

Sample data links:

<https://www.springboard.com/blog/free-public-data-sets-data-science-project/>

<https://www.kaggle.com/ratman/datasets-for-regression-analysis>

<https://guides.emich.edu/data/free-data>

yelp data

<https://scholars.unh.edu/cgi/viewcontent.cgi?article=1379&context=honors>

[https://www.researchgate.net/publication/259578317\\_Predicting\\_a\\_Business\\_Star\\_in\\_Yelp\\_from\\_Its\\_Reviews\\_Text\\_Alone](https://www.researchgate.net/publication/259578317_Predicting_a_Business_Star_in_Yelp_from_Its_Reviews_Text_Alone)

<https://rpubs.com/JeanReneN/132019>

<http://cs229.stanford.edu/proj2017/final-reports/5244334.pdf>

Regression type:

<https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>

Assumption:

<http://people.duke.edu/~rnau/testing.htm>

Regression with mtcars in R

[https://rstudio-pubs-](https://rstudio-pubs-static.s3.amazonaws.com/111995_0b63653147624f5c9223caf1c1bc0d33.html)

[static.s3.amazonaws.com/111995\\_0b63653147624f5c9223caf1c1bc0d33.html](https://rstudio-pubs-static.s3.amazonaws.com/111995_0b63653147624f5c9223caf1c1bc0d33.html)

<https://rpubs.com/davoodastarak/mtRegression>

Assumption for logistic:

<https://www.statisticssolutions.com/assumptions-of-logistic-regression/>

logistics in R

<https://www.datacamp.com/community/tutorials/logistic-regression-R>

Logistic use case:

<http://ucanalytics.com/blogs/case-study-example-banking-logistic-regression-3/>

Logistic generic:

<http://dataaspirant.com/2017/03/02/how-logistic-regression-model-works/>

Residual:

[https://gerardnico.com/data\\_mining/residual](https://gerardnico.com/data_mining/residual)

Bias – variance:

<https://elitedatascience.com/bias-variance-tradeoff>

<https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/>

Linear regression on Boston Housing data set: (python)

<https://towardsdatascience.com/linear-regression-on-boston-housing-dataset-f409b7e4a155>

<https://blog.goodaudience.com/linear-regression-on-the-boston-housing-data-set-d18c4ce4d0be>

<https://towardsdatascience.com/linear-regression-on-boston-housing-dataset-f409b7e4a155>

<https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9>

<https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9>

<http://ugrad.stat.ubc.ca/R/library/mlbench/html/BostonHousing.html>

<http://ugrad.stat.ubc.ca/R/library/mlbench/html/BostonHousing.html>

**boston housing (R)**

<https://www.kaggle.com/sukeshpabba/linear-regression-with-boston-housing-data>

<https://www.kaggle.com/andyxie/regression-with-r-boston-housing-price>

<https://rpubs.com/sukeshpabba/LR>

**data set:**

<https://www.kaggle.com/datasets>

**Red wine quality :** <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

<https://rpubs.com/jeknov/redwine>

<https://www.kaggle.com/sagarnildass/red-wine-analysis-by-r/report>

[https://rstudio-pubs-](https://rstudio-pubs-static.s3.amazonaws.com/274165_627a87883a534f15b42c4b879d369ac7.html)

[static.s3.amazonaws.com/274165\\_627a87883a534f15b42c4b879d369ac7.html](https://rstudio-pubs-static.s3.amazonaws.com/274165_627a87883a534f15b42c4b879d369ac7.html)

**FIFA player:**

<https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global#FullData.csv>

**UCI dataset:**

<http://mlr.cs.umass.edu/ml/datasets.html>

<https://data.world/uci>

**CA housing data set:**

<https://www.kaggle.com/thawatchai2018/california-housing-dataset>

**fuel consumption data:**

<https://carfueldata.vehicle-certification-agency.gov.uk/downloads/default.aspx>

## Regression assumptions

<https://www.statisticssolutions.com/assumptions-of-linear-regression/>  
<https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/>  
<http://r-statistics.co/Assumptions-of-Linear-Regression.html> (10 assumptions)  
<https://medium.com/datadriveninvestor/linear-regression-assumptions-f2252b8e2912>  
<http://thestatsgeek.com/2013/08/07/assumptions-for-linear-regression/>  
<https://dziganto.github.io/data%20science/linear%20regression/machine%20learning/python/Linear-Regression-101-Assumptions-and-Evaluation/>  
<https://stats.stackexchange.com/questions/362284/what-is-the-need-of-assumptions-in-linear-regression>  
<https://towardsdatascience.com/linear-regression-modeling-and-assumptions-dcd7a201502a>

## Boston Housing data:

<http://ugrad.stat.ubc.ca/R/library/mlbench/html/BostonHousing.html>  
<http://math.furman.edu/~dcs/courses/math47/R/library/mlbench/html/BostonHousing.html>

It's available from both R and Python library

```
from  
sklearn.datasets  
import  
load_boston  
  
boston_dataset = load_boston()
```

```
data(BostonHousing)  
data(BostonHousing2)
```

<http://ugrad.stat.ubc.ca/R/library/mlbench/html/BostonHousing.html>

## data archive directory:

<http://lib.stat.cmu.edu/datasets/>  
<ftp://ftp.ics.uci.edu/pub/machine-learning-databases>

## IQ and Brain size:

[http://lib.stat.cmu.edu/datasets/IQ\\_Brain\\_Size](http://lib.stat.cmu.edu/datasets/IQ_Brain_Size)

## Regression steps:

<https://www.theanalysisfactor.com/13-steps-regression-anova/>

<https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/>  
<https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/>

## **EDA**

<https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>

## **Statistics quote**

<https://stats.stackexchange.com/questions/726/famous-statistical-quotations>

## **cor not caus**

[https://commons.wikimedia.org/wiki/File:Correlation\\_vs\\_causation.png](https://commons.wikimedia.org/wiki/File:Correlation_vs_causation.png)

## **Logistic Regression**

<http://r-statistics.co/Logistic-Regression-With-R.html>

[http://uc-r.github.io/logistic\\_regression](http://uc-r.github.io/logistic_regression)

## **multiple dimension**

[http://reliawiki.org/index.php/Multiple\\_Linear\\_Regression\\_Analysis](http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis)

## **Multivariate**

<https://stats.stackexchange.com/questions/2358/explain-the-difference-between-multiple-regression-and-multivariate-regression>

<https://www.quora.com/What-is-multivariate-regression>

## **Polynomial**

<https://newonlinecourses.science.psu.edu/stat501/node/324/>

## **Logistics**

[https://ml-cheatsheet.readthedocs.io/en/latest/logistic\\_regression.html](https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html)

[https://en.wikipedia.org/wiki/Multinomial\\_logistic\\_regression](https://en.wikipedia.org/wiki/Multinomial_logistic_regression)

## **EDA**

<https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm>

[https://en.wikipedia.org/wiki/Exploratory\\_data\\_analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis)

## **90% cleaning**

<https://medium.com/datadriveninvestor/data-cleaning-for-data-scientist-363fbbf87e5f>

<https://hackernoon.com/data-cleaning-3c3e37f358dc>

## **80%**

## Data cleansing

<http://bretromero.com/data-science-kaggle-walkthrough-cleaning-data/>

## Rule of Thumb for Interpreting corr coefficient

<http://www.parvez-ahammad.org/blog/how-to-interpret-correlation-coefficients>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/>

## Correlation interpretation

<http://oak.ucc.nau.edu/rh232/courses/EP525/Handouts/Correlation%20Coefficient%20Handout%20-%20Hinkle%20et%20al.pdf>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/>

## Significance test

### For correlation

<http://www.opentextbooks.org.hk/ditatopic/9498>

<https://courses.lumenlearning.com/introstats1/chapter/testing-the-significance-of-the-correlation-coefficient/>

<https://www.google.com/search?q=what+is+null+htpotgesis&ie=utf-8&oe=utf-8&client=firefox-b-1-ab>

[https://www.statsdirect.com/help/basics/p\\_values.htm](https://www.statsdirect.com/help/basics/p_values.htm)

<https://en.wikipedia.org/wiki/P-value>

## missing data map

<https://dev.to/tomoyukiaota/visualizing-the-patterns-of-missing-value-occurrence-with-python-46dj>

<https://rpubs.com/sukeshpabba/LR>

## stepwise

### AIC

<https://stats.stackexchange.com/questions/347652/default-steapaic-in-r>

## Python

### REF for backward

[https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html)

[learn.org/stable/modules/generated/sklearn.feature\\_selection.RFE.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html)

<https://stackoverflow.com/questions/49493468/python-equivalent-for-r-steapaic-for-logistic-regression-direction-backwards>

**python REF**

[https://www.programcreek.com/python/example/86795/sklearn.feature\\_selection.RFE](https://www.programcreek.com/python/example/86795/sklearn.feature_selection.RFE)

[https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.RFE.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html)

<https://datascience.stackexchange.com/questions/937/does-scikit-learn-have-forward-selection-stepwise-regression-algorithm>

[https://planspace.org/20150423-forward\\_selection\\_with\\_statsmodels/](https://planspace.org/20150423-forward_selection_with_statsmodels/)

<http://trevor-smith.github.io/stepwise-post/>

**##model summary (multiple regression in python)**

<http://benalexkeen.com/linear-regression-in-python-using-scikit-learn/>

**OLS state models (pyton) vs. R lm**

<https://stats.stackexchange.com/questions/116825/different-output-for-r-lm-and-python-statsmodel-ols-for-linear-regression>

<https://stackoverflow.com/questions/43524756/difference-between-linear-regression-coefficients-between-python-and-r>

**difference between Difference between statsmodel OLS and scikit linear regression**

<https://stats.stackexchange.com/questions/249892/wildly-different-r2-between-statsmodels-linear-regression-and-sklearn-linear>

**Emulating R regression plots in Python**

<https://medium.com/@emredjan/emulating-r-regression-plots-in-python-43741952c034>

<https://medium.com/@emredjan/emulating-r-regression-plots-in-python-43741952c034>

<https://zhiyzuo.github.io/Linear-Regression-Diagnostic-in-Python/>  
<https://zhiyzuo.github.io/Linear-Regression-Diagnostic-in-Python/>

**normality and residual plots in python**

**Regression diagnostics**

<http://www.statsmodels.org/stable/diagnostic.html>

<https://data.library.virginia.edu/diagnostic-plots/>

<https://www.theanalysisfactor.com/linear-models-r-diagnosing-regression-model/>

**bp test for homoscedasticity**

**homoscedasticity**

<https://stats.stackexchange.com/questions/239060/interpretation-of-breusch-pagan-test-bptest-in-r>

**python model diagnostic**

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html>

**Shapiro test in python**

[https://www.statsmodels.org/dev/examples/notebooks/generated/regression\\_diagnostics.html](https://www.statsmodels.org/dev/examples/notebooks/generated/regression_diagnostics.html)

**#multicollinearity test**

**#Farrar Glauber Test**

<https://www.r-bloggers.com/multicollinearity-in-r/>

**python omni test for normality**

<https://pythonfordatascience.org/anova-python/>

**R normality test**

[https://cran.r-project.org/web/packages/olsrr/vignettes/residual\\_diagnostics.html](https://cran.r-project.org/web/packages/olsrr/vignettes/residual_diagnostics.html)

**normality hypothesis testing**

<http://web.space.ship.edu/pgmarr/Geo441/Lectures/Lec%205%20-%20Normality%20Testing.pdf>

[https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera\\_test](https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test)

## JB in Python

[https://www.statsmodels.org/dev/examples/notebooks/generated/regression\\_diagnostics.html](https://www.statsmodels.org/dev/examples/notebooks/generated/regression_diagnostics.html)

<https://pythonfordatascience.org/anova-python/>

## JB in R

<http://r.789695.n4.nabble.com/Diagnostic-Tests-Jarque-Bera-Test-RAMSEY-td819047.html>

## assumption test

<http://people.duke.edu/~rnau/testing.htm>

## ##multicollinearity

## VIF python

[https://etav.github.io/python/vif\\_factor\\_python.html](https://etav.github.io/python/vif_factor_python.html)

## VIF R

[https://cran.r-project.org/web/packages/olsrr/vignettes/regression\\_diagnostics.html](https://cran.r-project.org/web/packages/olsrr/vignettes/regression_diagnostics.html)

## R squared vs. adjusted r squared

[https://www.ibm.com/support/knowledgecenter/en/SSEP7J\\_11.1.0/com.ibm.swg.ba.cognos.ug\\_ca\\_dshb.doc/rsquared\\_adjusted.html](https://www.ibm.com/support/knowledgecenter/en/SSEP7J_11.1.0/com.ibm.swg.ba.cognos.ug_ca_dshb.doc/rsquared_adjusted.html)

<https://datascience.stackexchange.com/questions/14693/what-is-the-difference-of-r-squared-and-adjusted-r-squared>

<https://datascience.stackexchange.com/questions/14693/what-is-the-difference-of-r-squared-and-adjusted-r-squared>

<https://discuss.analyticsvidhya.com/t/difference-between-r-square-and-adjusted-r-square/264/2>

## DW test

<https://stats.stackexchange.com/questions/109234/durbin-watson-test-statistic>

18

In R, the function `durbinWatsonTest()` from `car` package verifies if the residuals from a linear model are correlated or not:

- The null hypothesis ( $H_0$ ) is that there is no correlation among residuals, i.e., they are independent.



- The alternative hypothesis ( $H_a$ ) is that residuals are autocorrelated. As the p value was near from zero it means one can reject the null.

[https://www.statsmodels.org/dev/generated/statsmodels.stats.stattools.durbin\\_watson.html](https://www.statsmodels.org/dev/generated/statsmodels.stats.stattools.durbin_watson.html)

### RFE vs. AIC

<https://discuss.analyticsvidhya.com/t/how-does-the-recursive-feature-elimination-rfe-works-and-how-it-is-different-from-backward-elimination/74199>

<https://www.scikit-yb.org/en/latest/api/features/rfecv.html>

<https://stats.stackexchange.com/questions/109234/durbin-watson-test-statistic>

From this website:

"The Hypotheses for the Durbin Watson test are:  $H_0$  = no first order autocorrelation.  $H_1$  = first order correlation exists.

The Durbin Watson test reports a test statistic, with a value from 0 to 4, where the rule of thumb is:

2 is no autocorrelation.

0 to <2 is positive autocorrelation (common in time series data).

>2 to 4 is negative autocorrelation (less common in time series data).

A rule of thumb is that test statistic values in the range of 1.5 to 2.5 are relatively normal. "

Note that to get a more precise conclusion, we should not just rely on the DW statistic, but rather look at the p-value. Software packages like SAS will give 2 p-values - one for test for positive first order autocorrelation and the second one for the test for negative first order autocorrelation (both p-values add upto 1). If both p-values are more than your selected Alpha (0.05 in most cases), then we can not reject the null hypothesis that "no first order autocorrelation exists.

If any one of the p-values is < 0.05 (or selected Alpha), then we know that the corresponding alternate hypothesis is true (with 1 - Alpha certainty).

I hope that helps.

The Durbin Watson test reports a test statistic, with a value from 0 to 4, where:

- 2 is no autocorrelation.
- 0 to <2 is positive autocorrelation (common in time series data).
- >2 to 4 is negative autocorrelation (less common in time series data).

A **rule of thumb** is that test statistic values in the range of 1.5 to 2.5 are relatively normal. Values outside of this range could be cause for concern. Field(2009) suggests that values under 1 or more than 3 are a definite cause for concern.

<https://www.statisticshowto.datasciencecentral.com/durbin-watson-test-coefficient/>

<https://newonlinecourses.science.psu.edu/stat501/node/366/>

Normality test

<https://www.r-bloggers.com/collinearity-and-stepwise-vif-selection/>

VIF