How to Validate Regression Models

Author: John Bonfardeci (john.bonfardeci@gmail.com)
Date: 2021-07-21

Introduction

With the prevalence of online courses in machine learning (ML) and the popularity of Python and R, there is no question as to the enthusiasm for ML. It's no longer the dark magic reserved only for "quants" who lurked within the shadows of fintech companies pre-2008. ML is an exciting subject that promises a lot of predictive power and there are countless introductory online tutorials and evangelists that make ML seem as easy as `model.fit(X).`

While amazing ML libraries, such as Scikit Learn and StatsModels, have abstracted the complexity of ML, it by no means excludes the aspiring data scientist from the burden of proving their trained ML models are a good fit for the data. Too many online tutorials do not address the vital topic of model validation. Determining model validity only by comparing the predictive accuracy on the validation data set is not enough. Deploying an invalid model to production can have disastrous results for an organization if decision makers heed faulty predictions.

In this article, we will examine validation methods for linear and logistic regression models.

Conventions Used for this Article

The following subsection category names have been defined based on the preferences of the audience. The nomenclature was selected based on my favorite race of humanoids in the Star Trek universe.

For Vulcans

Includes deep technical and academic content for ML practitioners.

For Non-Vulcans

Includes content at a high level for managers and other business-level professionals who may work with ML practitioners.

External Materials

All code used to produce the visuals in this article were written in Python for Jupyter Notebooks and are located in my personal GitHub repository. The project can be cloned or downloaded from https://github.com/jbonfardeci/model-validation-blog-post.

Model Evaluation

Model Evaluation is but one of a series of logical steps within the industry-recognized acronym **CRISP-DM**, which stands for the **Cross-Industry Standard Process for Data Mining.** While the "Data Mining" part of the acronym may seem irrelevant to ML, the steps defined within this standard are designed to ensure best practices are followed to produce valid models. See Figure 1 (below) for the main steps.

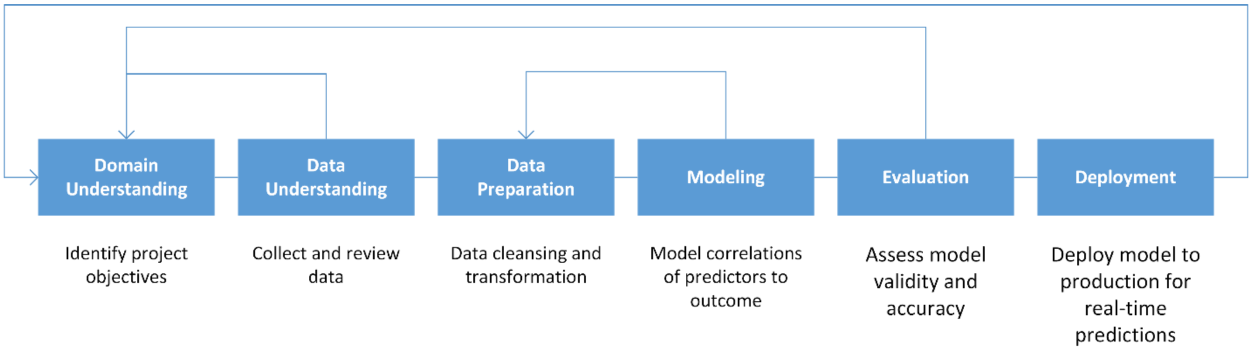

Domain Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Identify project objectives
Collect and review data
Data cleansing and transformation
Model correlations of predictors to outcome
Assess model validity and accuracy
Deploy model to production for real-time predictions

Figure 1. CRISP-DM Process. Each step is iterated until the most accurate, parsimonious (easy to explain), and valid model is achieved.

As indicated by the directional arrows in Figure 1 (above), each step can be reiterated until an optimal state is achieved. Discussing each step (and each step within each step) for CRISP-DM is out of scope for this article, but model validation falls under the Model Evaluation step. More information about CRISP-DM can be found at https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome

Why Use Graphs for Model Validation?

Anscombe's Quartet

There are many types of charts statisticians and data scientists use to describe models and assess their validity. In 1973, a well-known statistician named Francis Anscombe set out to prove the importance of graphing data. He created four datasets, now known as *Anscombe's Quartet* (Figure 2a below) – all with the same mean, standard deviation, and regression line. But each model is qualitatively different.

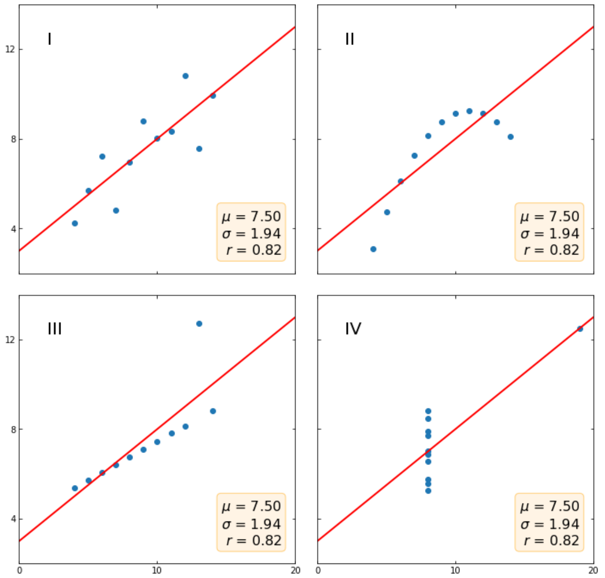


Figure 2a. Anscombe's Quartet. Each dataset above shares the same mean, standard deviation, regression line, and r-squared value.

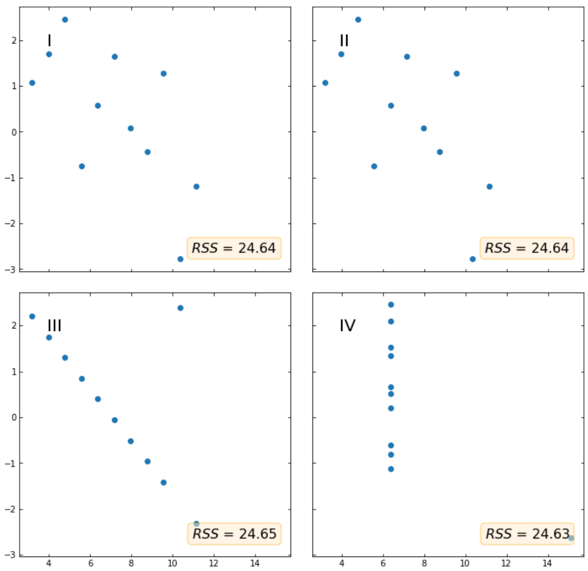


Figure 2b. Plots of Residuals (error terms) by Predicted Values for Anscombe's Quartet. Each dataset above shares a nearly identical RSS value but the error terms for each are very different. Only the top two models display homoscedasticity with no discernable patterns.

If we were to only evaluate the regression line and its r-squared value of 0.82 for each of the models, it would appear all four models are identical as far as model accuracy. But only the first model is a valid linear model. The second model is curvilinear. The third model has a bad leverage point that skews the linear model. The fourth model has all but one observation where X is a constant. When we plot the residuals (error terms) by the predicted values for each model (Figure 2b above), the top two models appear identical but the bottom two are very different even though their RSS values are very close – 24.65 and 24.63. respectively.

Validating Linear Regression Models

For Vulcans

For linear regression models, we test for *homoscedasticity* (same dispersion), aka constant variance, by plotting the model's residuals on the Y-axis by the model's predicted values on the X-axis (Figure 3 below). The model is valid if there is no discernable pattern. That is – the dots are randomly scattered from left to right, as shown in the first plot in Figure 3. The model is invalid if there is a pattern indicating *heteroscedasticity* (different dispersion), aka non-constant variance, as shown in the second and third plots in Figure 3. Heteroscedasticity in a plot indicates that one or more of the *Seven Classical Assumptions of Ordinary Least Squares (OLS)* have been violated. The assumptions of OLS are:

1. The regression model is linear in the coefficients and the error term.
2. The errors have a mean of zero.
3. The predictors are uncorrelated with the errors.
4. The errors are uncorrelated with each other.
5. The errors terms have constant variance.
6. No predictor is a perfect linear function of other predictors.
7. The errors are normally distributed.

Think of the 5th assumption of OLS about constant variance (above) as more of a symptom that manifests itself as a pattern in the errors when one or more of the six assumptions of OLS are not met.

For Non-Vulcans

For linear regression models, that is a model that estimates a numerical value from a set of predictors, data scientists and statisticians plot the model's errors (the vertical difference between actual and predicted values) by its predicted values. If there is no apparent pattern in the plot and all the dots appear randomly scattered from left to right, the model is valid as shown in the first plot in Figure 3 below. This is known as *homoscedasticity* which is a combination of the Greek root terms, "homo" (same) and "skedastikos" (dispersion). If this plot displays an evident pattern, such as the bowtie pattern or fan shape shown in the second and third plots in Figure 3 below, the model is invalid. The two aforementioned plots display *heteroscedasticity* (different dispersion).

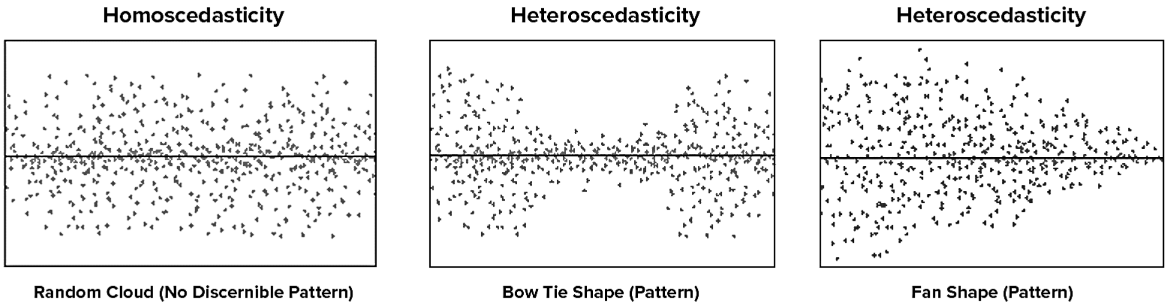
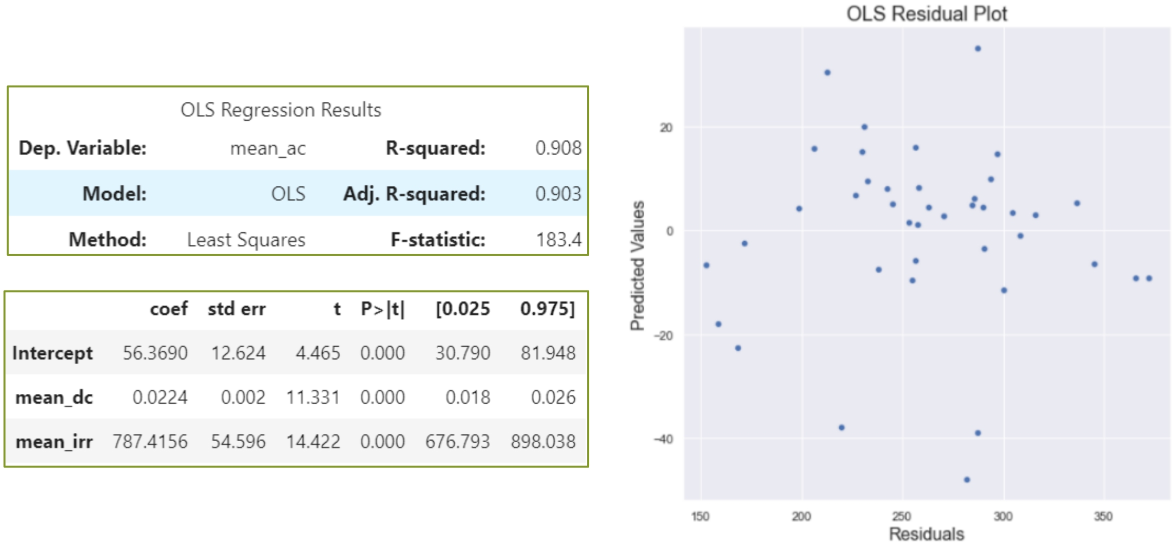


Figure 3. Plots of a linear regression model's residuals (error terms) by its fitted values should display no discernable pattern. It should look like a random cloud as shown in the first plot which displays *homoscedasticity*, meaning same dispersion.

Please note that linear regression models are not limited to OLS models. The assumptions about error terms for OLS hold true for any model that predicts a continuous numerical target, including: decision trees, random forests, boosted trees, and neural networks.

Graphs for Model Evaluation

Consider the StatsModels OLS model in Figure 4 (below) with an Adjusted R-squared value of 0.90 and a residual plot that shows constant variance with no discernible pattern. The R-squared value simply means that 90% of the variance in the target variable can be explained by its predictors.



OLS Regression Results

Dep. Variable:	mean_ac	**R-squared:**	0.908
Model:	OLS	**Adj. R-squared:**	0.903
Method:	Least Squares	**F-statistic:**	183.4

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	56.3690	12.624	4.465	0.000	30.790	81.948
mean_dc	0.0224	0.002	11.331	0.000	0.018	0.026
mean_irr	787.4156	54.596	14.422	0.000	676.793	898.038

Figure 4. Output from a StatsModels OLS Model

While residual plots are used to examine linear regression models for violations of the assumptions of OLS, they don't provide information on *how well* a model matches the data (Sanford Weisberg, Applied Linear Regression, 3rd Edition, pp. 198, 2005).

A relatively recent innovation and alternative is the *marginal model plot* as discussed by Weisberg (Sanford Weisberg, Applied Linear Regression, 3rd Edition, pp. 185-190, 2005). The marginal model plot is a practical graphical tool for visualizing how well a model fits the data by comparing a model's predicted line of fit (ŷ pronounced "y-hat") to the actual line of fit (Y).

For Vulcans

To create a marginal model plot, we overlap a model's predicted value (ŷ) for each observation on top of the actual values (Y) on the Y-axis, and any one of the continuous numerical predictor values on the X-axis. We then employ the LOESS (aka LOWESS) function to "smooth" both the Y and predicted values (ŷ). More about LOESS can be found at https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm

Figure 5a and 5b (below) shows a marginal model plot for each of the two predictors for the OLS model in Figure 4 (above). As Figure 5a reveals, X_1 is a close fit for Y given the predicted blue line is very close to the red line for actual values. However even though X_2 is statistically significant given its p-value < 0.05, it is not a good predictor of Y as evidenced by the distance between the blue predicted line and the red line for the actual values.

For Non-Vulcans

In Figures 5a and 5b (below), a marginal model plot was created for each predictor (X_1 and X_2) in the OLS model specified in Figure 4. The blue lines represent the predicted values. The red lines represent the actual target values in the data. Because the blue line in Figure 5a is very close to the red line, the predictor X_1 is an excellent predictor of that target variable. But as we see in Figure 5b, the blue line is not very close to the red line meaning that X_2 is not a very good predictor of the target variable.

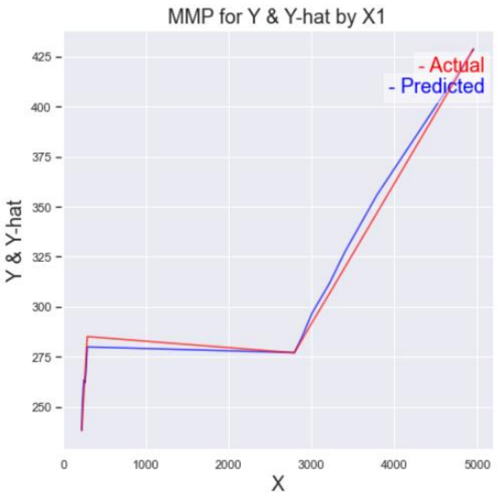


Figure 5a. Marginal Model Plot for X_1.

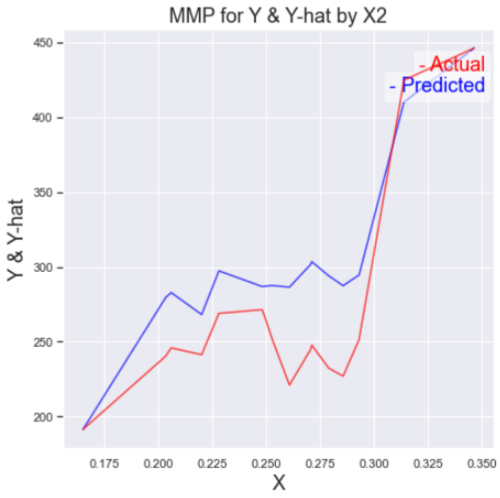


Figure 5b. Marginal Model Plot for X_2.

Validating Classification Models

For classification models, the convention is to employ a goodness-of-fit (GoF) test to determine if the model has been specified correctly. If the GoF test results in a p-value (probability value) that is less than the significance level, say alpha=0.05, the model is rejected. Otherwise, the model is accepted.

For Vulcans

The GoF test that is commonly applied to classification models is the Hosmer-Lemeshow (HL) test (Hosmer D.W. and Lemeshow S. (1980) "A goodness-of-fit test for the multiple logistic regression model." Communications in Statistics A10:1043-1069). But the HL test has serious problems, especially that it's subject to providing false negatives or false positives for GoF with even slight changes to the test's arbitrary hyperparameter for group size. (Allison, Paul. "Hosmer-Lemeshow Test for Logistic Regression: Statistical Horizons." Statistical Horizons | Statistics Training That Makes Sense, Statistical Horizons, 27 Nov. 2019, statisticalhorizons.com/hosmer-lemeshow).

Furthermore, the HL test has shown to produce wild swings in p-values due to large data sets. (Journal of Palliative Medicine. Volume 12, Number 2, 2009).

> The Hosmer-Lemeshow test detected a statistically significant degree of miscalibration in both models, due to the extremely large sample size of the models, as the differences between the observed and expected values within each group are relatively small.

Per the quote above, the errors were relatively small, meaning the model explained the target variable Y reasonably well. But the HL GoF said otherwise! This is a false negative or what's known as a Type II Error. In other words, we failed to reject the null hypothesis (H_0) when we should have.

While Weisberg described marginal model plots in the context of linear regression models, they also work very well for classification models that predict the probability of an observation belonging to a class. This applies to popular classifier models such as logistic regression, decision tree, random forest, boosted trees, and Support Vector Machine.

To create a marginal model plot for classification models we can utilize the same function described for linear regression models. In Figures 6a-6bc below, we overlap a model's predicted probability value for each observation on top of the actual Y values on the Y-axis, and any one of the continuous numerical predictor values on the X-axis. Even though Y consists of only finite integer values (0, 1, 2, …n) indicative of class labels, the LOESS (aka LOWESS) function "smooths" both the Y and predicted Y ($\hat{y}$) values so we can compare apples to apples.

For Non-Vulcans

While Weisberg described marginal model plots in the context of linear regression models, they also work very well for classification models. As shown in Figures 6a-6c (below) for a two-class logistic regression model.

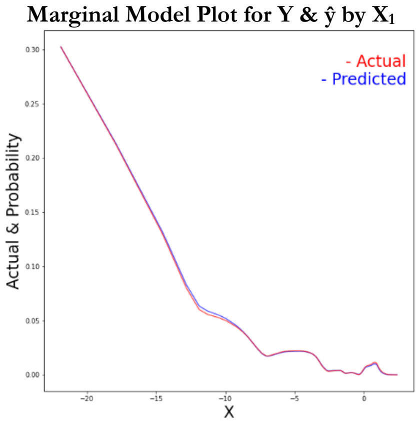


Figure 6a. X_1 is a very good predictor of Y.

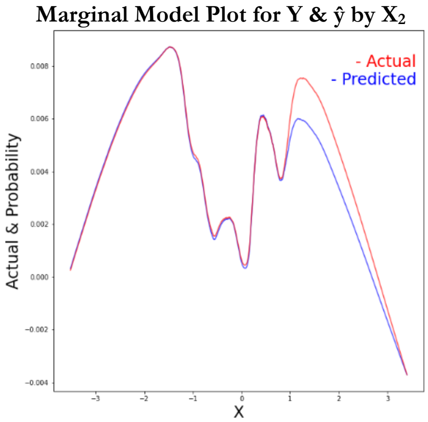


Figure 6b. X_2 is a good predictor of Y except between values between ~1.0 and ~3.5.

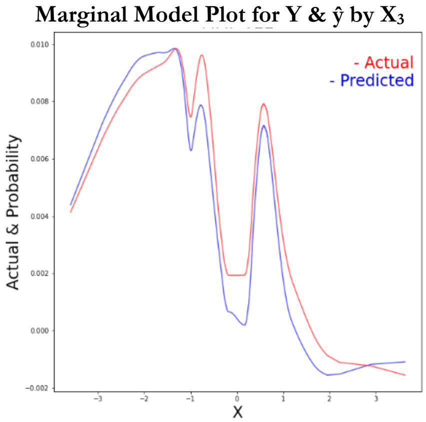


Figure 6c. X_3 is a poor predictor of Y.

For All Carbon-based Lifeforms

In Figure 6a (above), the blue 'Predicted' line is very close to the red 'Actual' line. In this case, variable X_1 for all values is an extremely good predictor of Y. The plot of X_2 (Figure 6b above) is also a very good predictor of Y except within the range of values on the X-axis between approximately 1.0 and 3.5. In this case we would investigate why X_2 is a poor predictor in this range, such as the effects of outliers. We may also look for any interactions X_2 may have with another X_n variable. If an interaction with another variable is found we would include a new variable in the linear formula for $X_2 * X_n$, which may or may not improve the fit. And for Figure 6c for predictor X_3, which is clearly a poor fit given the distances between the predicted values and actual values, we may decide it isn't useful to keep in the model.

Key Takeaways

In summary, model validation is a serious responsibility. If ignored, the consequences for key decision makers can be disastrous. Just because a model appears to be as accurate for the validation data set as it was for the training data set, does not mean the model is a good fit for the data. Visual analysis of diagnostic plots for goodness-of-fit (GoF) is the most reliable method for evaluating model fit over the use of R-squared or p-values.