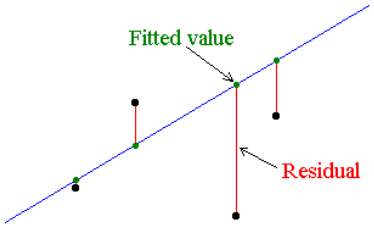
- The Model, what you're presuming the world looks like
- β_0 and β_1 are unknown constants that represent the intercept and slope.
- ε is the error term. ε ~ i.i.d. N(0, σ^2)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are model coefficient estimates for world presumed
- $\hat{y}$ indicates the prediction of Y based on X=x

Simple Linear Regression

$$e_i = y_i - \hat{y}_i$$



Want these to be small

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Typically square them!
(though absolute value is an alternative)

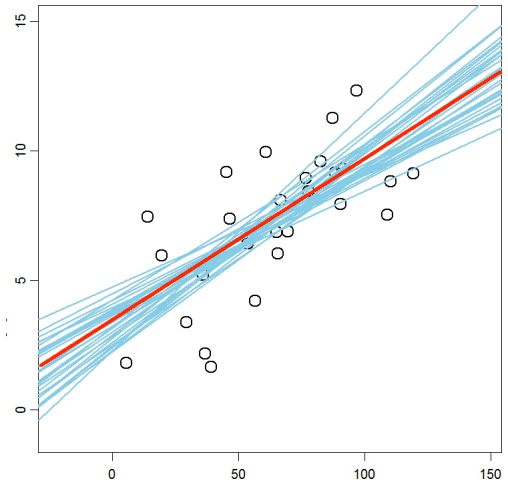
$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

These are the estimates that minimize RSS

Simple Linear Regression



$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

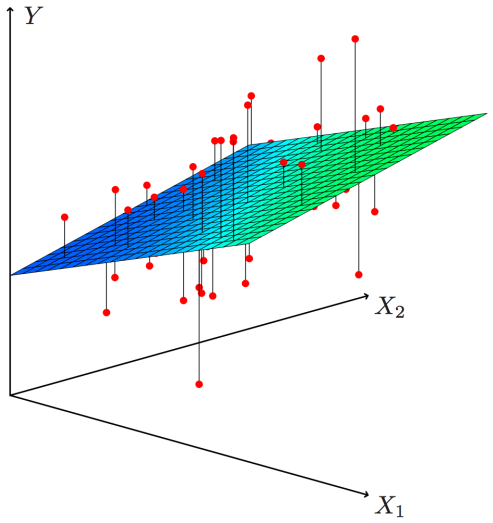
$$\sigma^2 = \mathrm{Var}(\epsilon)$$

	Recall	Here
Setup Hypothesis	$H_0 : \mu = \mu_0 = 100$	$H_0 : \beta_1 = 0$ ← Test if X has effect on Y
Sample Statistic	$\bar{x}$	$\hat{\beta}_1$
Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$
Confidence Interval	$(\bar{x} - t_{\alpha/2} * \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2} * \frac{s}{\sqrt{n}})$	$[\hat{\beta}_1 - t_{\alpha/2} * SE(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2} * SE(\hat{\beta}_1)]$

Multiple Linear Regression

Model

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



Fitted Value

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Residual Sum of Squares

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

Coefficient Estimates

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Multiple Linear Regression

Model in Matrix Form

$$\begin{aligned} \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \\ \epsilon &\sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n\times n}) \\ \mathbf{Y} &\sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}) \end{aligned}$$

Design Matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}$$

Coefficient matrix β:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Assessing Accuracy

So your **RSS** is **1,520,123.11**.

This is a really meaningless number…

- Grows with n
- Measured in the units of your response, y
  - Think y in dollars vs. y in millions of dollars

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Assessing Accuracy

Residual Sum of Squares

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Not great…

Residual Standard Error

$$RSE = \sqrt{\frac{1}{n-p-1}RSS} = \sqrt{\frac{(y_i - \hat{y}_i)^2}{n-p-1}}$$

Better…can roughly think of as average amount that response will deviate from regression line

R-Squared, or "Proportion of Variance Explained"

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

☺ Nice interpretation
Independent of scale of y

Comparing Models

- F-test can be used to compare any one model, *m_reduced*, nested within another model, *m_full*

- Ex. Suppose trying to predict Y, MPG. Suspect height and color might not really be important variables.

  m_reduced: $Y \sim \beta_0 + \beta_{weight} + \beta_{modelyear} + \beta_{cartype}$

  m_full: $Y \sim \beta_0 + \beta_{weight} + \beta_{modelyear} + \beta_{cartype} + \underline{\beta_{height}} + \underline{\beta_{color}}$

$$H_0 : \beta_{height} = \beta_{color} = 0$$
$$H_A : at\ least\ one\ of\ \beta_{height}\ or\ \beta_{color}\ is\ non-zero$$

Comparing Models

(1) Set up comparison

m_reduced: $Y \sim \beta_0 + \beta_{weight} + \beta_{modelyear} + \beta_{cartype}$

m_full: $Y \sim \beta_0 + \beta_{weight} + \beta_{modelyear} + \beta_{cartype} + \underline{\beta_{height}} + \underline{\beta_{color}}$

(2) Compute F-statistic

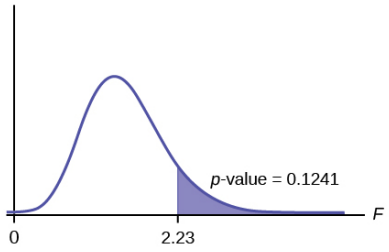
$$F = \frac{(RSS_{reduced} - RSS_{full})/(p_{full} - p_{reduced})}{RSS_{full}/(n - p_{full} - 1)}$$

where F has degrees of freedom (p_full - p_reduced), (n - p_full – 1)

Notice that if *height* and *color* really don't matter much…

$(RSS_{reduced} - RSS_{full})$ will be small → F-statistic will be small

(3) Compute p-value



Assuming α=0.05,

- if p < 0.05 reject null (that height and color don't matter)
- If p >= 0.05, fail to reject null (that height and color don't matter)

Comparing Models

- F-test can be used super generally
- Two special use cases

① Is my model useful at all? i.e. Is at least one of my predictors X_1, X_2, … X_p useful in predicting the response?

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_A : at\ least\ one\ of\ \beta_j\ is\ non-zero$$
$$\longrightarrow F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

② Equivalence to t-test in the Regression Output!

m_reduced: $Y \sim \beta_0 + \beta_{weight} + \beta_{height} + \beta_{modelyear}$

m_full: $Y \sim \beta_0 + \beta_{weight} + \beta_{height} + \beta_{modelyear} + \underline{\beta_{cartype}}$

Interpretation

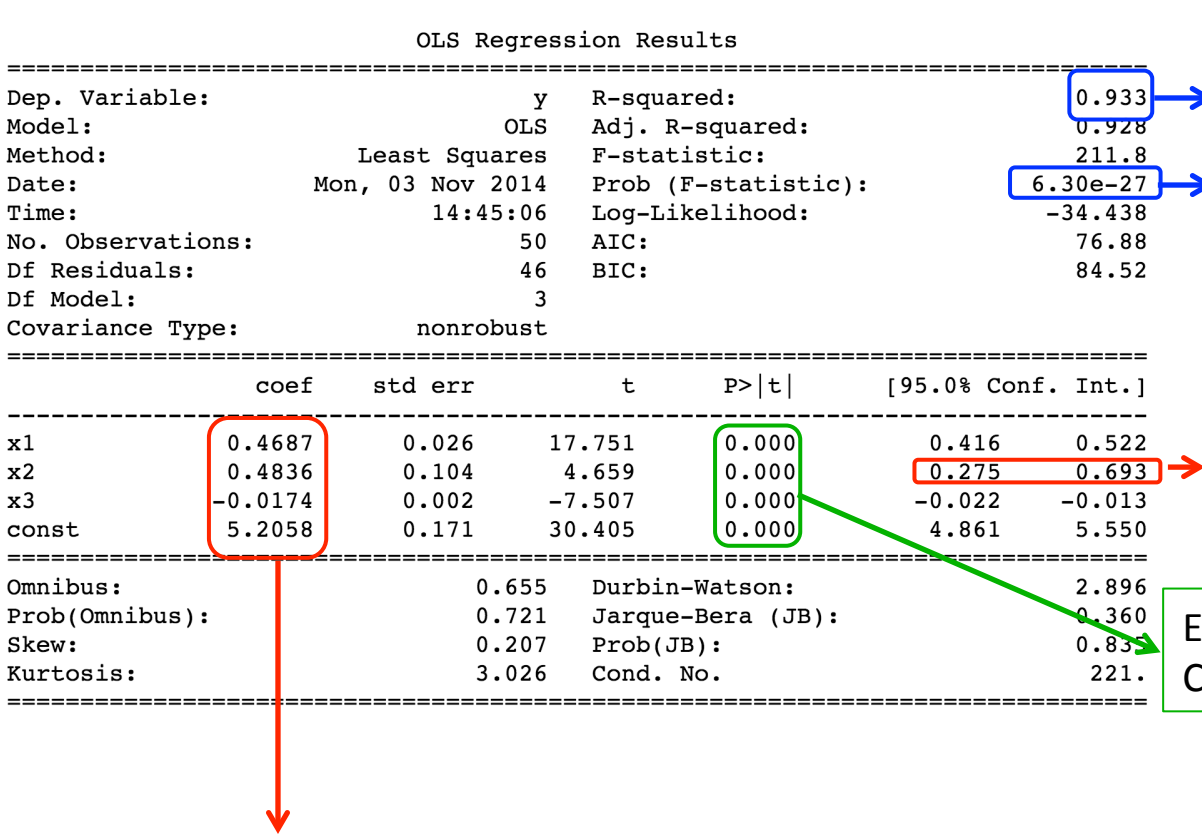
OLS Regression Results

Dep. Variable:	y	R-squared:	0.933
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	211.8
Date:	Mon, 03 Nov 2014	Prob (F-statistic):	6.30e-27
Time:	14:45:06	Log-Likelihood:	-34.438
No. Observations:	50	AIC:	76.88
Df Residuals:	46	BIC:	84.52
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	0.4687	0.026	17.751	0.000	0.416	0.522
x2	0.4836	0.104	4.659	0.000	0.275	0.693
x3	-0.0174	0.002	-7.507	0.000	-0.022	-0.013
const	5.2058	0.171	30.405	0.000	4.861	5.550

Omnibus:	0.655	Durbin-Watson:	2.896
Prob(Omnibus):	0.721	Jarque-Bera (JB):	0.360
Skew:	0.207	Prob(JB):	0.835
Kurtosis:	3.026	Cond. No.	221.

Interpretation



OLS Regression Results

Dep. Variable:	y	R-squared:	0.933
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	211.8
Date:	Mon, 03 Nov 2014	Prob (F-statistic):	6.30e-27
Time:	14:45:06	Log-Likelihood:	-34.438
No. Observations:	50	AIC:	76.88
Df Residuals:	46	BIC:	84.52
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	0.4687	0.026	17.751	0.000	0.416	0.522
x2	0.4836	0.104	4.659	0.000	0.275	0.693
x3	-0.0174	0.002	-7.507	0.000	-0.022	-0.013
const	5.2058	0.171	30.405	0.000	4.861	5.550

Omnibus:	0.655	Durbin-Watson:	2.896
Prob(Omnibus):	0.721	Jarque-Bera (JB):	0.360
Skew:	0.207	Prob(JB):	0.835
Kurtosis:	3.026	Cond. No.	221.

Proportion of Variance Explained by model is 93.3%

Measure of the significance of the fit …my model isn't utterly useless ☺

There is an approximately 95% chance that [0.275, 0.693] will contain the true value of β_2

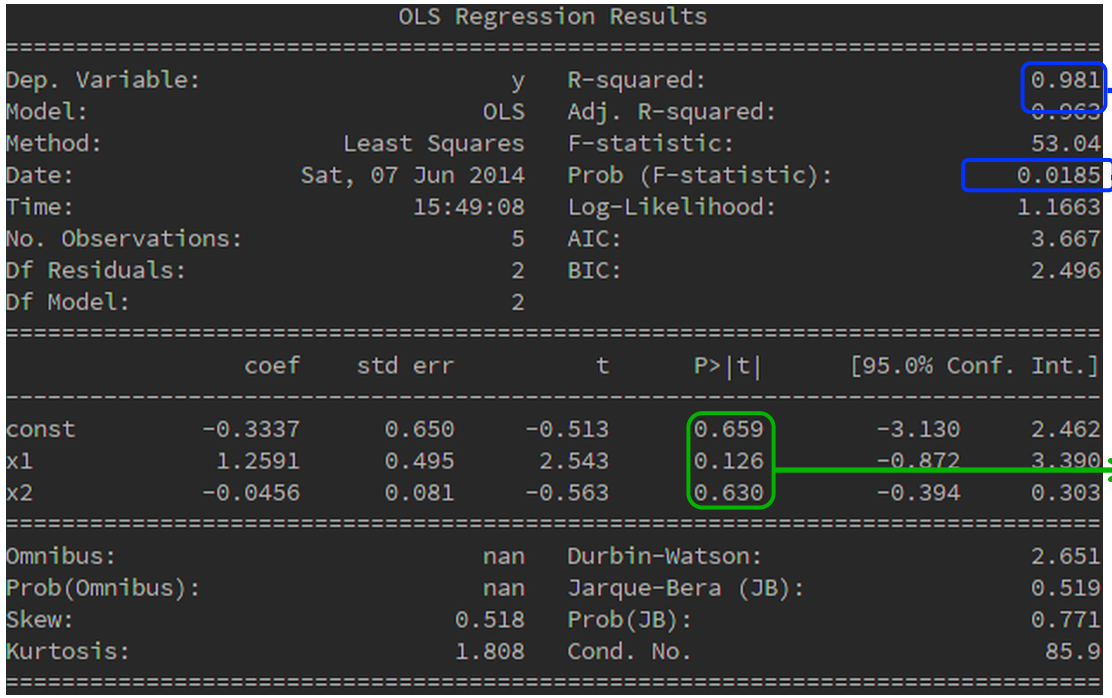
Each coefficient is really significant. Can also think of this as a Partial F-test.

"The average effect on Y of a one unit increase in X_2, holding all other predictors (X_1 & X_3) fixed, is 0.4836"

- However, interpretations are generally pretty hazardous due to correlations among predictors.
- p-values for each coefficient ≈ 0, so might be okay here

Note: Magnitude of the Beta coefficients is NOT how to determine whether predictor contributes. Why?

Interpretation



OLS Regression Results

Dep. Variable:	y	R-squared:	0.981
Model:	OLS	Adj. R-squared:	0.963
Method:	Least Squares	F-statistic:	53.04
Date:	Sat, 07 Jun 2014	Prob (F-statistic):	0.0185
Time:	15:49:08	Log-Likelihood:	1.1663
No. Observations:	5	AIC:	3.667
Df Residuals:	2	BIC:	2.496
Df Model:	2		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-0.3337	0.650	-0.513	0.659	-3.130	2.462
x1	1.2591	0.495	2.543	0.126	-0.872	3.390
x2	-0.0456	0.081	-0.563	0.630	-0.394	0.303

Omnibus:	nan	Durbin-Watson:	2.651
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.519
Skew:	0.518	Prob(JB):	0.771
Kurtosis:	1.808	Cond. No.	85.9

Proportion of Variance Explained by model is 98.1%

Model fit seems pretty good

But there appears to be some issues with p-values. **Interpreting coefficients hazardous** (may have correlation issues)

May want to…

- Plot x1 and x2
- Try modeling with just x1
- Try modeling with just x2
- We'll look into other techniques later

Questions

- Describe or interpret each of the components below.

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

- How to assess accuracy of model above?
  - RSS? RSE? R^2? How are they related?

- How would you compare a model nested within another model?
  - How is this related to the p-value for the t-statistic in the usual model output?
  - How is this related to the F-statistic in the usual linear regression output?

Questions

- Describe or interpret each of the components below.

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

β is the average effect on Y of a one unit increase in X_k, holding all other predictors fixed, is 0.4836

ε is the error term. ε ~ i.i.d. N(0, σ^2)

- How to assess accuracy of model above?
  - RSS? RSE? R^2? How are they related? See slide 24

- How would you compare a model nested within another model?
  - How is this related to the p-value for the t-statistic in the usual model output?
  - How is this related to the F-statistic in the usual linear regression output?

See slide 27