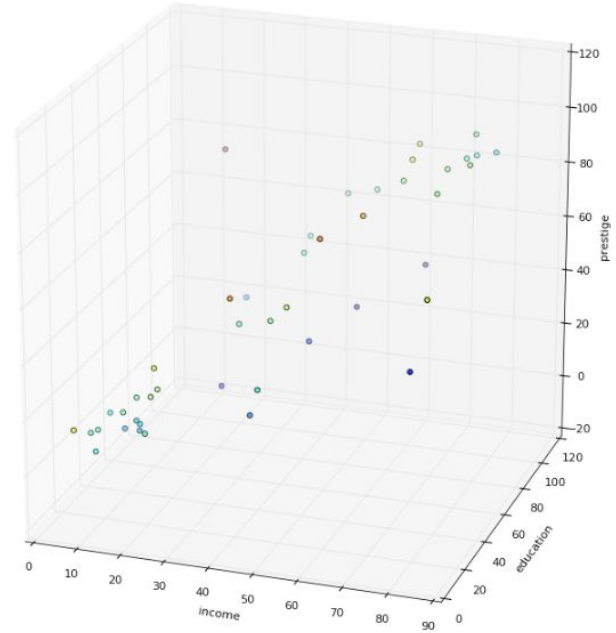


# Linear Regression

DSI SEA5, jf.omhover, Sep 21, 2016



# Linear Regression

DSI SEA5, jf.omhover, Sep 21, 2016

## STANDARDS

- **Describe**, **interpret**, and **visualize** the model form of linear regression:  $Y = B_0 + B_1X_1 + B_2X_2 + \dots$
- **Relate** Beta vector solution of Ordinary Least Squares to the cost function (residual sum of squares)
- **State** and troubleshoot the assumptions of linear regression model
- **Perform** OLS with statsmodels and interpret the output: Beta coefficients, p-values,  $R^2$
- How can one **detect** outliers?



# Linear Regression

DSI SEA5, jf.omhover, Sep 21, 2016

## OBJECTIVES

- **Relate** linear regression with general machine learning
- **State** assumptions of linear regression
- **Estimate** a linear regression model
- **Evaluate** a linear regression model
- **Recognize and fix** common problems





# What's the Big Idea ?

Learning / Estimating FUNCTIONS

# Reality VS Model : assumption, learning and error



## REALITY

	type	income	education	prestige
accountant	prof	62	86	82
pilot	prof	72	76	83
architect	prof	75	92	90
author	prof	55	90	76
chemist	prof	64	86	90
minister	prof	21	84	87
professor	prof	64	93	93
dentist	prof	80	100	90
reporter	wc	67	87	52
engineer	prof	72	86	88
undertaker	prof	42	74	57
lawyer	prof	76	98	89

data

$(x_1, y_1)$

...

$(x_n, y_n)$

$x \ y$

**OBJECTIVE:**  
descriptive  
predictive  
normative

...

$$\sum (y_i - \hat{f}(x_i))^2$$

COST FUNCTION

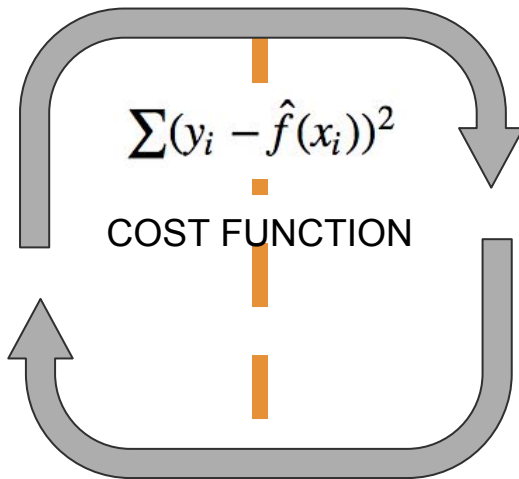
## MODEL

$$y = f(x) + \epsilon$$

take a function as  
an assumption

$$\hat{y} = \hat{f}(x)$$

Estimator  
of the function



# Reality VS Model : LINEAR functions



## REALITY

	type	income	education	prestige
accountant	prof	62	86	82
pilot	prof	72	76	83
architect	prof	75	92	90
author	prof	55	90	76
chemist	prof	64	86	90
minister	prof	21	84	87
professor	prof	64	93	93
dentist	prof	80	100	90
reporter	wc	67	87	52
engineer	prof	72	86	88
undertaker	prof	42	74	57
lawyer	prof	76	98	89

data

$(x_1, y_1)$

...

$(x_n, y_n)$

$x \ y$

## OBJECTIVE:

descriptive  
predictive  
normative

...

## MODEL

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

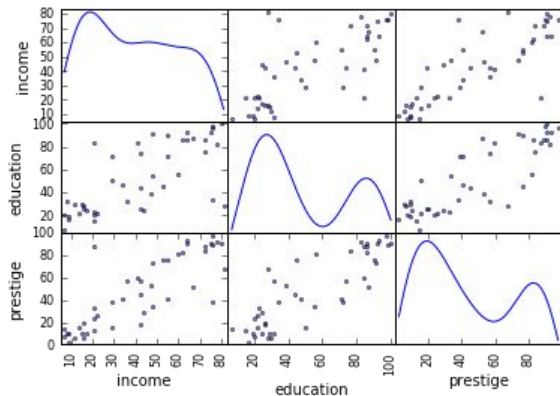
We make the assumption that we  
have a linear relation

# From reality to model : general process



## REALITY

- 1) Having a data sample  
Observing an underlying behavior



## MODEL

- 2) Make an assumption  
on the model underlying the data

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

linear relation  
(+ assumptions)

- 3) Find the instance of the model  
that fits with data sample

# Framing the problem : linear regression



## REALITY

	type	income	education	prestige
accountant	prof	62	86	82
pilot	prof	72	76	83
architect	prof	75	92	90
author	prof	55	90	76
chemist	prof	64	86	90
minister	prof	21	84	87
professor	prof	64	93	93
dentist	prof	80	100	90
reporter	wc	67	87	52
engineer	prof	72	86	88
undertaker	prof	42	74	57
lawyer	prof	76	98	89

data

X : independant variables

Y : dependant variable

**DEFINE A CRITERIA  
OPTIMIZE IT  
OVER PARAMETERS**

## MODEL

**PROBLEM 1:**  
identify model class  
verify assumptions

$$y \approx \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \dots + \beta_p x_p$$

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

**PROBLEM 2:**  
find actual  
coefficients/parameters  
values





# The Simple Linear Case

# Framing the problem : SIMPLE linear regression



**COST FUNCTION (Residual Sum of Squares)**

$$RSS = \sum (y_i - \hat{f}(x))^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x)^2$$

O.L.S.

**REALITY**

**DATA**

$(x_1, y_1)$

...

$(x_n, y_n)$

$x \ y$

**MODEL**

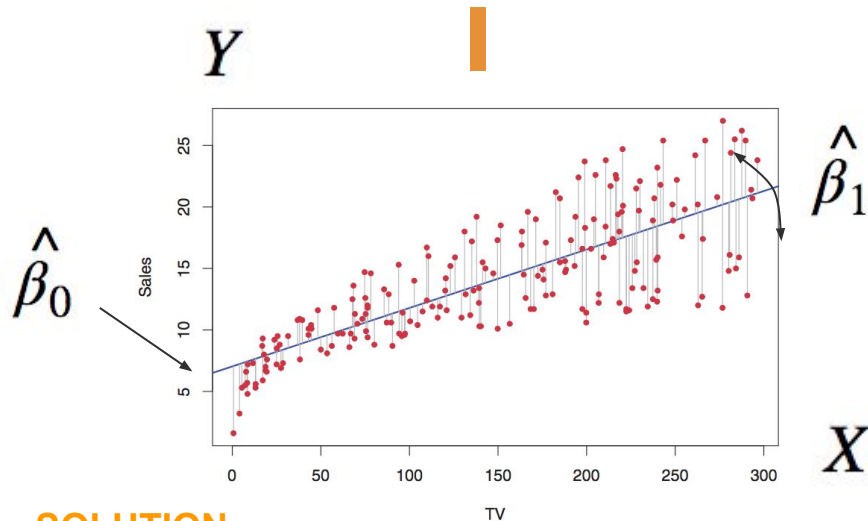
**model class**

$$y \approx \beta_0 + \beta_1 x$$

**PROBLEM**

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

**model instance  
estimator  
parameters**



**SOLUTION**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# Evaluation as a model explaining the outcome



## Residual Sum of Squares

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

## Total Sum of Squares

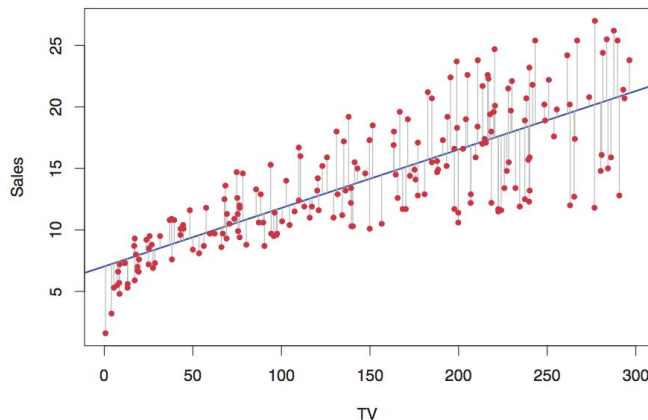
$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

## Explained Sum of Squares

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

## R squared statistic

$$R^2 = \frac{TSS - RSS}{TSS} \quad \text{in } [0, 1]$$



$$y = f(x) + \epsilon$$

**class**  $y \approx \beta_0 + \beta_1 x$

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

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

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i, \forall i < n$$

**residual**  $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

# Evaluation as for hypothesis



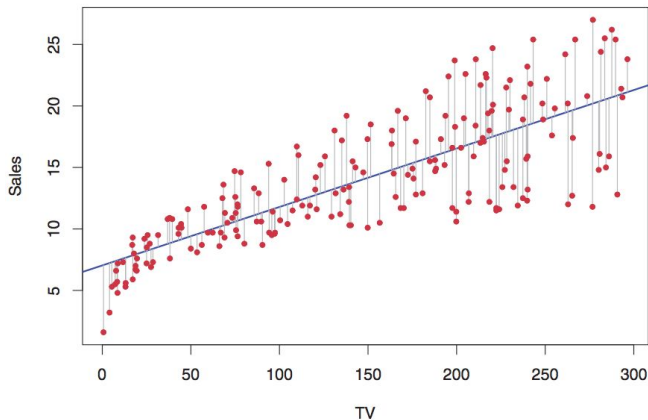
## Stating the null hypothesis

H0 : there is no relation between X and Y

H1 : there is some linear relation between X and Y

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$



$$y = f(x) + \epsilon$$

**class**  $y \approx \beta_0 + \beta_1 x$

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

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

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i, \forall i < n$$

**residual**  $e_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i$



# The Multi-Linear Case

# Framing the problem : multi-linear regression



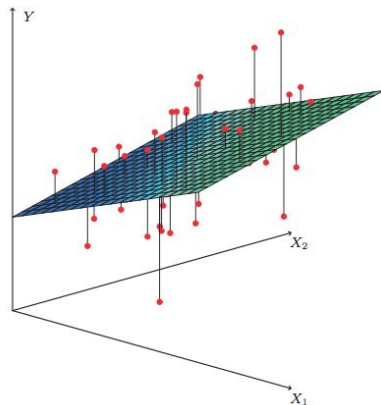
**COST FUNCTION (Residual Sum of Squares)**

**O.L.S.**

**REALITY**

**DATA**

	type	income	education	prestige
accountant	prof	62	86	82
pilot	prof	72	76	83
$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y$				
lawyer	prof	76	98	89



**MODEL**

**model class**

$$y \approx \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \dots + \beta_p x_p$$

**PROBLEM**

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

**model instance  
estimator  
parameters**

**SOLUTION**



# Framing the problem : multi-linear regression



**COST FUNCTION (Residual Sum of Squares)**

$$RSS(\beta) = (y - X\beta)^T(y - X\beta)$$

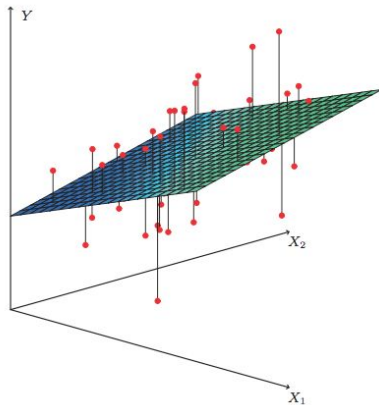
O.L.S.

**REALITY**

**DATA**

	type	income	education	prestige
accountant	prof	62	86	82
pilot	prof	72	76	83
lawyer	prof	76	98	89

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



**MODEL**

**model class**

$$y \approx X\beta$$

**PROBLEM**

$$\hat{y} = X\hat{\beta}$$

**model instance  
estimator  
parameters**

**SOLUTION**

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

# Evaluation as a model explaining the outcome



## Residual Sum of Squares

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

## Total Sum of Squares

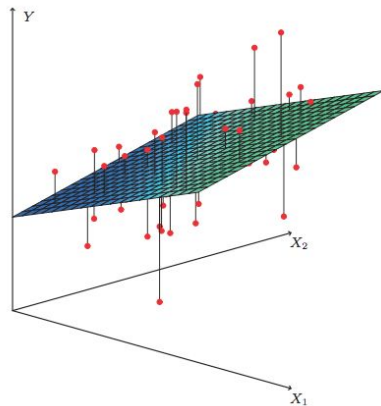
$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

## Explained Sum of Squares

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

## R squared statistic

$$R^2 = \frac{TSS - RSS}{TSS} \quad \text{in } [0, 1]$$



$$y = f(x) + \epsilon$$

class  $y \approx X\beta$

instance  $\hat{y} = X\hat{\beta}$

$$y = X\hat{\beta} + \epsilon$$

$$y_i = X\hat{\beta} + \epsilon_i$$

residual  $e_i = y_i - X\hat{\beta}$



# Evaluation as for hypothesis



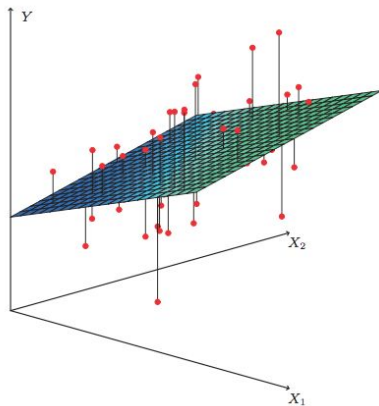
## Stating the null hypothesis

$H_0$  : there is no relation between  $X$  and  $Y$

$H_1$  : there is some linear relation between  $X$  and  $Y$

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1 : \exists \beta_i \neq 0$$



$$y = f(x) + \epsilon$$

class  $y \approx X\beta$

instance  $\hat{y} = X\hat{\beta}$

$$y = X\hat{\beta} + \epsilon$$

$$y_i = X\hat{\beta} + \epsilon_i$$

residual  $e_i = y_i - X\hat{\beta}$

# Fitting a linear regression model in Python



```
y = prestige['prestige']
x = prestige[['income', 'education']].astype(float)
x['const'] = 1

prestige_model = statsmodels.api.OLS(endog=y, \
                                     exog=x).fit()

prestige_model.summary()
```

OLS Regression Results

<b>Dep. Variable:</b>	prestige	<b>R-squared:</b>	0.828
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.820
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	101.2
<b>Date:</b>	Tue, 20 Sep 2016	<b>Prob (F-statistic):</b>	8.65e-17
<b>Time:</b>	15:00:41	<b>Log-Likelihood:</b>	-178.98
<b>No. Observations:</b>	45	<b>AIC:</b>	364.0
<b>Df Residuals:</b>	42	<b>BIC:</b>	369.4
<b>Df Model:</b>	2		
<b>Covariance Type:</b>	nonrobust		

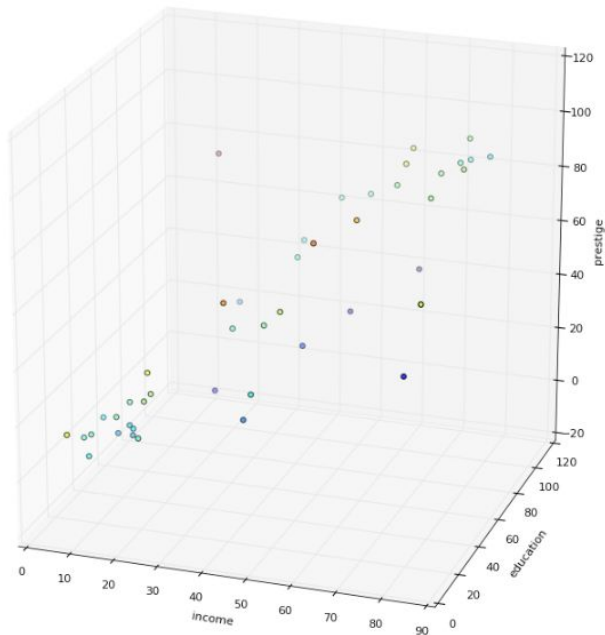
	coef	std err	t	P> t	[95.0% Conf. Int.]
<b>income</b>	0.5987	0.120	5.003	0.000	0.357 0.840
<b>education</b>	0.5458	0.098	5.555	0.000	0.348 0.744
<b>const</b>	-6.0647	4.272	-1.420	0.163	-14.686 2.556

<b>Omnibus:</b>	1.279	<b>Durbin-Watson:</b>	1.458
<b>Prob(Omnibus):</b>	0.528	<b>Jarque-Bera (JB):</b>	0.520
<b>Skew:</b>	0.155	<b>Prob(JB):</b>	0.771
<b>Kurtosis:</b>	3.426	<b>Cond. No.</b>	163.



# Assumptions

# Assumptions underlying Linear Regression



Linearity

Independence

Normal distribution of residuals

Homoscedasticity

Lack of multicollinearity

# Assumptions // Linearity

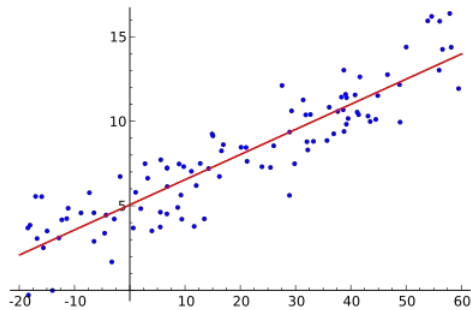


Take the function component of your model

$$y = f(x) + \epsilon$$

This function is assumed to be linear.

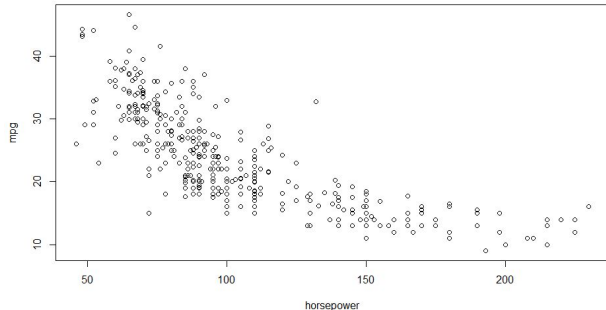
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1} x_{p-1} + \epsilon$$



OK

## TROUBLESHOOTING

If it's not x, it might be  $1/x$  or  $\log(x)$  or  $x^2$ ...



NOT OK

# Assumptions // Linearity



Take the function component of your model

$$y = f(x) + \epsilon$$

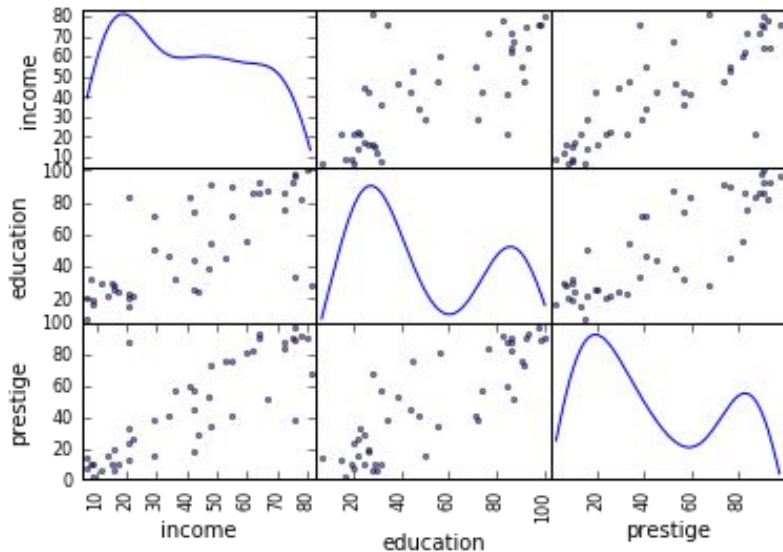
This function is assumed to be linear.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1} x_{p-1} + \epsilon$$

## TROUBLESHOOTING

If it's not  $x$ , it might be  $1/x$  or  $\log(x)$  or  $x^2$ ...

Scatter plots



```
pd.scatter_matrix(df, diagonal = 'kde')
```

# Assumptions // Linearity



Take the function component of your model

$$y = f(x) + \epsilon$$

This function is assumed to be linear.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1} x_{p-1} + \epsilon$$

## TROUBLESHOOTING

If it's not x, it might be  $1/x$  or  $\log(x)$  or  $x^2$ ...

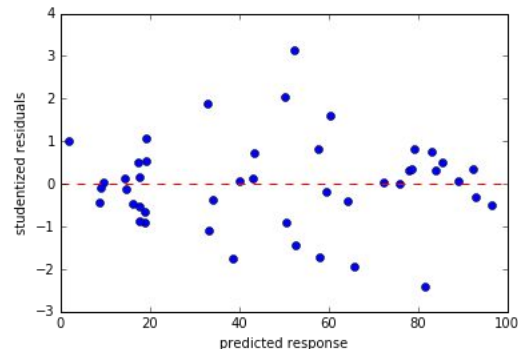
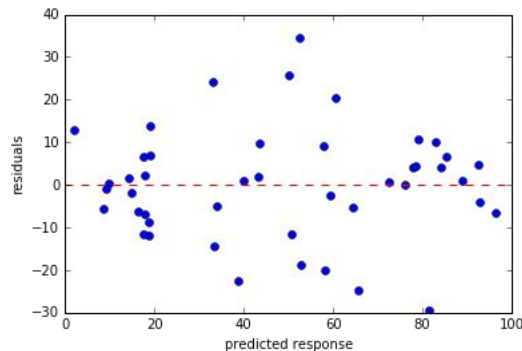
Scatter plots

Residual / Studentized residual plot

$$e_i = y_i - X\hat{\beta}$$

$$s_e^2 = \frac{1}{n-1} \sum e_i^2$$

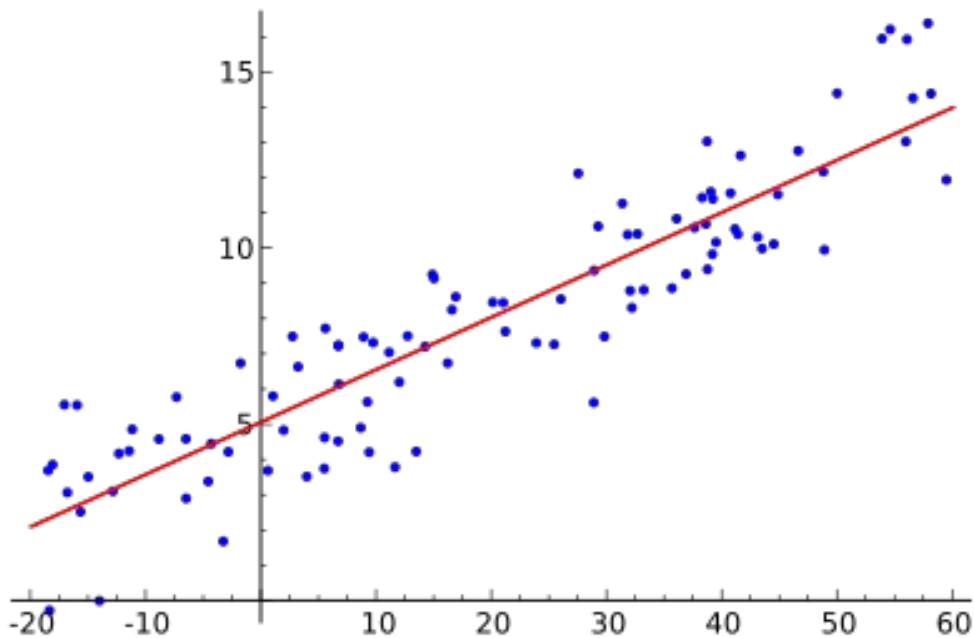
$$r_i = \frac{e_i}{s_e}$$



# Assumptions // Independence



Each observation is independent





# Assumptions // Residuals are normally distributed

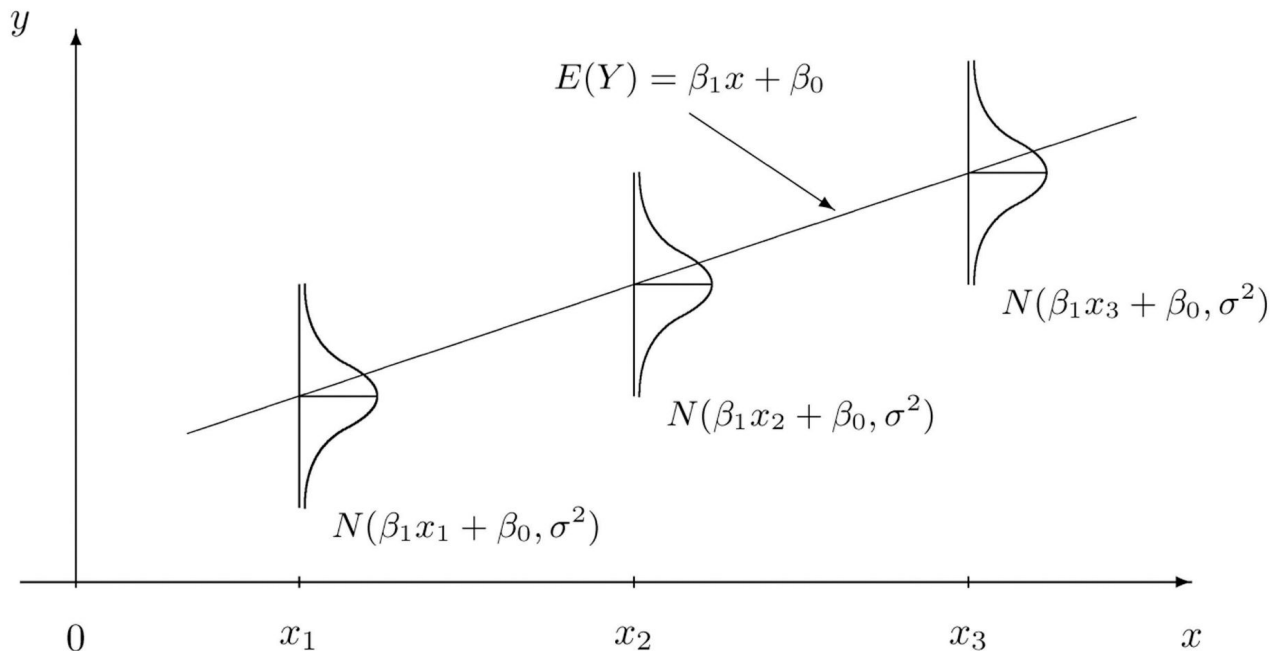


Figure from Tammy Lee's slides

# Assumptions // Residuals are normally distributed



the quantile-quantile (q-q) plot

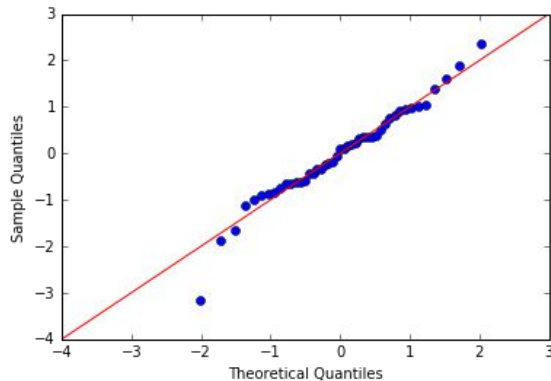
determining if two data sets come from populations with a common distribution.

plot of the quantiles of the first data set against the quantiles of the second data set.

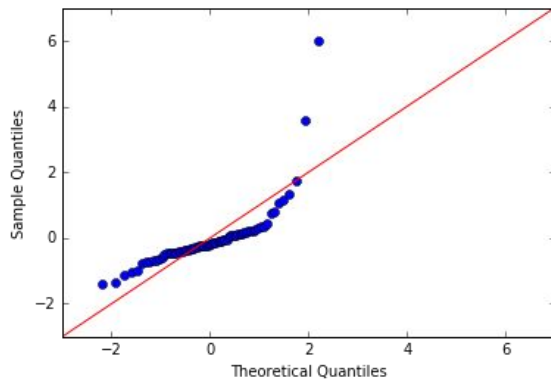
in our case, the second data set is based on the normal pdf

Qqplots should align on a 45-degree reference line

```
statsmodels.graphics.gofplots.qqplot(residuals  
, dist='norm', line='45', fit=True)
```



OK



NOT OK

# Assumptions // Homoscedasticity of residuals



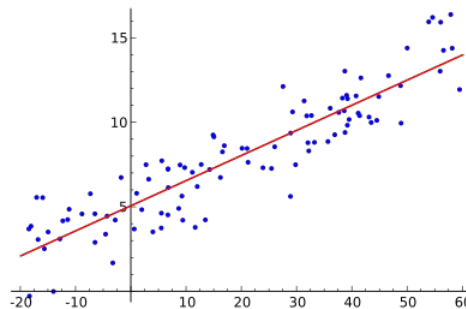
Same Variance, Constant Variance

$$y = X\hat{\beta} + \epsilon$$

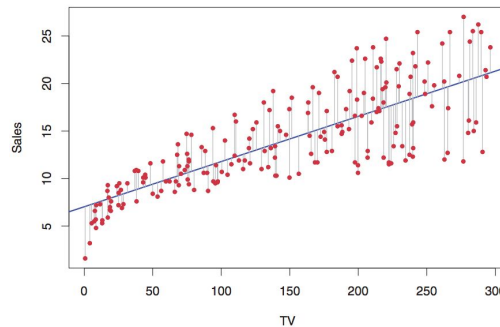
the error term is the same  
across all values of the independent variables

## TROUBLESHOOTING

Using the log of the response-y ?



OK



NOT OK

# Assumption // NO multicollinearity



We assume that there is

no underlying linear relationship

between independent variables

How to challenge that ?

**Solution 1 : correlation matrix**

	DJIA	S&P 500	Nasdaq	Canada	Mexico	Brazil	Stoxx 50	FTSE 100	CAC 40	DAX	IBEX	Italy	Netherlands	Sweden	Switzerland	Nikkei	Hang Seng	Australia
DJIA	1.00	0.97	0.85	0.57	0.56	0.52	0.52	0.48	0.51	0.56	0.49	0.50	0.50	0.42	0.42	0.09	0.11	0.05
S&P 500	0.97	1.00	0.91	0.62	0.58	0.55	0.50	0.47	0.50	0.55	0.48	0.50	0.49	0.41	0.41	0.09	0.11	0.05
Nasdaq	0.85	0.91	1.00	0.58	0.56	0.52	0.48	0.43	0.48	0.54	0.47	0.48	0.48	0.42	0.38	0.14	0.16	0.07
Canada	0.57	0.62	0.58	1.00	0.53	0.53	0.42	0.45	0.41	0.41	0.42	0.42	0.39	0.37	0.35	0.17	0.22	0.17
Mexico	0.56	0.58	0.56	0.53	1.00	0.56	0.42	0.42	0.44	0.43	0.43	0.44	0.39	0.38	0.38	0.17	0.25	0.17
Brazil	0.52	0.55	0.52	0.53	0.56	1.00	0.33	0.35	0.32	0.34	0.34	0.34	0.29	0.30	0.28	0.17	0.22	0.15
Stoxx 50	0.52	0.50	0.48	0.42	0.42	0.33	1.00	0.92	0.94	0.89	0.87	0.88	0.92	0.78	0.86	0.26	0.30	0.24
FTSE 100	0.48	0.47	0.43	0.45	0.42	0.35	0.92	1.00	0.86	0.80	0.80	0.82	0.84	0.73	0.78	0.26	0.30	0.26
CAC 40	0.51	0.50	0.48	0.41	0.44	0.32	0.94	0.86	1.00	0.89	0.88	0.89	0.92	0.78	0.84	0.28	0.32	0.25
DAX	0.56	0.55	0.54	0.41	0.43	0.34	0.89	0.80	0.89	1.00	0.83	0.84	0.86	0.75	0.77	0.26	0.29	0.21
IBEX	0.49	0.48	0.47	0.42	0.43	0.34	0.87	0.80	0.88	0.83	1.00	0.84	0.83	0.75	0.77	0.27	0.32	0.26
Italy	0.50	0.50	0.48	0.42	0.44	0.34	0.88	0.82	0.89	0.84	0.84	1.00	0.85	0.74	0.78	0.24	0.29	0.23
Netherlands	0.50	0.49	0.48	0.39	0.39	0.29	0.92	0.84	0.92	0.86	0.83	0.85	1.00	0.75	0.82	0.27	0.30	0.23
Sweden	0.42	0.41	0.42	0.37	0.38	0.30	0.78	0.73	0.78	0.75	0.75	0.74	0.75	1.00	0.75	0.29	0.33	0.27
Switzerland	0.42	0.41	0.38	0.35	0.38	0.28	0.86	0.78	0.84	0.77	0.77	0.78	0.82	0.75	1.00	0.29	0.32	0.29
Nikkei	0.09	0.09	0.14	0.17	0.17	0.17	0.26	0.26	0.28	0.26	0.27	0.24	0.27	0.29	0.29	1.00	0.52	0.49
Hang Seng	0.11	0.11	0.16	0.22	0.25	0.22	0.30	0.30	0.32	0.29	0.32	0.29	0.30	0.33	0.32	0.52	1.00	0.48
Australia	0.07	0.05	0.07	0.17	0.17	0.15	0.24	0.26	0.25	0.21	0.26	0.23	0.23	0.27	0.29	0.49	0.48	1.00

# Assumption // NO multicollinearity



We assume that there is

no underlying linear relationship

between independent variables

**How to challenge that ?**

**Solution 2 : ???**

**We have an algorithm that and its name is...**

**Linear Regression !**

$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + \dots + \alpha_k X_k + c_0 + e$$

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

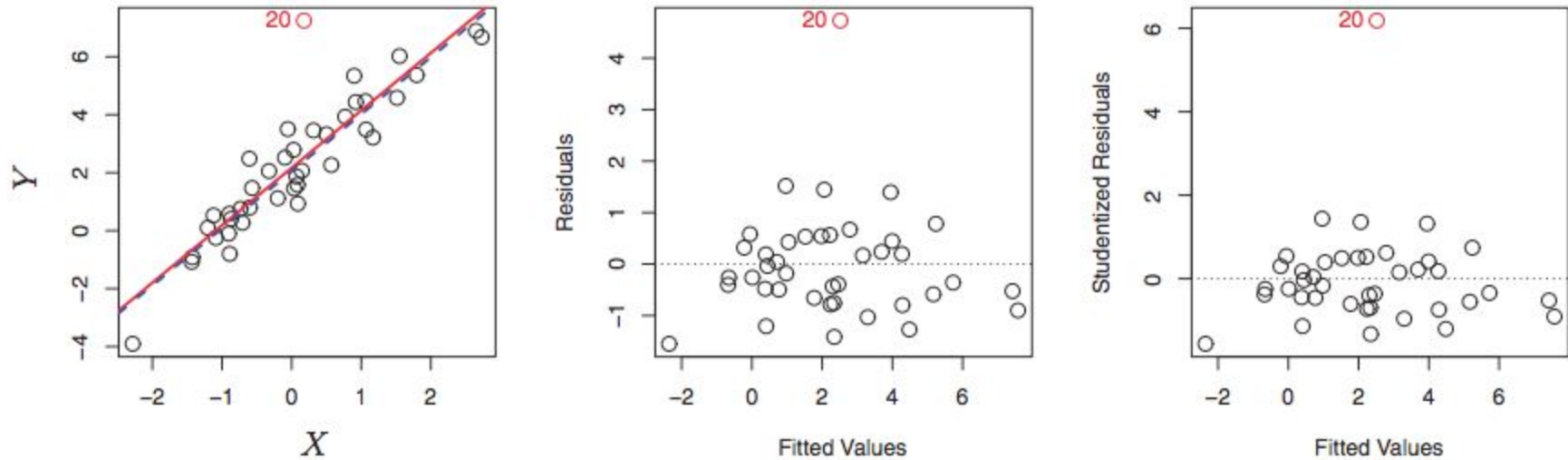
**Rule of thumb, if VIF > 10, problem !**

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
variance_inflation_factor(x.values, index)
```

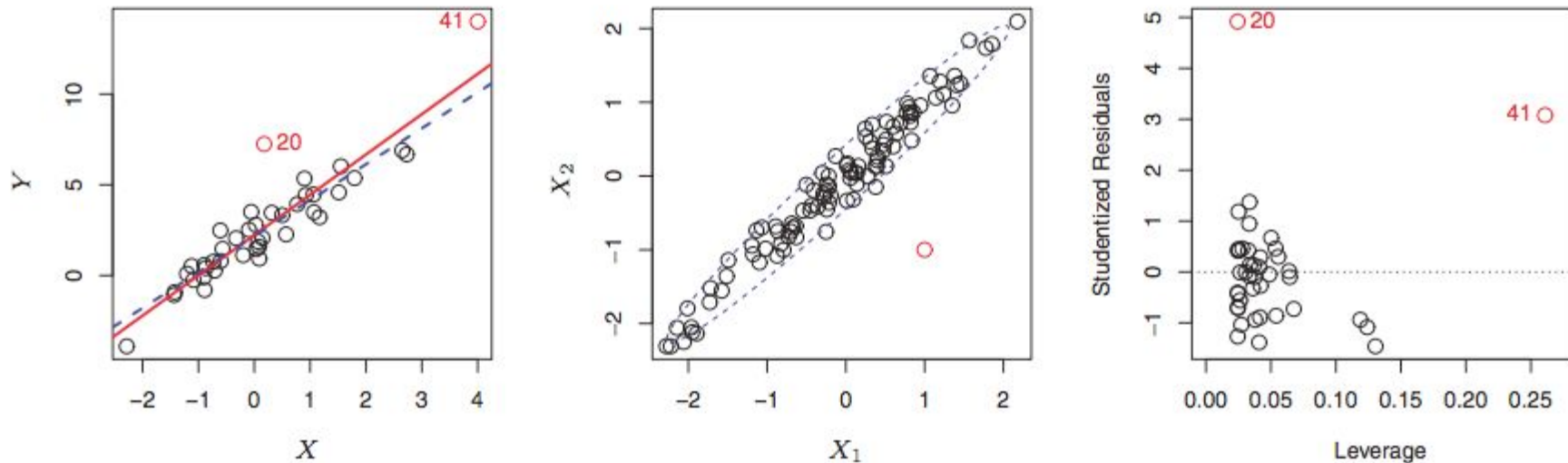


# Outliers

# What is an outlier ?



# Leverage of data points



To measure influence of an outlier:

- Compute the 'hat matrix':  $H = X^T(X^T X)^{-1}X$
- $i$ -th element on the diagonal,  $h_{ii} \equiv (H)_{ii}$ , is  $i$ -th feature's 'leverage'
- An observation with a large residual may not have a lot of influence



# Linear Regression

DSI SEA5, jf.omhover, Sep 21, 2016

## STANDARDS

- **Describe**, **interpret**, and **visualize** the model form of linear regression:  $Y = B_0 + B_1X_1 + B_2X_2 + \dots$
- **Relate** Beta vector solution of Ordinary Least Squares to the cost function (residual sum of squares)
- **State** and troubleshoot the assumptions of linear regression model
- **Perform** OLS with statsmodels and interpret the output: Beta coefficients, p-values,  $R^2$
- How can one **detect** outliers?





Useful snippets (soon on slack)

# Ordinary Least Squares fit



```
import statsmodels
```

```
y = prestige['prestige']
```

```
x = prestige[['income', 'education']].astype(float)
```

```
x['const'] = 1
```

```
prestige_model = statsmodels.api.OLS(endog=y, exog=x).fit()
```

```
prestige_model.summary()
```

OLS Regression Results

<b>Dep. Variable:</b>	prestige	<b>R-squared:</b>	0.828
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.820
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	101.2
<b>Date:</b>	Tue, 20 Sep 2016	<b>Prob (F-statistic):</b>	8.65e-17
<b>Time:</b>	15:00:41	<b>Log-Likelihood:</b>	-178.98
<b>No. Observations:</b>	45	<b>AIC:</b>	364.0
<b>Df Residuals:</b>	42	<b>BIC:</b>	369.4
<b>Df Model:</b>	2		
<b>Covariance Type:</b>	nonrobust		

	coef	std err	t	P> t	[95.0% Conf. Int.]
<b>income</b>	0.5987	0.120	5.003	0.000	0.357 0.840
<b>education</b>	0.5458	0.098	5.555	0.000	0.348 0.744
<b>const</b>	-6.0647	4.272	-1.420	0.163	-14.686 2.556

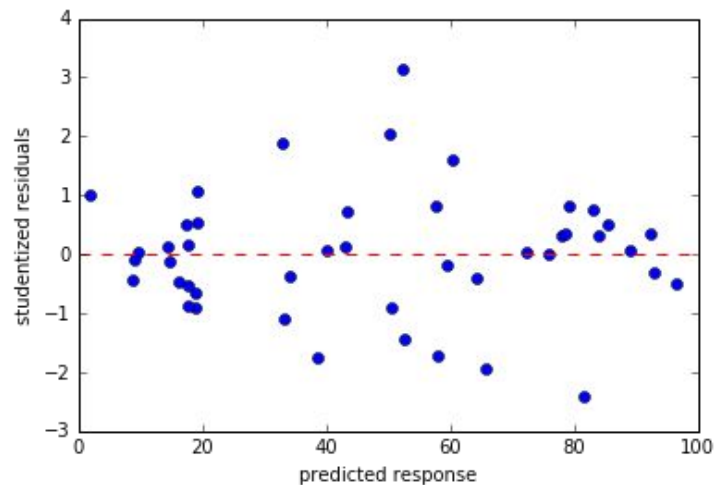
<b>Omnibus:</b>	1.279	<b>Durbin-Watson:</b>	1.458
<b>Prob(Omnibus):</b>	0.528	<b>Jarque-Bera (JB):</b>	0.520
<b>Skew:</b>	0.155	<b>Prob(JB):</b>	0.771
<b>Kurtosis:</b>	3.426	<b>Cond. No.</b>	163.

# Studentized residual plot



```
stdresids = prestige_model.outlier_test()['student_resid']
```

```
plt.plot(prestige_model.fittedvalues, resid, 'o')  
plt.xlabel('predicted response')  
plt.ylabel('studentized residuals')  
plt.axhline(0, c='r', linestyle = '--')
```



# QQPlot



```
statsmodels.graphics.gofplots.qqplot(residuals, dist=norm, line='45', fit=True)
```

