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

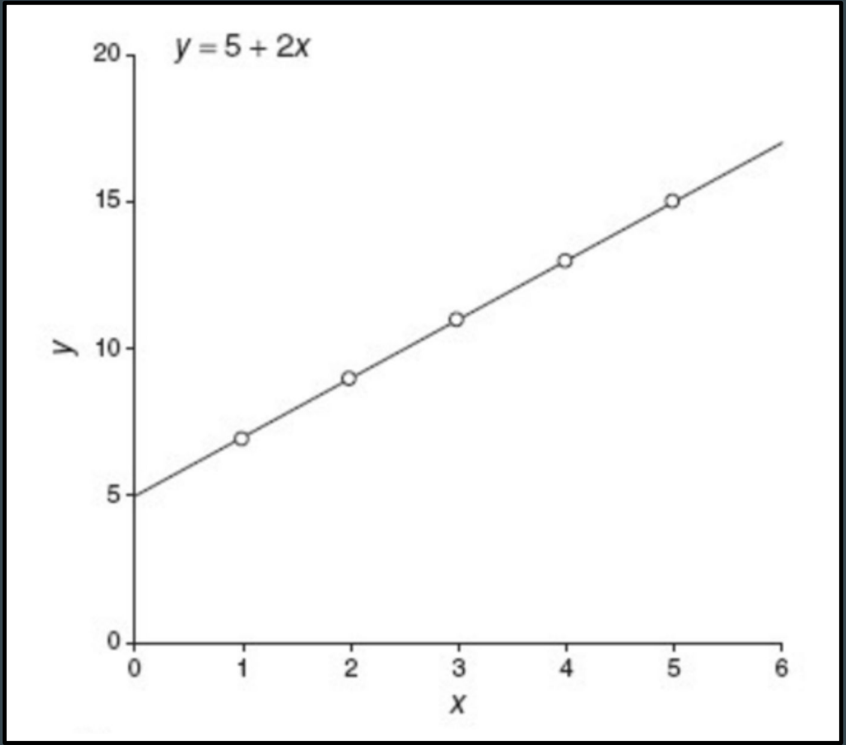
- Logistic Regression
- Multilayer Perceptrons (Deep Learning)

Getting Started

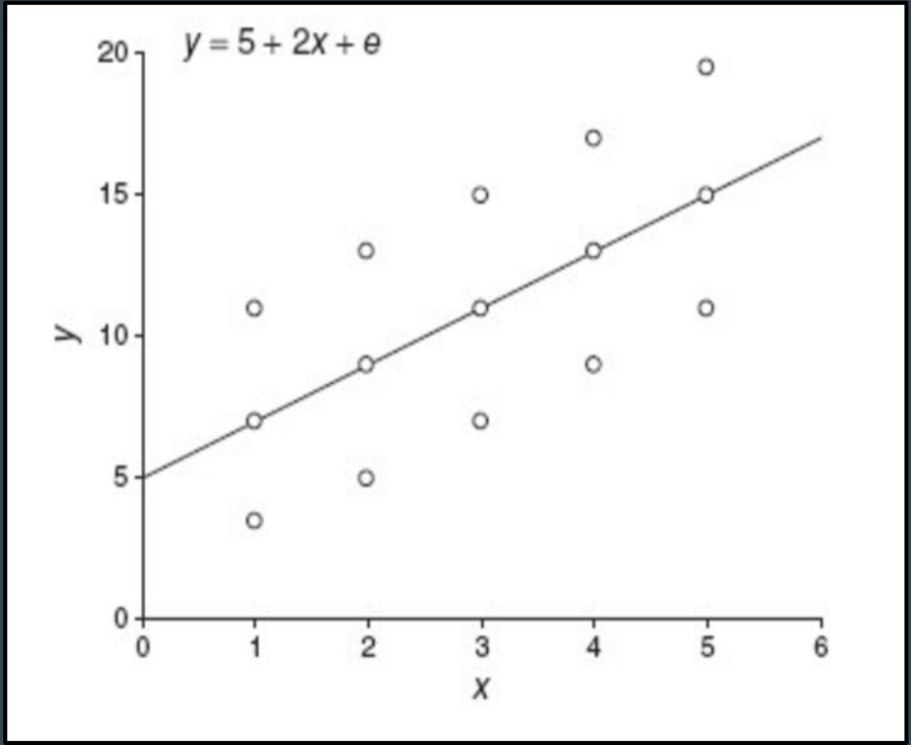
Interactive Linear Regression Demonstrations

- 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

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

Target:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Coefficient Matrix β

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

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

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

- results.summary()
- Sklearn does not provide these results
- Will discuss these values in later slides

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

Interpreting Beta coefficients

2 When the X_1 variable **increases by one unit** then the Y variable **increases by β_1 units**, all other variables in the model being **kept at the same level**. 3

That is, if we do not change other variables and only change X_1 by increasing it by one unit, then the Y variable will increase by β_1 units.

Beta coefficients

Importance of the Normality assumption about errors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k$$

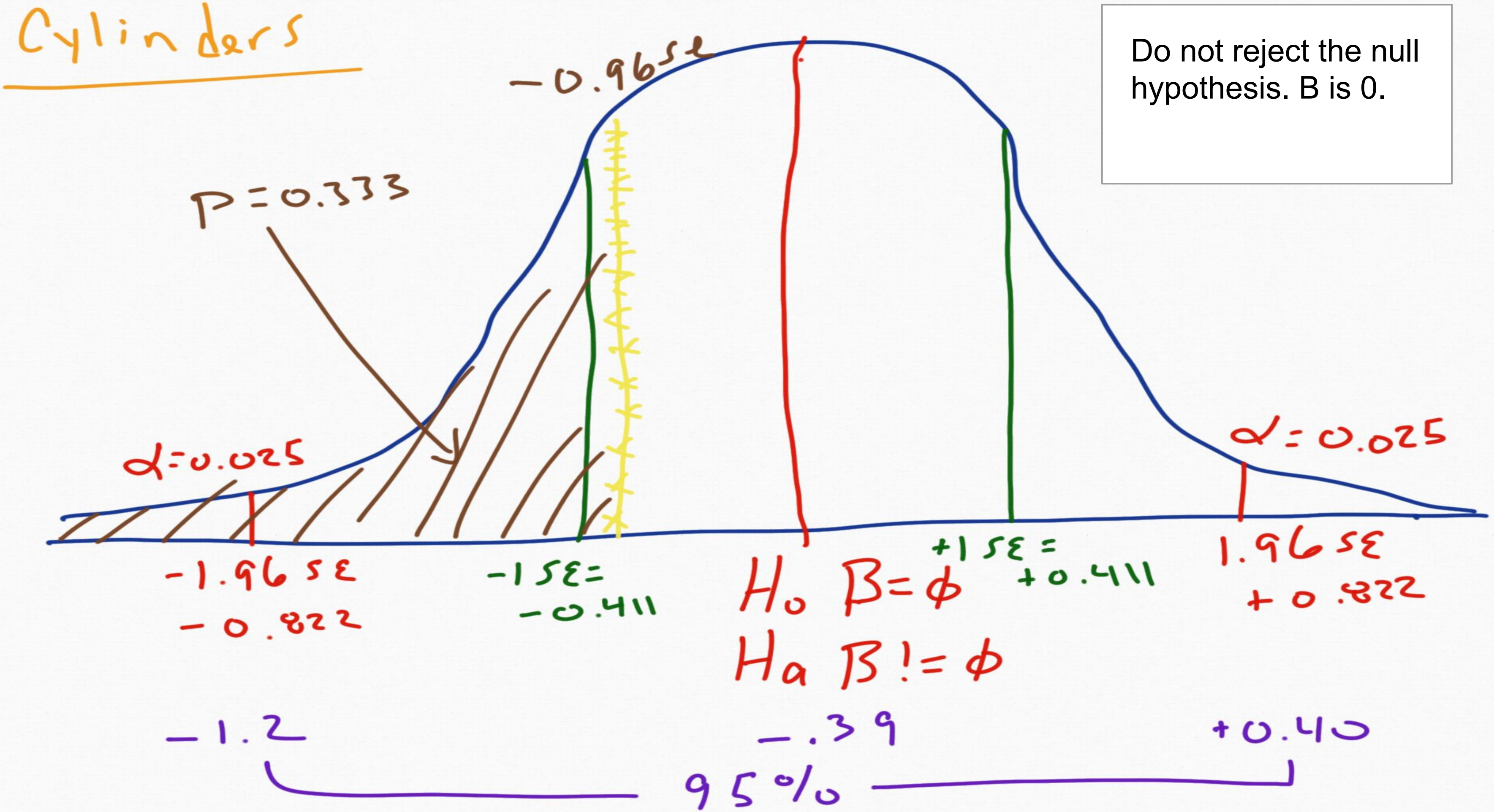
Estimated Model : $$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_k X_k$$

The *b*'s can be considered as random variables…,

$$b_0 \sim Normal(\beta_0,\quad some\ std)$$

$$b_1 \sim Normal(\beta_1,\quad some\ std)$$

Do not reject the null hypothesis. B is 0.



RMSE or Loss, during gradient descent

RSE (aka RMSE)

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

R2

R-squared is the “percent of variance explained” by the model.

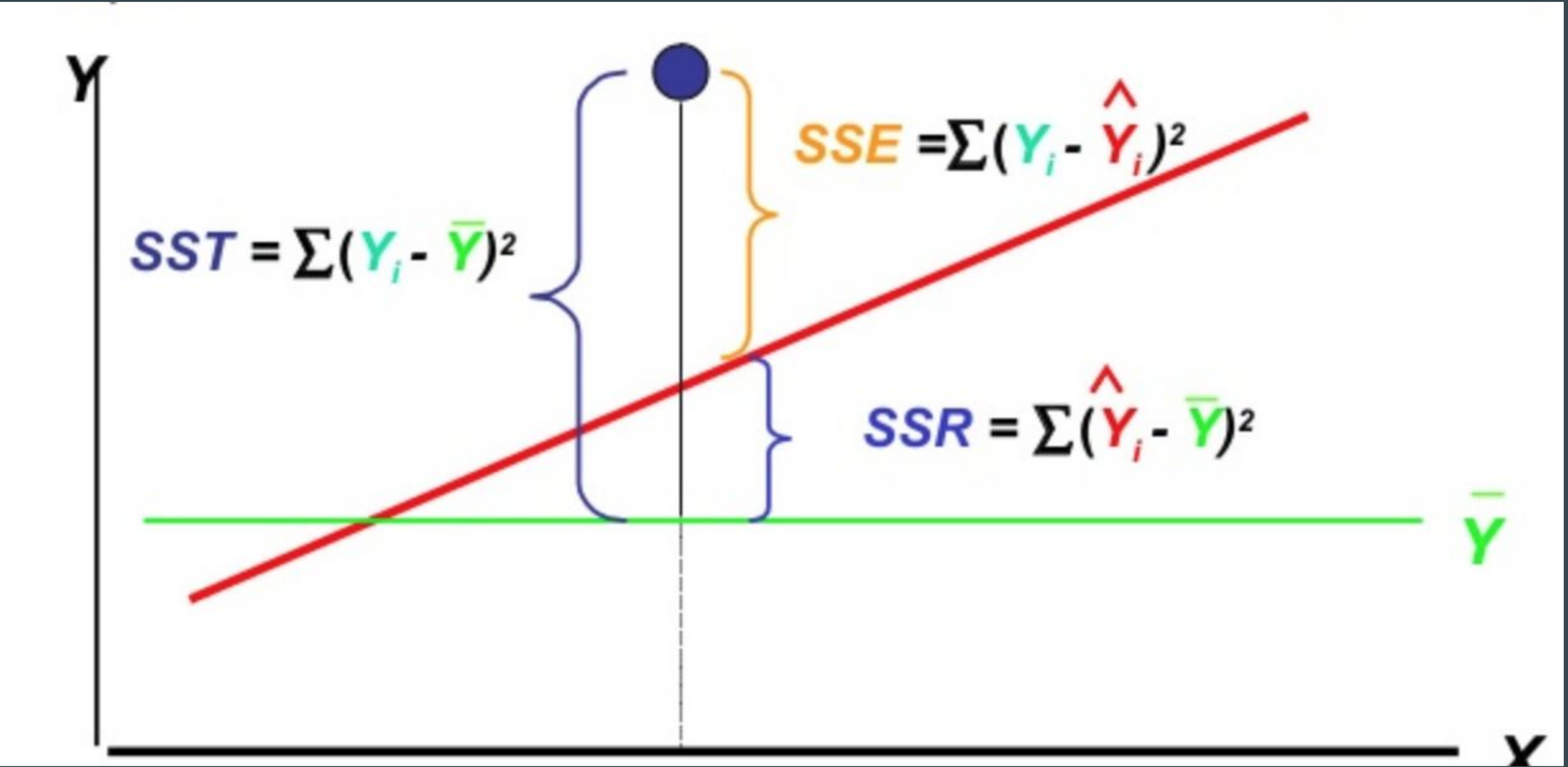
$$\text{Coefficient of Determination} \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

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

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

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

R2



Adjusted R2

Regression Statistics	
Multiple R	0.6888117
R Square	0.47446156
Adjusted R Square	0.44973034
Standard Error	7353.74751
Observations	90

- Mere addition of X variables always increases R-square.
- Adj. R-square adjusts the R-square for the number of X variables in the model.

F-statistic

That being said, the null hypothesis of the F-test is that the data can be modeled accurately by setting the regression coefficients to zero. The alternative hypothesis is that at least one of the regression coefficients should be non-zero. If the F-distribution provides a p-value that is lower than some threshold $\alpha = 0.05, 0.01$, then we reject the null hypothesis, and and say that our model is, in fact, "doing something with its life." The $F-$ statistic is computed as the ratio of two χ^2 distributed variables, discussed below.

AIC/BIC

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are based on the log-likelihood described in the previous section. Both measures introduce a penalty for model complexity, but the AIC penalizes complexity less severely than the BIC. The AIC and BIC are given by,

$$AIC = 2k - 2\ln(\mathcal{L}) \tag{12}$$

$$BIC = k\ln(N) - 2\ln(\mathcal{L}) \tag{13}$$

Skew and Kurtosis

Skew and kurtosis refer to the shape of a (normal) distribution. Skewness is a measure of the asymmetry of a distribution, and kurtosis is a measure of its curvature, specifically how peaked the curve is. These values are calculated as,

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)^{3/2}} \tag{18}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)^{2}} \tag{19}$$

Omnibus

The Omnibus test uses skewness and kurtosis to test the null hypothesis that a distribution is normal. In this case, we're looking at the distribution of the residual. If we obtain a very small value for $\Pr(\text{ Omnibus })$, then the residuals are not normally distributed about zero, and we should maybe look at our model more closely.

DEMO #2

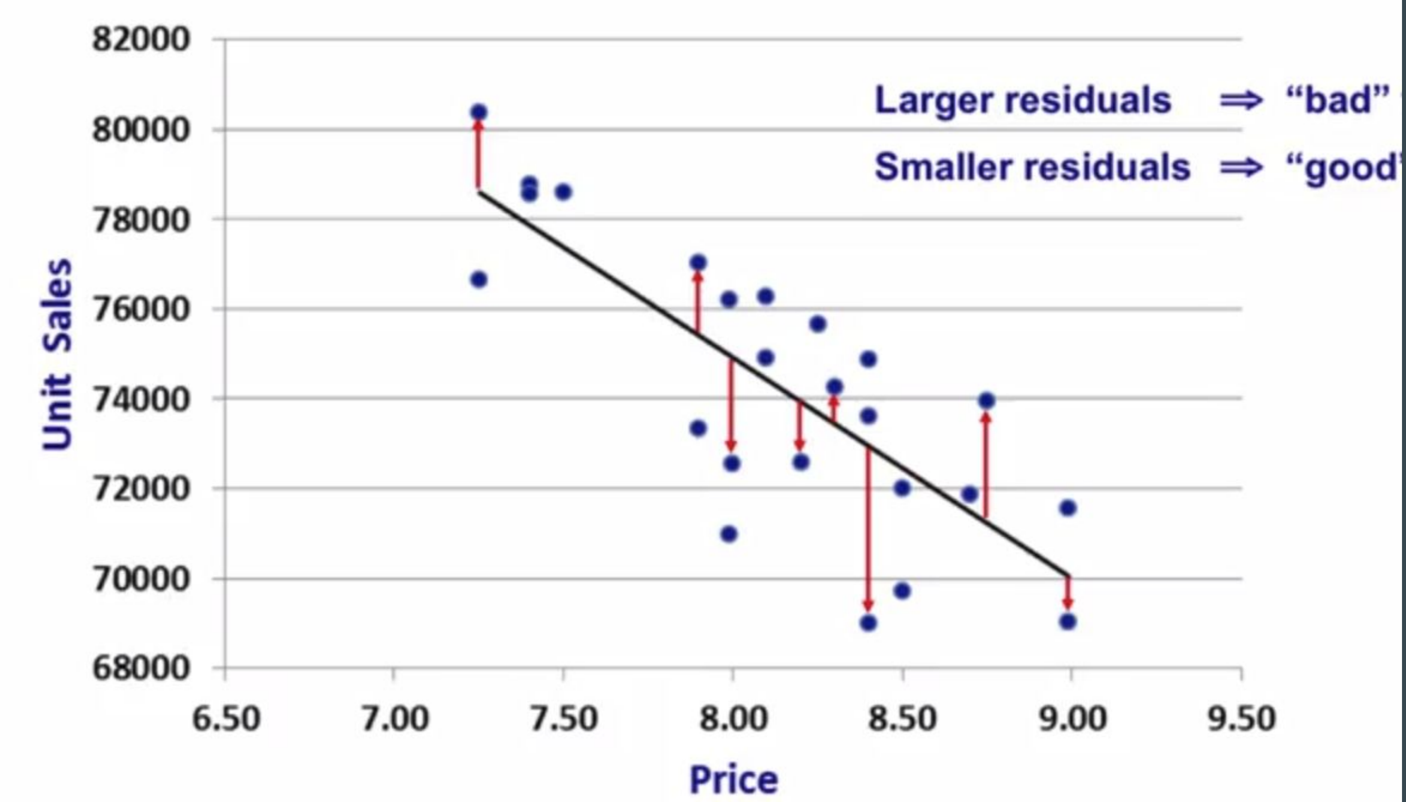
Assumptions of Linear Regression

- 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.

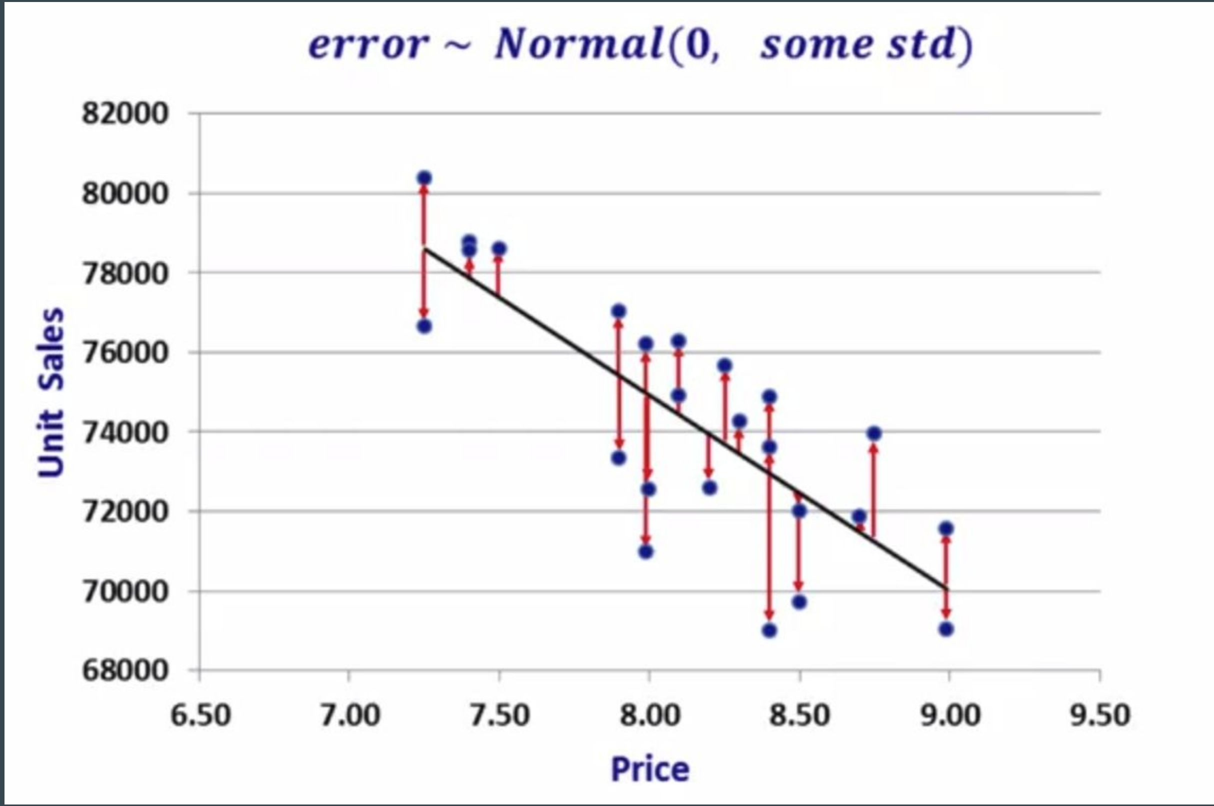
Residuals

Residuals



Larger residuals ⇒ "bad"
Smaller residuals ⇒ "good
Unit Sales
Price
82000
80000
78000
76000
74000
72000
70000
68000
6.50
7.00
7.50
8.00
8.50
9.00
9.50

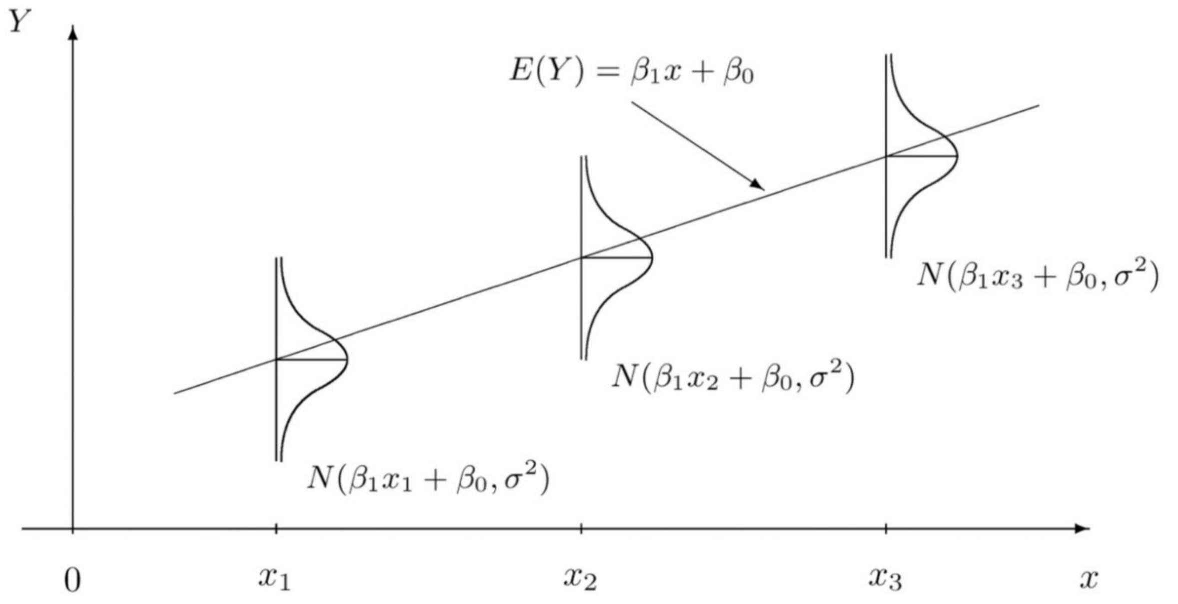
Normally distributed residuals



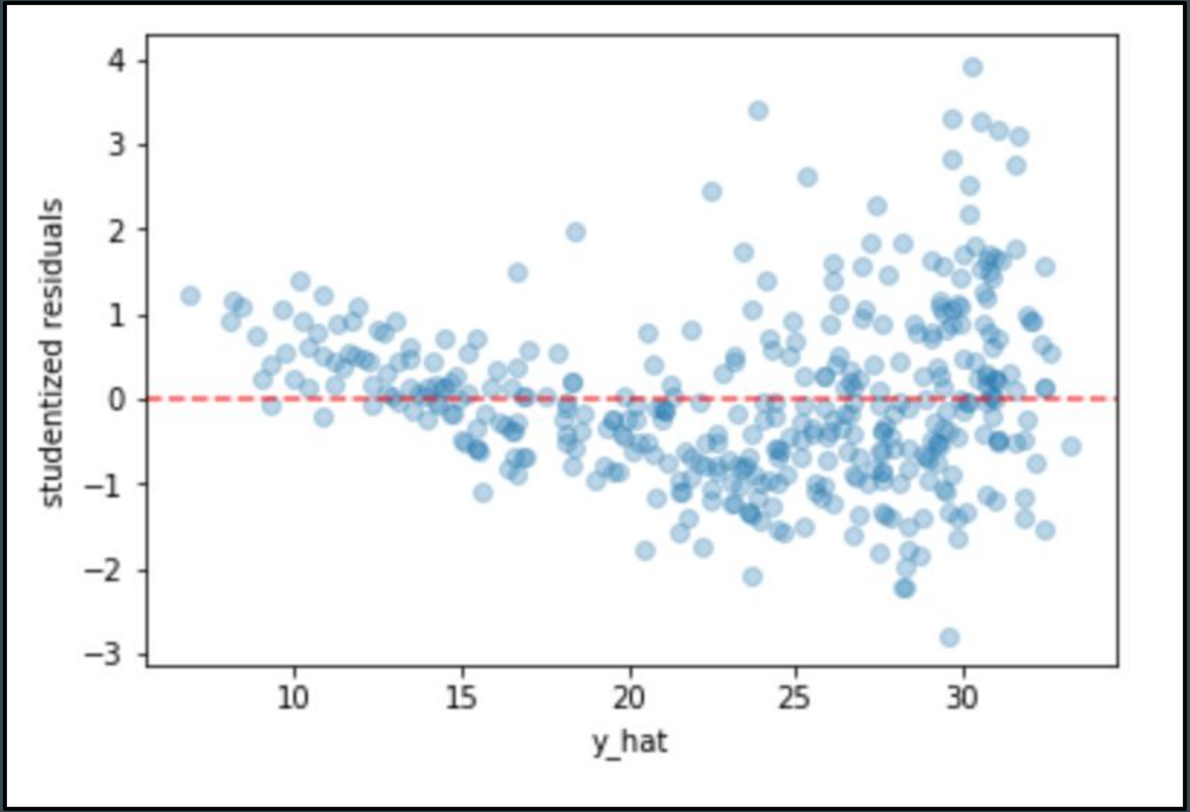
Residuals

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

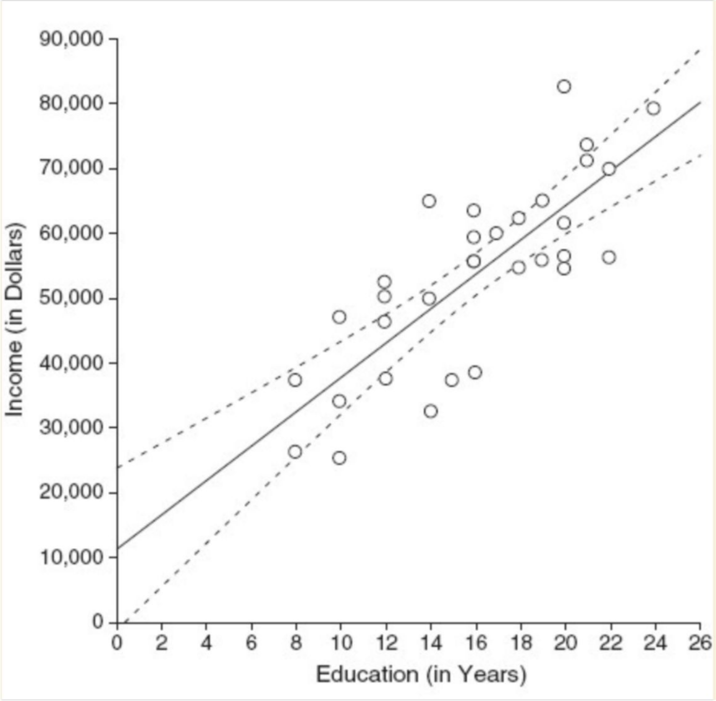
Intercept Error/Noise



Studentized Residuals

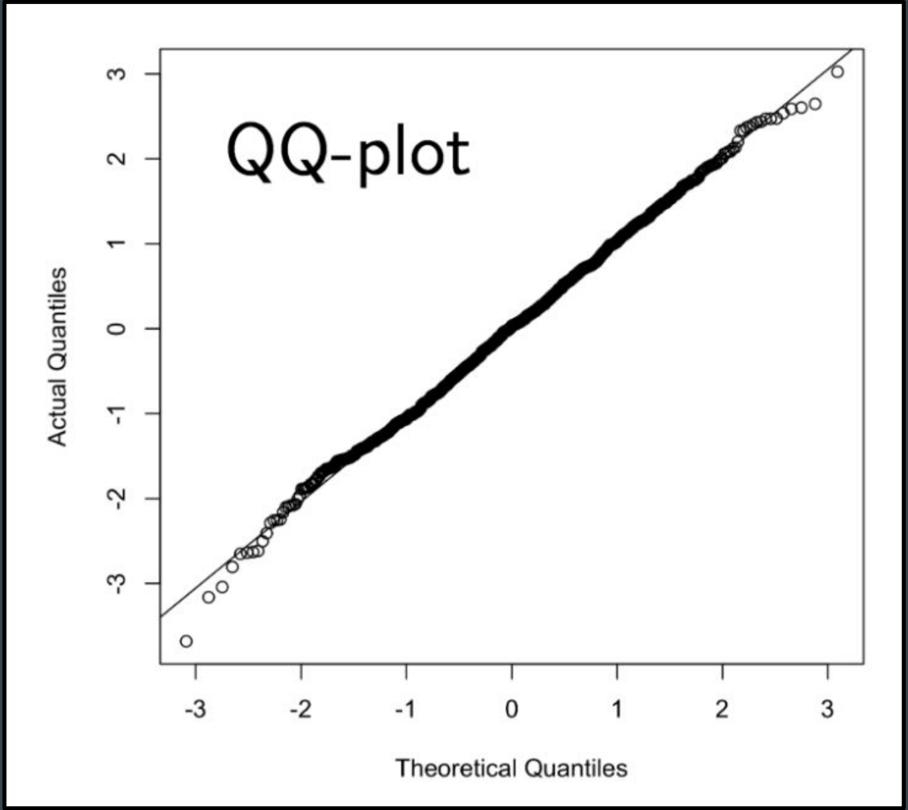


Even better still, we can **"Studentize"** the errors by dividing, not by the "global" standard error for our model, but by the standard error of our model at the particular value of y where the residual occurred. Our confidence intervals change depending on how much data we have seen in a particular 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

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