Linear Regression II

Presented by David John Baker
November 2019

FLATIRON SCHOOL

Why Linear Regression II ?

- Today we take second pass at Linear Regression
- Goals: Have better understanding of core components of model
- Procede through entire checklist of running linear regression model
- Clearly define more specific terms and explain how they relate to NHST

Linear Regression

- Can you predict X given Y if we assume a linear relationship between the variables in question?
- Dependant variable is continuous
- Independent variables can be either continuous or categorical

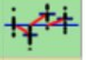
Format

- Assumptions
- Fitting
- Multiple Models
- Interpretation

Common statistical tests are linear models

Last updated: 29 June, 2019. Also check out the R version!

See worked examples and more details at the accompanying notebook: https://github.com/eigenfoo/tests-as-linear

	Common name	Function in scipy.stats	Equivalent linear model in smf.ols	Exact?	The linear model in words	Icon
Simple Regression: (y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	scipy.stats.ttest_1samp(y) scipy.stats.wilcoxon(y)	smf.ols("y ~ 1", data) smf.ols("y ~ 1", signed_rank(data))	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	scipy.stats.ttest_rel(y1, y2) scipy.stats.wilcoxon(y1, y2)	smf.ols("y2_sub_y1 ~ 1", data) smf.ols("y2_sub_y1 ~ 1", signed_rank(data))	✓ for N >14	One intercept predicts the pairwise y_2-y_1 differences. - (Same, but it predicts the *signed rank* of y_2-y_1.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	scipy.stats.pearsonr(x, y) scipy.stats.spearmanr(x, y)	smf.ols("y ~ 1 + x", data) smf.ols("y ~ 1 + x", rank(data))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	scipy.stats.ttest_ind(y1, y2) N/A in Python, but see R version scipy.stats.mannwhitneyu(y1, y1)	smf.ols("y ~ 1 + group", data)[A] N/A in Python, but see R version smf.ols("y ~ 1 + group", signed_rank(data))[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: (y ~ 1 + x_1 + x_2 +...)	P: One-way ANOVA N: Kruskal-Wallis	scipy.stats.f_oneway(a, b, c) scipy.stats.kruskal(a, b, c)	smf.ols(y ~ 1 + G_2 + G_3 +...+ G_N)[A] smf.ols(rank(y) ~ 1 + G_2 + G_3 +...+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	N/A in Python, but see R version	smf.ols("y ~ 1 + G_2 + G_3 +...+ G_N + x", data)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	N/A in Python, but see R version	smf.ols("y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2+G_3*S_3+...+G_N*S_K", data)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: **$G_{2 \text{ to } N}$** is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for **$S_{2 \text{ to } K}$** for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	scipy.stats.chisquare(data)	**Equivalent log-linear model** sm.GLM(y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2+G_3*S_3+...+G_N*S_K, family=...)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments:* `glm(model, family=poisson())` *As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	scipy.stats.chi2_contingency(data)	sm.GLM(y ~ 1 + G_2 + G_3 +...+ G_N, family=...)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank(df) = np.sign(df) * df.rank()`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://eigenfoo.xyz/tests-as-linear/.

[A] See the note to the two-way ANOVA for explanation of the notation.

Jonas Kristoffer Lindeløv, George Ho
https://lindeloev.net, https://eigenfoo.xyz

https://eigenfoo.xyz/tests-as-linear/

Basic Linear Regression Assumptions

1. Independence of Data Points
2. Normality of Residuals
3. Homoscedasticity

Linear Regression Assumptions

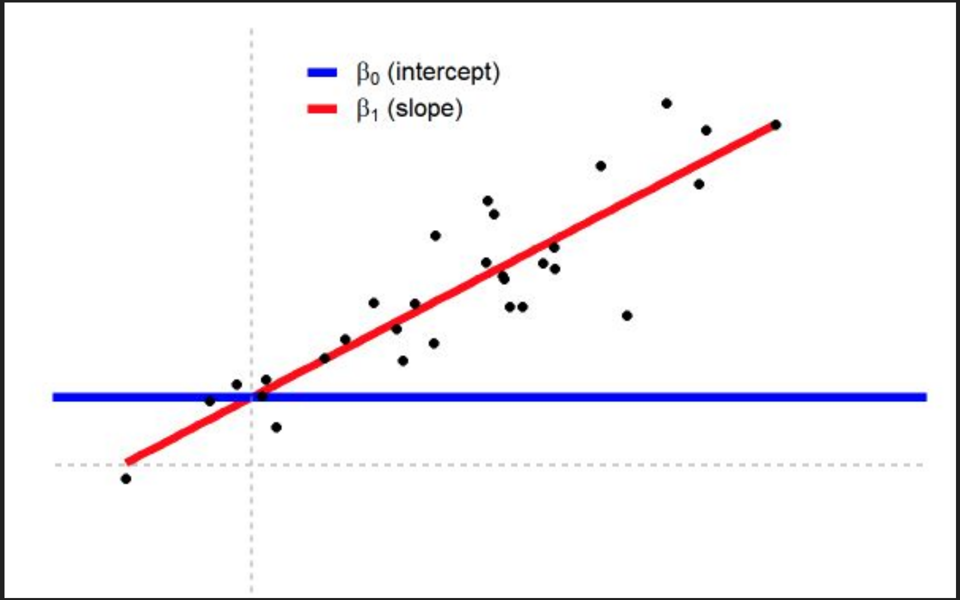
(0. Linearity)

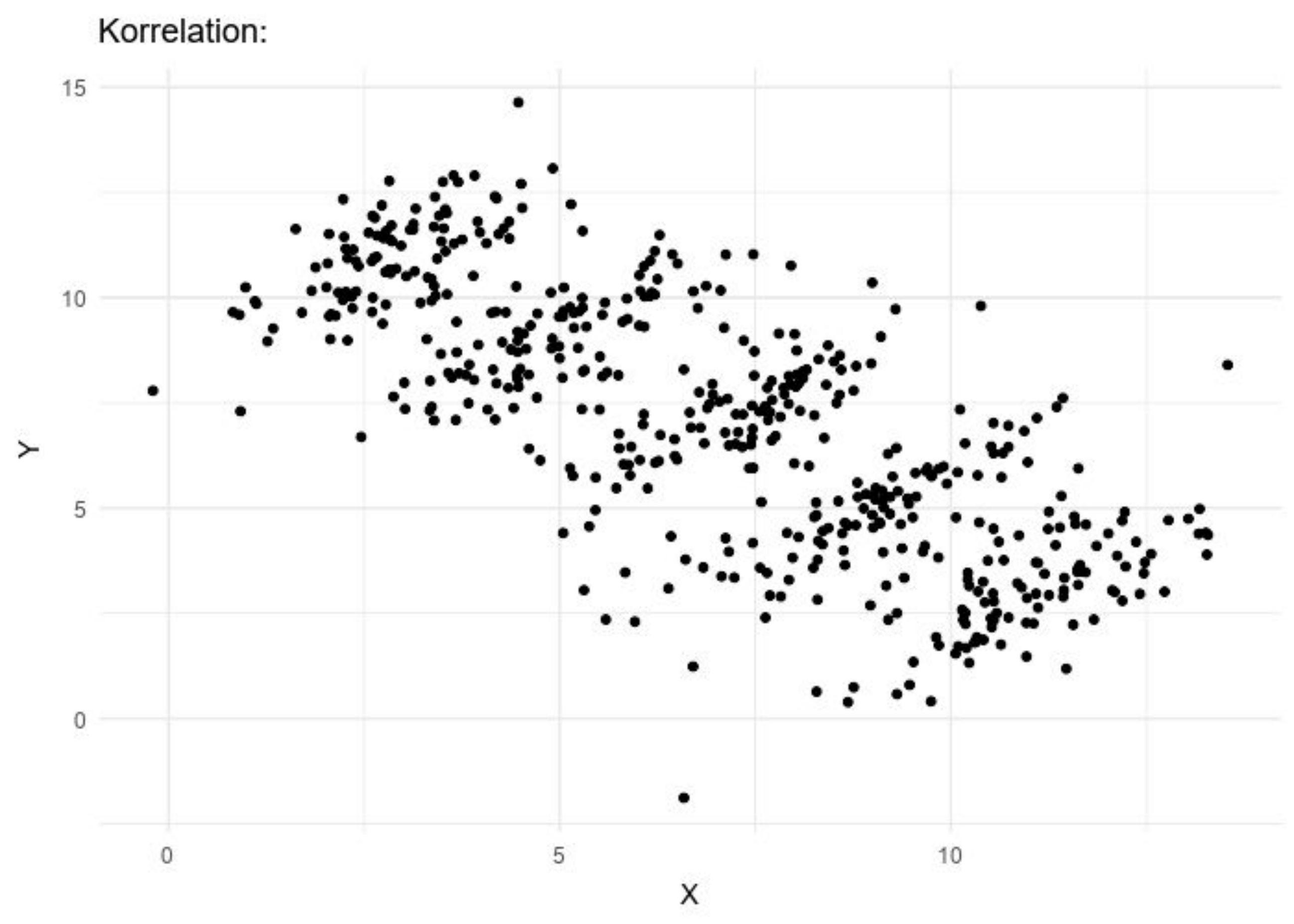
1. Independence of Data Points
2. Normality of Residuals
3. Homoscedasticity
4. Multicollinearity
5. You DO NOT need Normally distributed variables, you DO NEED to Run Diagnostic Plots

Independence of Data Points

- Each data point needs to be independently sampled from the parent population

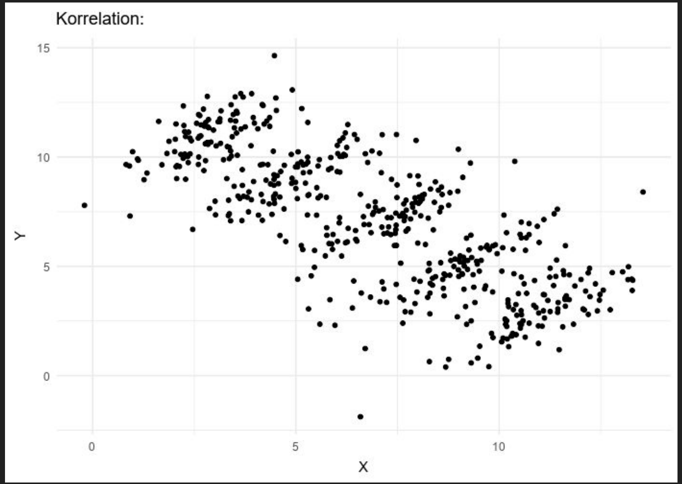
- AKA Data points shouldn't be able to influence/talk to one another!!!





https://en.wikipedia.org/wiki/Simpson%27s_paradox#/media/File:Simpsons_paradox_-_animation.gif

Think of situation in your area of expertise where you might come across a violation of the the assumption of independence.



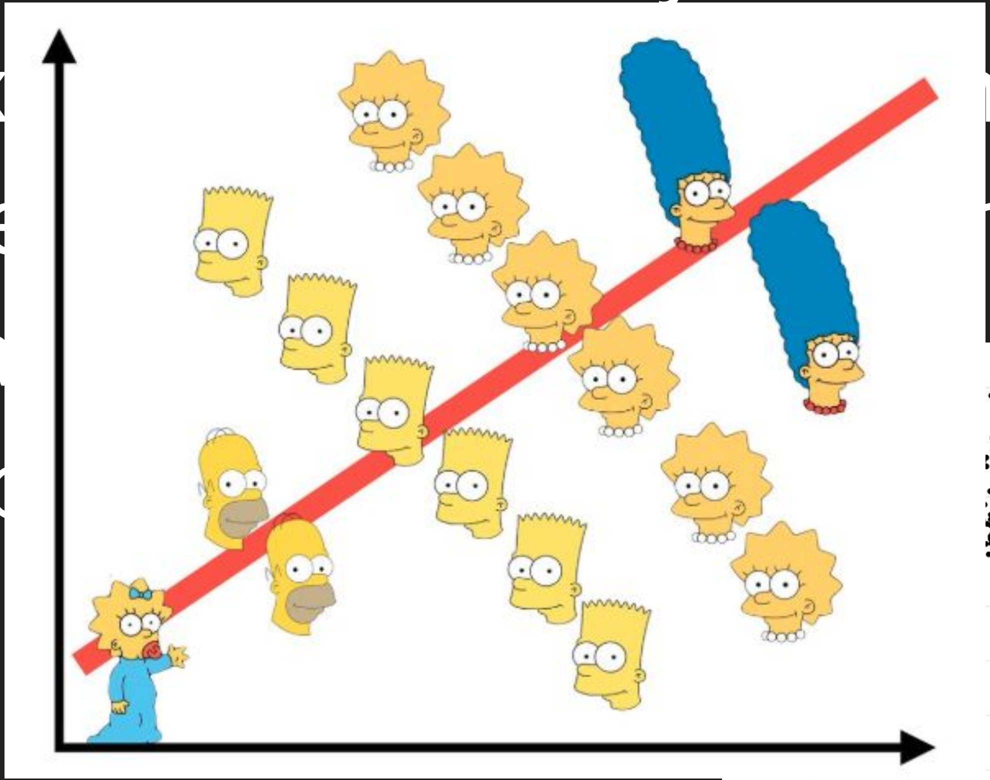
Think of situation in your area

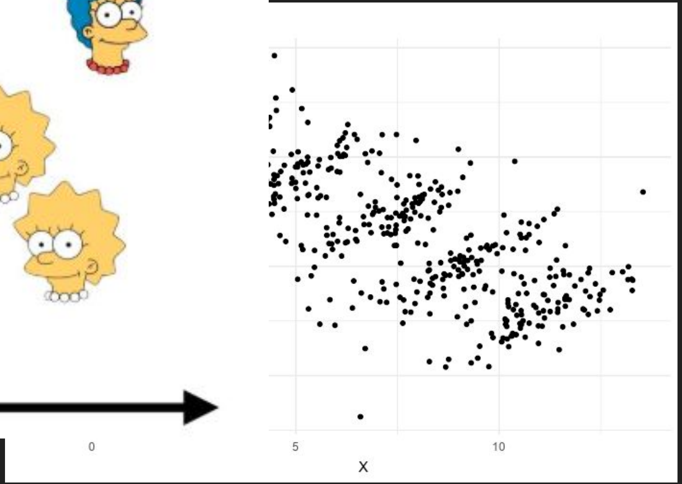
of ex... nt

come... ne

the a...

indep...

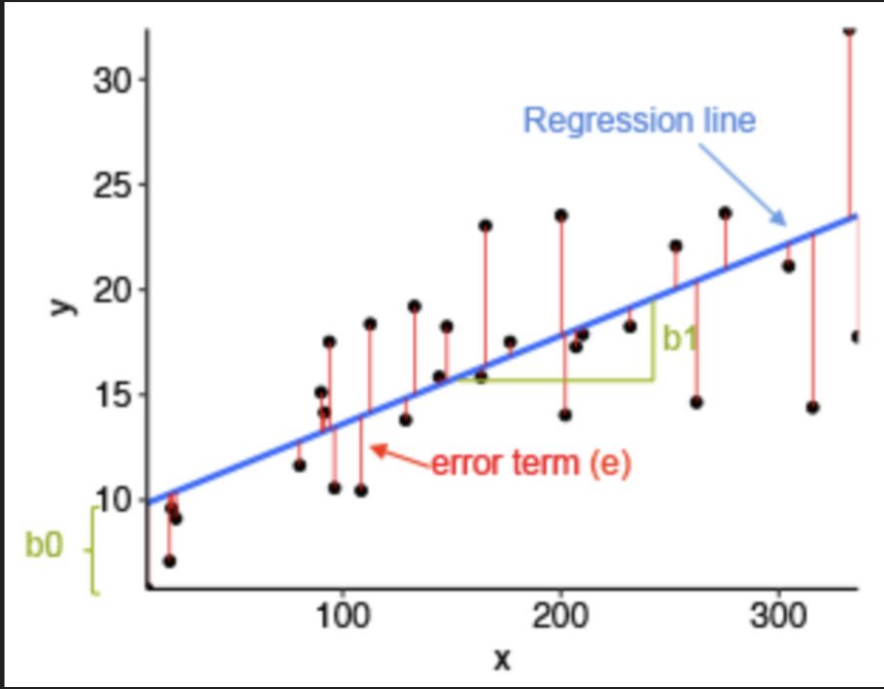




Normally Distributed Residuals

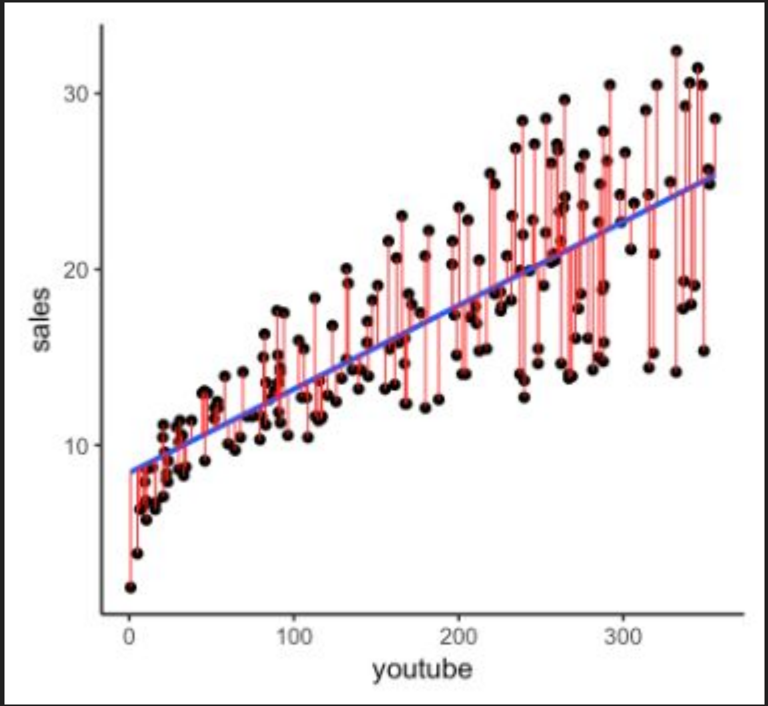
Residuals, or the distances between your regression line and observed data, when plotted should create a normal distribution

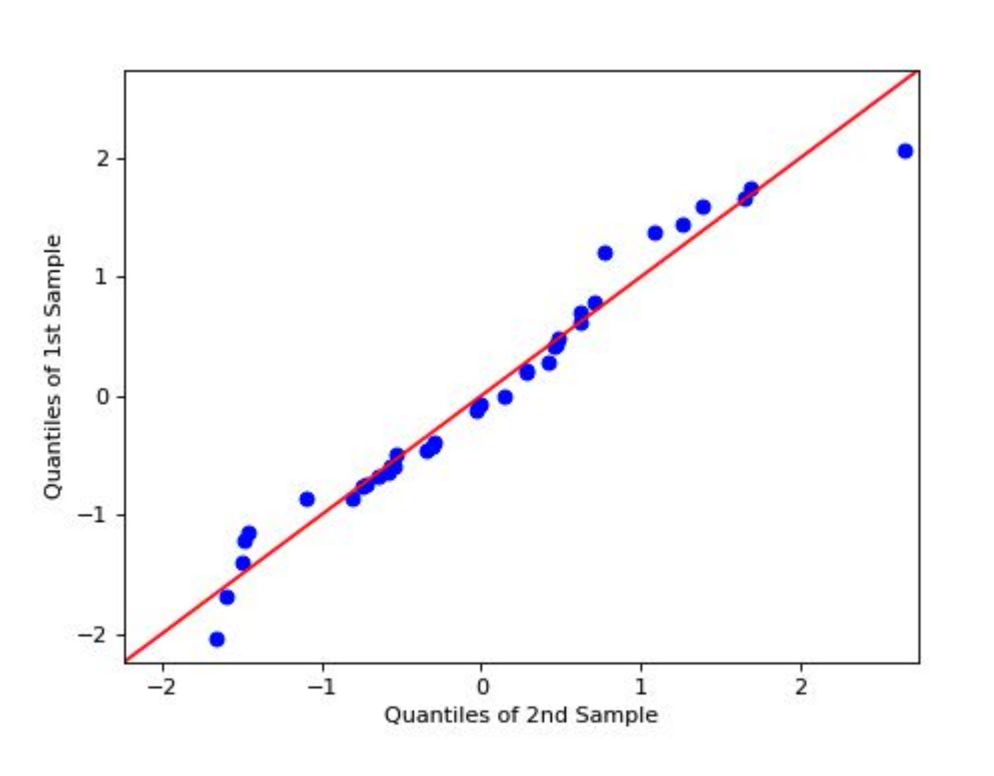
Inspect this with a QQ plot



http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

Normally Distributed Residuals





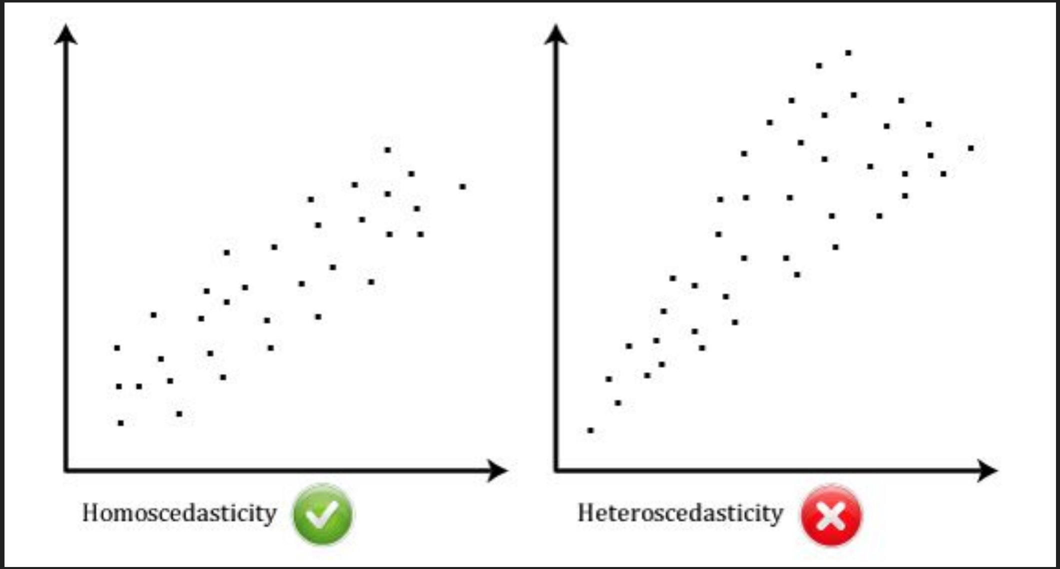
Homoscedasticity

Homoscedasticity (ho-mo-ske-das-ti-ci-tee)

Heteroscedasticity (het-er-o-ske-das-tis-i-tee)

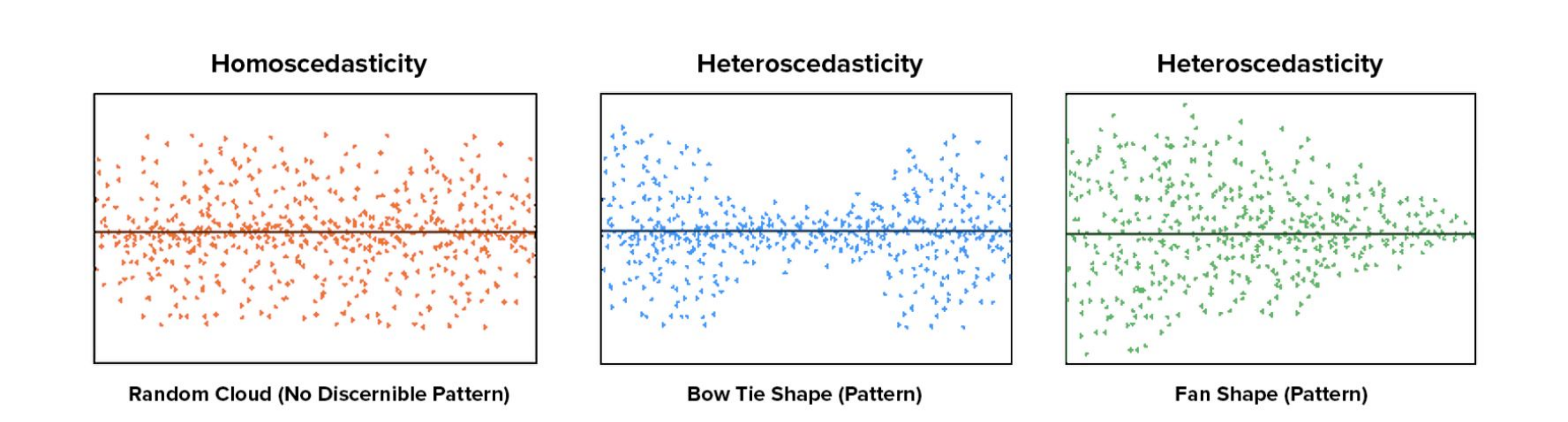
Homoscedasticity (ho-mo-ske-das-ti-ci-tee)
Homogeneity of Variance

- Variance around errors is uniform



https://stats.stackexchange.com/questions/76151/what-is-an-intuitive-explanation-of-why-we-want-homoskedasticity-in-a-regression

Residual Plots



https://clevertap.com/blog/a-brief-primer-on-linear-regression-part-ii/

Linear Regression Checklist

- Plot all variables
- Check for Multicollinearity
- Check for Outliers (Univariate and Multivariate)

Regression Output

Out[12]:

OLS Regression Results

Dep. Variable:	Sepal.Length	**R-squared:**	0.867
Model:	OLS	**Adj. R-squared:**	0.862
Method:	Least Squares	**F-statistic:**	155.8
Date:	Mon, 18 Nov 2019	**Prob (F-statistic):**	3.86e-60
Time:	11:57:26	**Log-Likelihood:**	-32.558
No. Observations:	150	**AIC:**	79.12
Df Residuals:	143	**BIC:**	100.2
Df Model:	6		
Covariance Type:	nonrobust		

Regression Output

R-Squared:

Coefficient of

Determination

0 -- 1

Variance explained

Literally, r, squared

//

Out[12]:

OLS Regression Results

Dep. Variable:	Sepal.Length	**R-squared:**	0.867
Model:	OLS	**Adj. R-squared:**	0.862
Method:	Least Squares	**F-statistic:**	155.8
Date:	Mon, 18 Nov 2019	**Prob (F-statistic):**	3.86e-60
Time:	11:57:26	**Log-Likelihood:**	-32.558
No. Observations:	150	**AIC:**	79.12
Df Residuals:	143	**BIC:**	100.2
Df Model:	6		
Covariance Type:	nonrobust		

https://rpsychologist.com/d3/correlation/

Regression Output

F statistic:

Omnibus test,

Just like ANOVA

Out[12]:

OLS Regression Results

Dep. Variable:	Sepal.Length	**R-squared:**	0.867
Model:	OLS	**Adj. R-squared:**	0.862
Method:	Least Squares	**F-statistic:**	155.8
Date:	Mon, 18 Nov 2019	**Prob (F-statistic):**	3.86e-60
Time:	11:57:26	**Log-Likelihood:**	-32.558
No. Observations:	150	**AIC:**	79.12
Df Residuals:	143	**BIC:**	100.2
Df Model:	6		
Covariance Type:	nonrobust		

Variance / Error

Regression Output

Model Fit Comparison Metrics

In statistics, the **likelihood function** (often simply called the **likelihood**) expresses how likely particular values of statistical model parameters are for a given sample of data.[a]

Out[12]:

sults

	Sepal.Length	**R-squared:**	0.867
	OLS	**Adj. R-squared:**	0.862
	Least Squares	**F-statistic:**	155.8
Date:	Mon, 18 Nov 2019	**Prob (F-statistic):**	3.86e-60
Time:	11:57:26	**Log-Likelihood:**	-32.558
No. Observations:	150	**AIC:**	79.12
Df Residuals:	143	**BIC:**	100.2
Df Model:	6		
Covariance Type:	nonrobust		

Regression Output

Model Fit Comparison Metrics

In statistics, the **likelihood function** (often simply called the **likelihood**) expresses how likely particular values of statistical model parameters are for a given sample of data.[a]

The **Akaike information criterion** (**AIC**) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data.[1][2] Given a

In statistics, the **Bayesian information criterion** (**BIC**) or **Schwarz information criterion** (also **SIC**, **SBC**, **SBIC**) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

```
Out[12]:
```

sults

Sepal.Length	**R-squared:**	0.867
OLS	**Adj. R-squared:**	0.862
Least Squares	**F-statistic:**	155.8
Mon, 18 Nov 2019	**Prob (F-statistic):**	3.86e-60
11:57:26	**Log-Likelihood:**	-32.558
150	**AIC:**	79.12
143	**BIC:**	100.2
6		
nonrobust		

Regression Output -- Unstandardized Beta

For every change in the indicator variable, you have y changes in the dependant variable, all others equal

	coef	td err	t	P>|t|	[0.025	0.975]
Unnamed: 0	-3.606e-05	0.002	-0.020	0.984	-0.004	0.003
Sepal.Width	0.4960	0.086	5.737	0.000	0.325	0.667
Petal.Length	0.8290	0.070	11.888	0.000	0.691	0.967
Petal.Width	-0.3150	0.152	-2.075	0.040	-0.615	-0.015
Species_setosa	2.1722	0.285	7.628	0.000	1.609	2.735
Species_versicolor	1.4511	0.327	4.443	0.000	0.805	2.097
Species_virginica	1.1531	0.441	2.614	0.010	0.281	2.025

Regression Output -- Standardized Beta

Everything on z score

Need to standardize
Variables ahead
Of time!!!

	coef	std err	t	P>|t|	[0.025	0.975]
Unnamed: 0	[illegible]	0.002	-0.020	0.984	-0.004	0.003
Sepal.Width	[illegible]	0.086	5.737	0.000	0.325	0.667
Petal.Length	[illegible]	0.070	11.888	0.000	0.691	0.967
Petal.Width	[illegible]	0.152	-2.075	0.040	-0.615	-0.015
Species_setosa	2.1722	0.285	7.628	0.000	1.609	2.735
Species_versicolor	1.4511	[illegible]	4.443	0.000	0.805	2.097
Species_virginica	1.1531	[illegible]	2.614	0.010	0.281	2.025

//

Regression Output

	coef	std err	t	P>|t|	[0.025	0.975]
Unnamed: 0	-3.606e-05	0.002	-0.020	0.984	-0.004	0.003
Sepal.Width	0.4960	0.086	5.737	0.000	0.325	0.667
Petal.Length	0.8290	0.070	11.888	0.000	0.691	0.967
Petal.Width	-0.3150	0.152	-2.075	0.040	-0.615	-0.015
Species_setosa	2.1722	0.285	7.628	0.000	1.609	2.735
Species_versicolor	1.4511	0.327	4.443	0.000	0.805	2.097
Species_virginica	1.1531	0.441	2.614	0.010	0.281	2.025

Regression Output

	coef	std err	t	P>|t|	[0.025	0.975]
Unnamed: 0	-3.606e-05	0.002	-0.020	0.984	-0.004	0.003
Sepal.Width	0.4960	0.086	5.737	0.000	0.325	0.667
Petal.Length	0.8290	0.070	11.888	0.000	0.691	0.967
Petal.Width	-0.3150	0.152	-2.075	0.040	-0.615	-0.015
Species_setosa	2.1722	0.285	7.628	0.000	1.609	2.735
Species_versicolor	1.4511	0.327	4.443	0.000	0.805	2.097
Species_virginica	1.1531	0.441	2.614	0.010	0.281	2.025

Regression Output

	coef	std err	t	P>|t|	[0.025	0.975]
Unnamed: 0	-3.606e-05	0.002	-0.020	0.984	-0.004	0.003
Sepal.Width	0.4960	0.086	5.737	0.000	0.325	0.667
Petal.Length	0.8290	0.070	11.888	0.000	0.691	0.967
Petal.Width	-0.3150	0.152	-2.075	0.040	-0.615	-0.015
Species_setosa	2.1722	0.285	7.628	0.000	1.609	2.735
Species_versicolor	1.4511	0.327	4.443	0.000	0.805	2.097
Species_virginica	1.1531	0.441	2.614	0.010	0.281	2.025

Regression Output

	coef	std err	t	P>|t|	[0.025	0.975]
Unnamed: 0	-3.606e-05	0.002	-0.020	0.984	-0.004	0.003
Sepal.Width	0.4960	0.086	5.737	0.000	0.325	0.667
Petal.Length	0.8290	0.070	11.888	0.000	0.691	0.967
Petal.Width	-0.3150	0.152	-2.075	0.040	-0.615	-0.015
Species_setosa	2.1722	0.285	7.628	0.000	1.609	2.735
Species_versicolor	1.4511	0.327	4.443	0.000	0.805	2.097
Species_virginica	1.1531	0.441	2.614	0.010	0.281	2.025

Regression Output

Omnibus:	0.414	**Durbin-Watson:**	1.966
Prob(Omnibus):	0.813	**Jarque-Bera (JB):**	0.567
Skew:	-0.060	**Prob(JB):**	0.753
Kurtosis:	2.723	**Cond. No.**	1.98e+03

In statistics, the **Durbin–Watson statistic** is a test statistic used to detect the presence of autocorrelation at lag 1 in the residuals (prediction errors) from a regression analysis. It is named after James Durbin and Geoffrey Watson. The small

Regression Output

Omnibus:	0.414	**Durbin-Watson:**	1.966
Prob(Omnibus):	0.813	**Jarque-Bera (JB):**	$\boxed{0.567}$
Skew:	-0.060	**Prob(JB):**	0.753
Kurtosis:	2.723	**Cond. No.**	1.98e+03

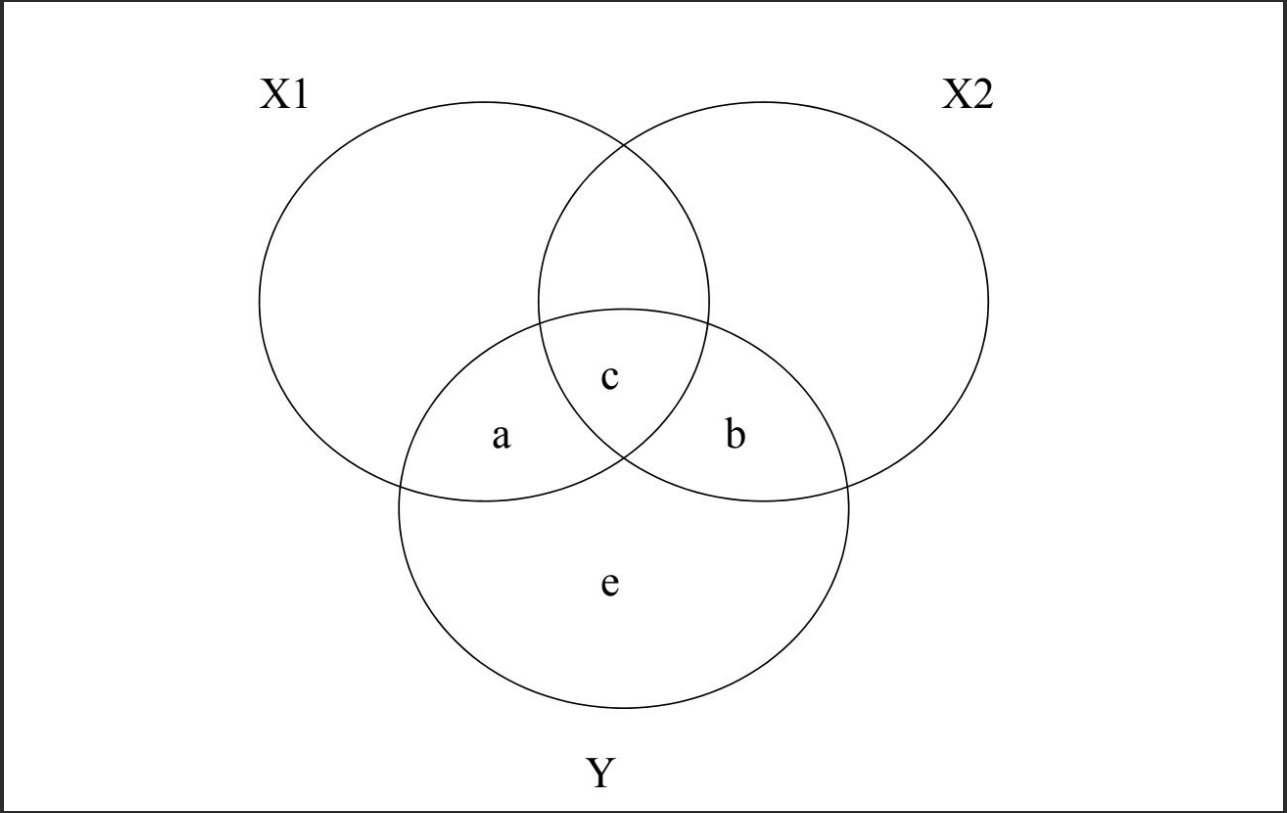
In statistics, the **Jarque–Bera test** is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The test is named after Carlos Jarque and Anil K. Bera. The test statistic is always nonnegative. If it is far from zero, it signals the data do not have a normal distribution.

Regression Output

Omnibus:	0.414	**Durbin-Watson:**	1.966
Prob(Omnibus):	0.813	**Jarque-Bera (JB):**	0.567
Skew:	-0.060	**Prob(JB):**	0.753
Kurtosis:	2.723	**Cond. No.**	1.98e+03

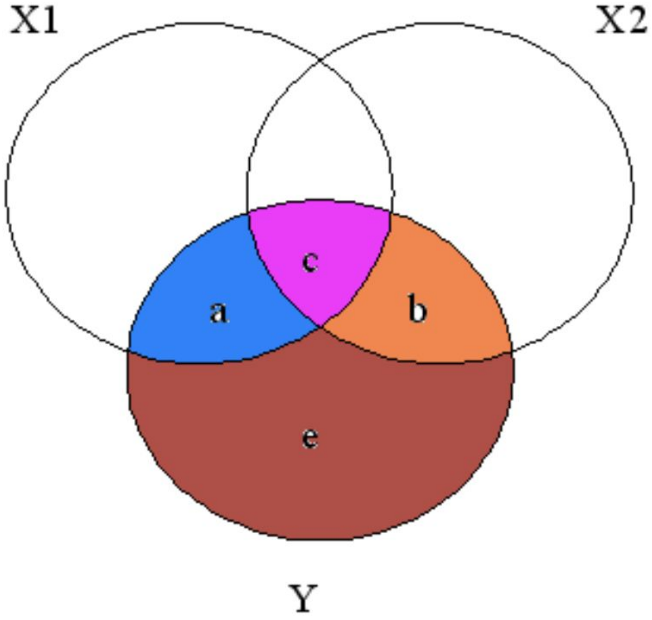
In the field of numerical analysis, the **condition number** of a function measures how much the output value of the function can change for a small change in the input argument. This is used to measure how sensitive a function is to changes or errors in the input, and how much error in the output

Regression Output



Regression Output

Concept of multiple R^2 for the full regression model



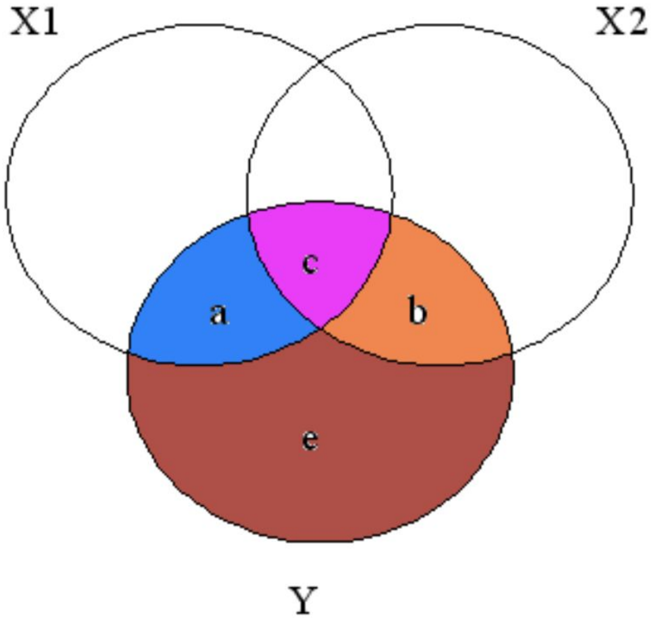
$$r^2_{Y \bullet 12} = \frac{a+b+c}{a+b+c+e=1} = a+b+c$$

What of Y can be accounted for by X1, X2, and any redundancy among X1 and X2?

Area *c* is not unique to either X1 or X2 alone, but contributes to the full R^2

Regression Output

Concept of multiple R^2 for the full regression model



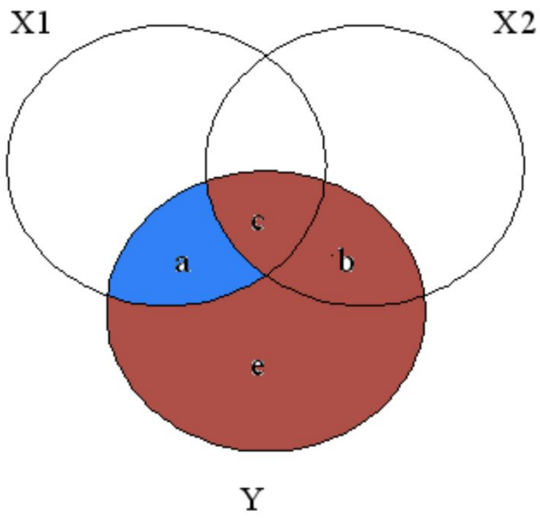
$$r^2_{Y\bullet 12} = \frac{a+b+c}{a+b+c+e=1} = a+b+c$$

What of Y can be accounted for by X1, X2, and any redundancy among X1 and X2?

Area *c* is not unique to either X1 or X2 alone, but contributes to the full R^2

Regression Output

Influence of predictor X1 when X2 is already in the model: variance shared between X1 & Y **beyond** that accounted for by X2



"part" or "semipartial" correlation of X1 with Y

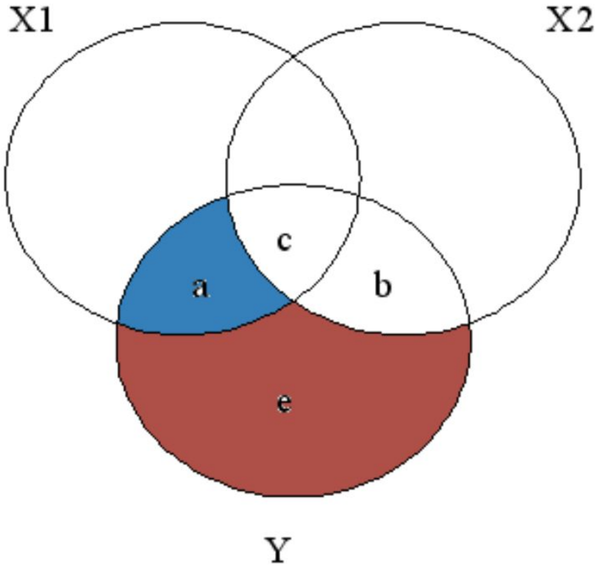
$$r^2_{1(Y\bullet 2)} = \frac{a}{a+b+c+e=1} = a$$

Part correlation: the correlation of Y with that part of X1 that is independent of X2

The squared semipartial correlation is ONE way to determine how much influence (i.e., importance) each predictor has to the full equation

Regression Output

Partial correlation: X1 with Y when X2's overlap with Y AND X1 is controlled



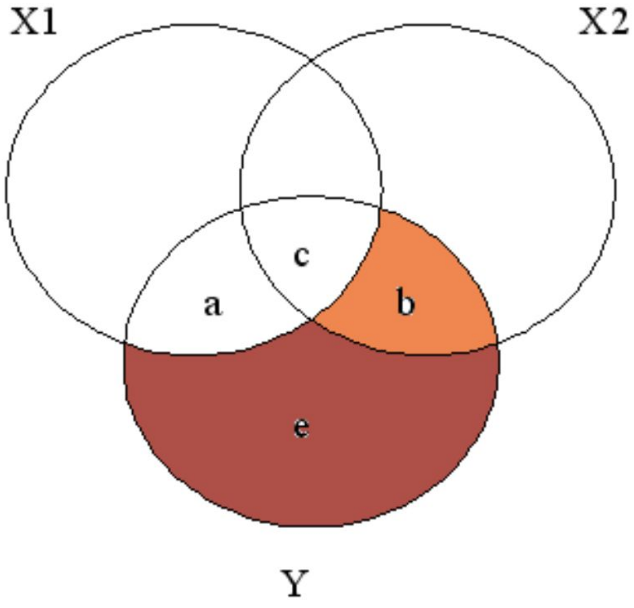
$$r^2_{1Y \bullet 2} = \frac{a}{a+e}$$

Influence of *b* and *c* are removed from consideration altogether

Partial correlation: the correlation between the residual of one relationship (e_{2Y}) and the residual of another (e_{12}) **in the variance of Y**.

Regression Output

Partial correlation: X2 with Y when X1's overlap with Y AND X2 is controlled



$$r^2_{2Y\bullet 1} = \frac{b}{b+e}$$

Influence of *a* and *c* are removed from consideration altogether

Partial correlation: the correlation between the residual of one relationship (e_{1Y}) and the residual of another (e_{12}) **in the variance of Y**.

//