Linear Regression

*** *More slides here*
https://github.com/gSchool/DSI_Lectures/tree/master/linear-regression

Overview

Machine Learning

- Regression vs Classification
- Supervised vs Unsupervised

Other models that use linear regression

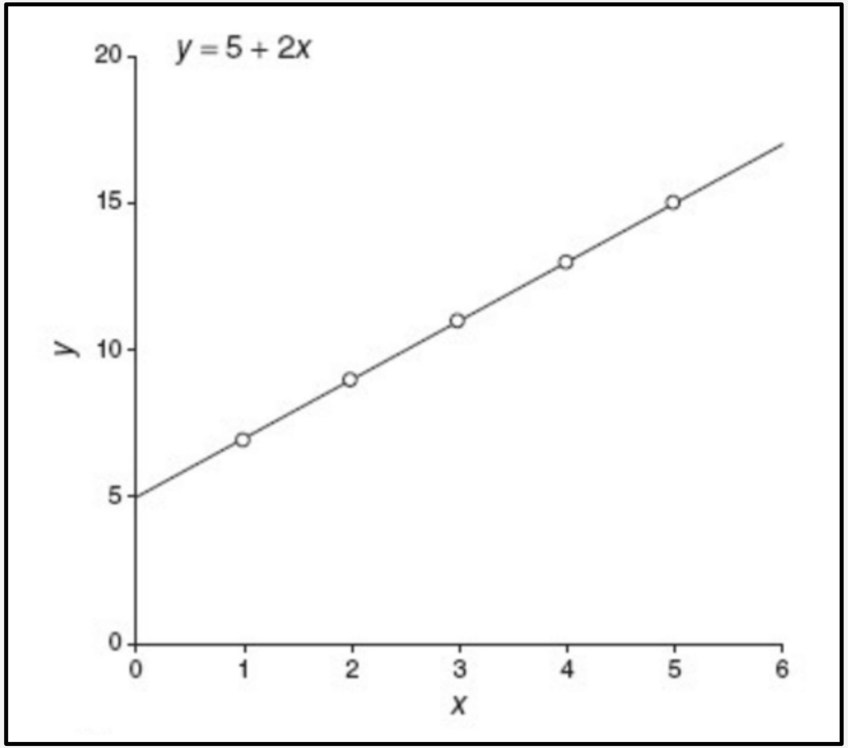
- Logistic Regression
- Multilayer Perceptrons (Deep Learning)

Getting Started

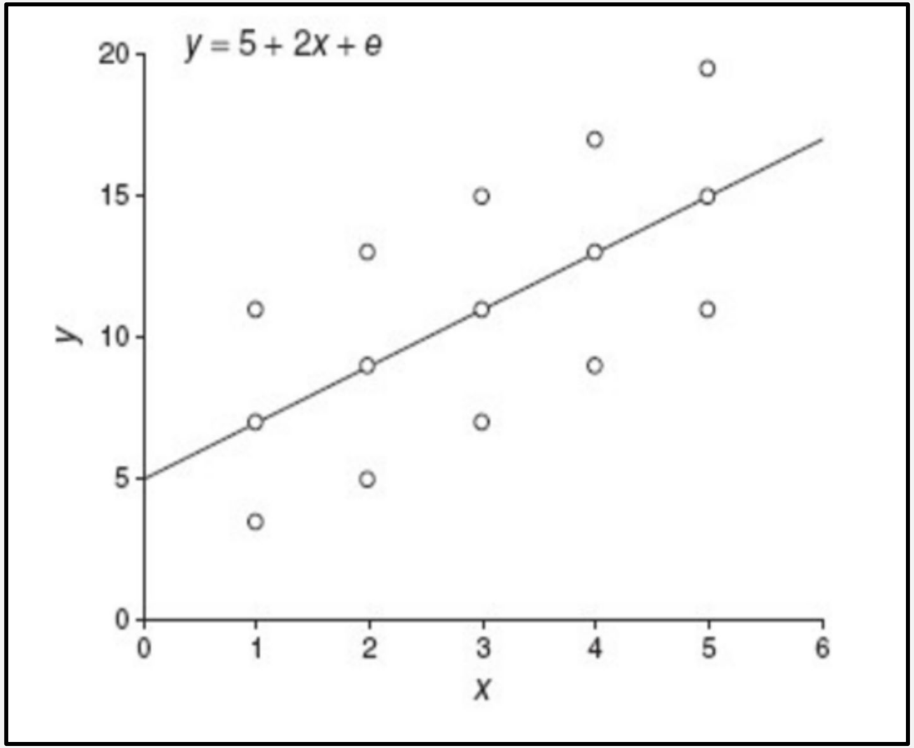
Interactive Linear Regression Demonstrations

- http://setosa.io/ev/ordinary-least-squares-regression/
- https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html
- http://miabellaai.net/demo.html

Exact Fit



Inexact Fit



The Model

Simple Linear Regression

- The World
  - what you're presuming the world looks like:

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

- β_0 and β_1 are unknown constants that represent the intercept and slope
- ϵ, the error term, is i.i.d $N(0, \sigma^2)$

- The Model
  - what you've created from data to estimate the world:

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

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

Matrix Form

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

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

Target:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Coefficient Matrix β

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

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

Linear Regression Libraries

- StatsModels
  - http://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html
- Scikit Learn
  - http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

DEMO #1

Model Evaluation

Statsmodels Summary

OLS Regression Results

Dep. Variable:	mpg	**R-squared:**	0.708
Model:	OLS	**Adj. R-squared:**	0.704
Method:	Least Squares	**F-statistic:**	186.9
Date:	Mon, 05 Mar 2018	**Prob (F-statistic):**	9.82e-101
Time:	12:20:06	**Log-Likelihood:**	-1120.1
No. Observations:	392	**AIC:**	2252.
Df Residuals:	386	**BIC:**	2276.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	46.2643	2.669	17.331	0.000	41.016	51.513
cylinders	-0.3979	0.411	-0.969	0.333	-1.205	0.409
displacement	-8.313e-05	0.009	-0.009	0.993	-0.018	0.018
weight	-0.0052	0.001	-6.351	0.000	-0.007	-0.004
acceleration	-0.0291	0.126	-0.231	0.817	-0.276	0.218
hp	-0.0453	0.017	-2.716	0.007	-0.078	-0.012

Omnibus:	38.561	**Durbin-Watson:**	0.865
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	52.737
Skew:	0.706	**Prob(JB):**	3.53e-12
Kurtosis:	4.111	**Cond. No.**	3.87e+04

RMSE

RSE (aka RMSE)

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

R squared

$$Coefficient\ of\ Determination \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$Sum\ of\ Squares\ Total \rightarrow \quad SST = \sum (y - \bar{y})^2$$

$$Sum\ of\ Squares\ Regression \rightarrow \quad SSR = \sum (y' - \overline{y'})^2$$

$$Sum\ of\ Squares\ Error \rightarrow \quad SSE = \sum (y - y')^2$$

t-statistic

Suppose we wish to test

$$H_0\colon \beta_1 = \beta_{1,0}$$

$$H_1\colon \beta_1 \neq \beta_{1,0}$$

$$\frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

P-value

From the df and the t-statistic, find the p-value:
https://faculty.washington.edu/heagerty/Books/Biostatistics/TABLES/t-Tables/

Confidence interval

From the df and percent, find the t-statistic:
https://faculty.washington.edu/heagerty/Books/Biostatistics/TABLES/t-Tables/

Beta +/- critical_value * SE

DEMO #2

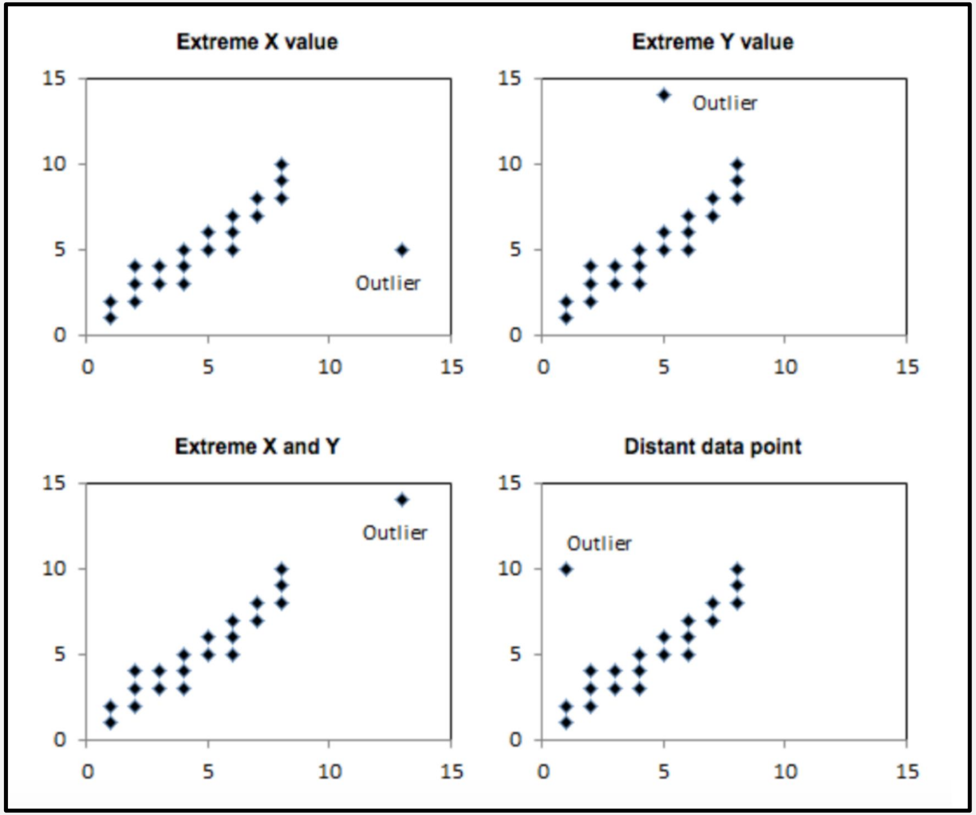
Assumptions of Linear Regression

- Assumptions of Linear Regression
  - Linearity
    - We assume it's possible
  - Constant Variance (Homoscedasticity)
    - Our variance shouldn't change as y or X gets bigger
  - Independence of Errors
    - We should gain no information from knowing the error of a different data point
  - Normality of Errors
    - Errors should be normally distributed
  - Lack of Multicollinearity
    - We shouldn't be measuring the same thing in multiple ways

We can't always meet these assumptions, and often have to find ways to combat that reality.

Residuals

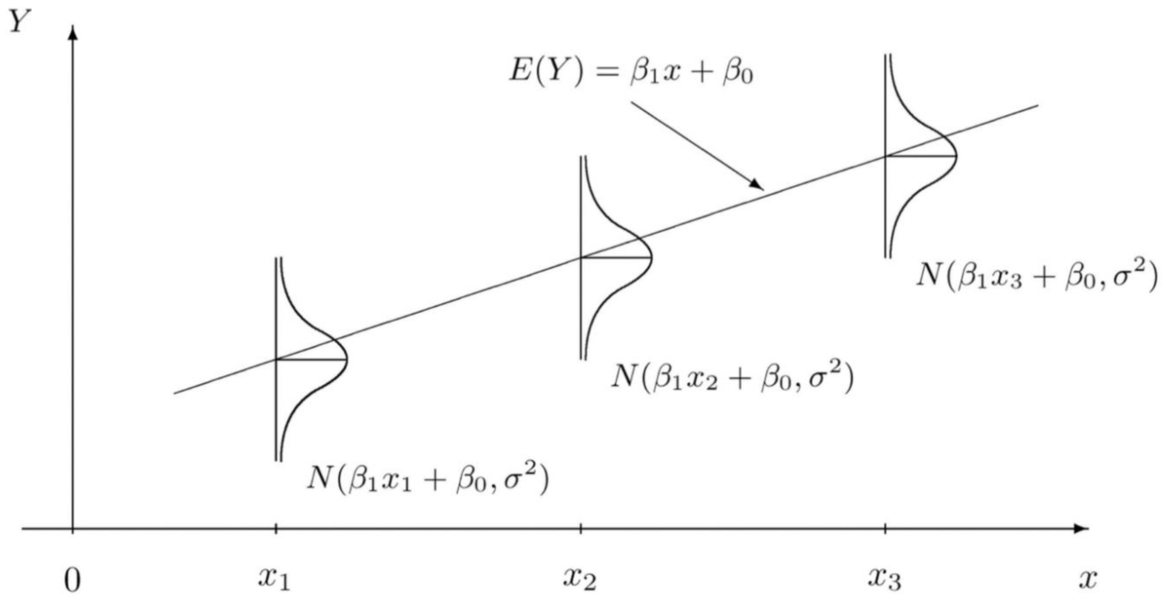
Types of Outliers



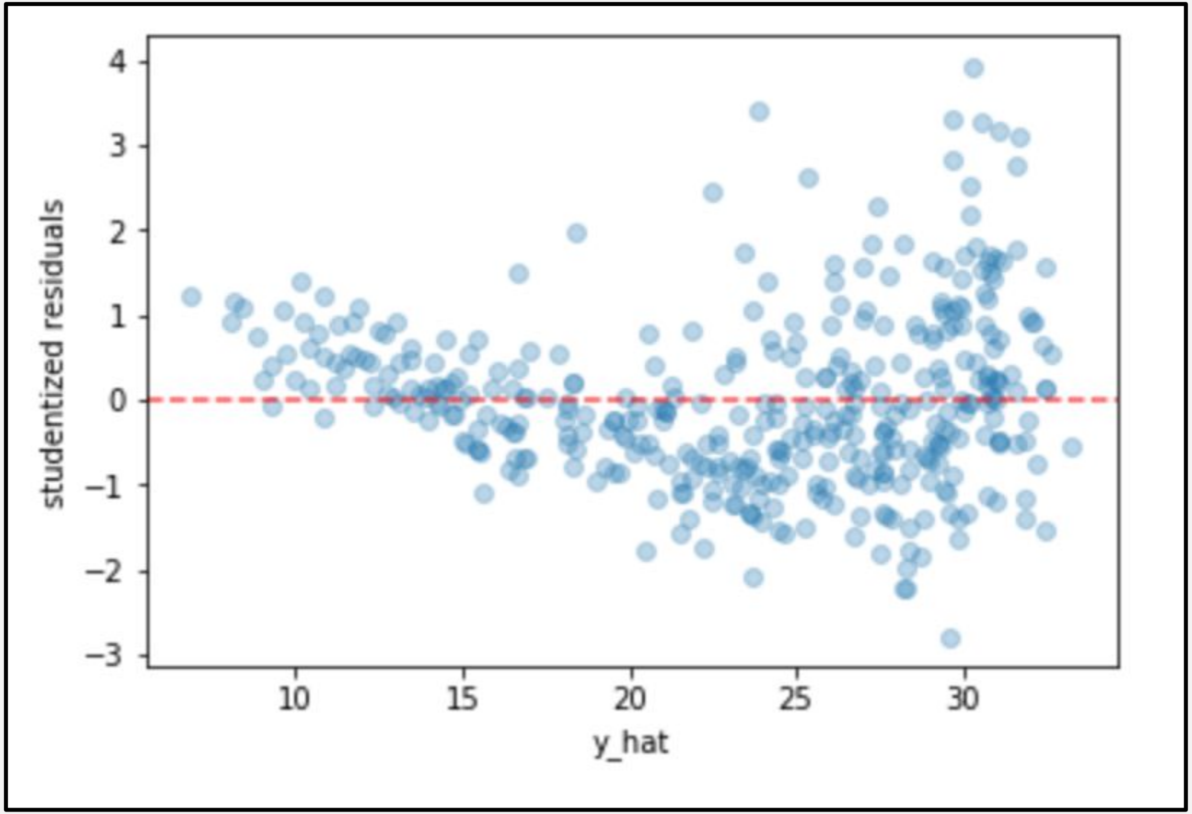
Residuals

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \epsilon_i \overset{i.i.d.}{\sim} Normal(0, \sigma^2)$$

Intercept   Error/Noise



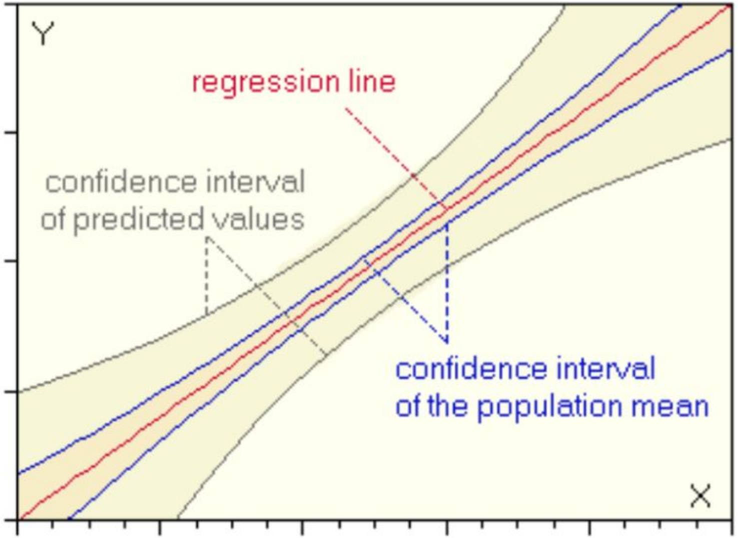
Studentized Residuals



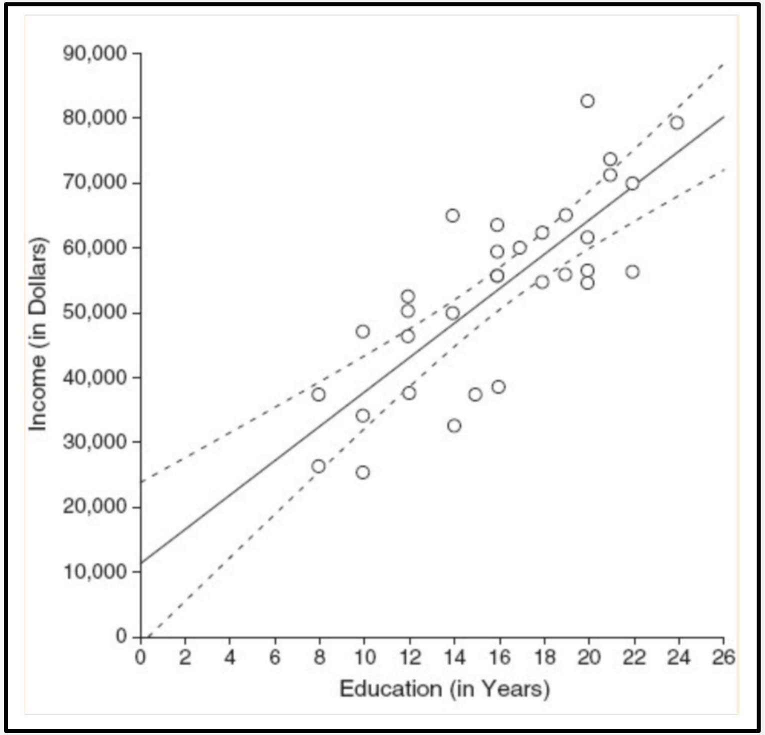
Studentized Residuals

Studentized Residuals have a t-distribution...

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$



Even better still, we can **"Studentize"** the errors by dividing, not by the "global" standard error for our model, but by the standard error of our model at the particular value of y where the residual occurred. Our confidence intervals change depending on how much data we have seen in a particular region. If we've seen a lot of data, our intervals are tight; otherwise, they are wide. So, it takes "more" for a data point to be considered an outlier if it is in a region in which we have little data.



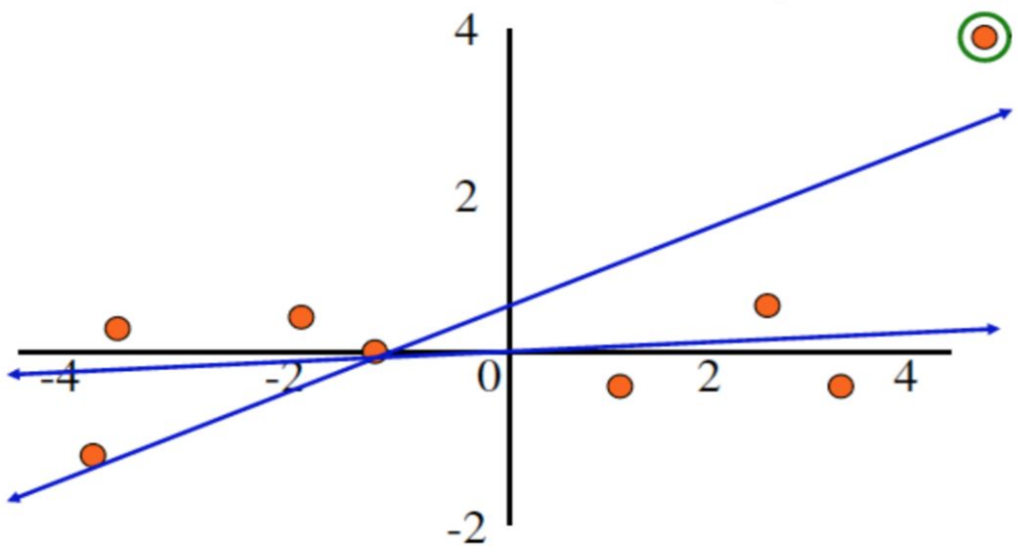
Studentized Residuals

Leverage

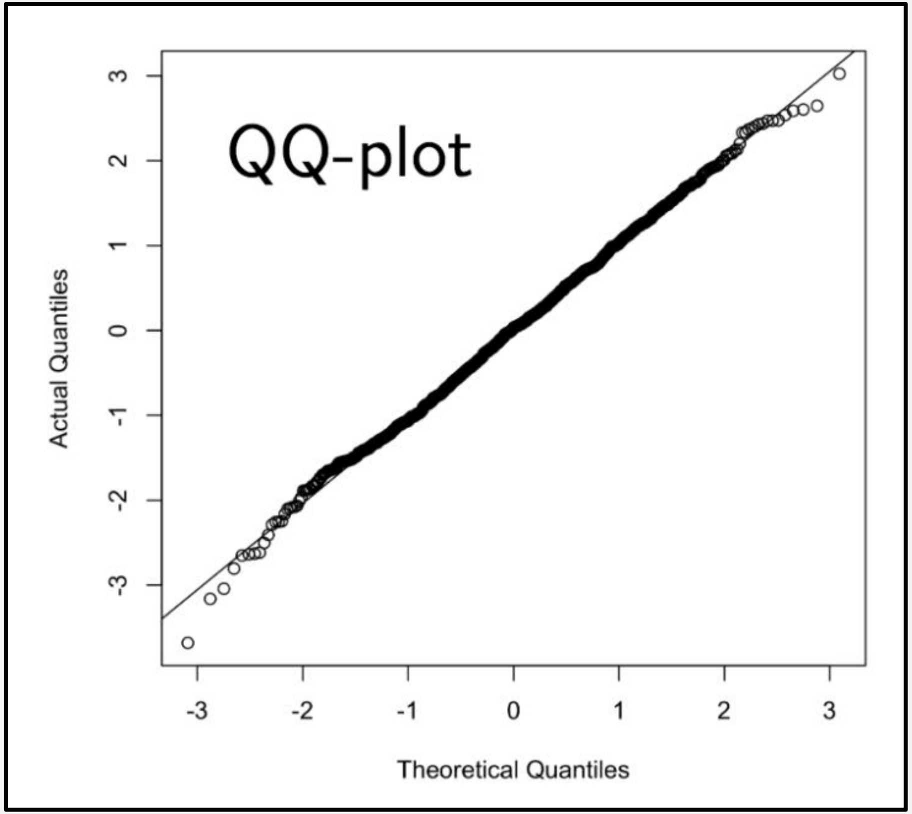
The *hat* matrix H *"puts the hat on"* $\boldsymbol{Y}$ projecting $\boldsymbol{Y}$ onto the (least squares) closest vector to $\boldsymbol{Y}$ in the column space of $\boldsymbol{x}$, $\hat{\mathbf{Y}} \in \mathcal{R}(\boldsymbol{x})$

$$H = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$$

$$\hat{\mathbf{Y}} = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} = H\mathbf{Y}$$



QQ Plots



If a set of observations is approximately normally distributed, a normal quantile-quantile (QQ) plot of the observations will result in an approximately straight line.

DEMO #3

Categorical Values

DEMO #4