A Practical Guide to Regression and Model Selection

Varun Varanasi

August 24, 2023

Contents

1 Introduction

Over the course of my undergraduate degree I've come across a plethora of different regressions, model selection methods, and other general practices which I've had trouble keeping straight. I often found myself wishing for a practical reference guide with summaries of regression methods, what their assumptions were, and when to best use them. With that in mind, this guide is meant to be a quick and dirty refresher for myself when approaching different data problems. This is by no means comprehensive nor rigorous, but I hope you find it useful. Please feel free to reach out to varun.varanasi@yale.edu if you have any questions, recommendations, or concerns. Happy reading!

2 Classical Regression

This section is focused on parametric regression methods in which we assume a form of the relationship between the dependent and independent variables.

2.1 Classical Regression Assumptions

- Sample is representative of larger population
- Independent variables are measured with no errors
- Deviations from the model have an expected value of 0
- Variance of residuals is constant across observations (homoscedasticity)
- Residuals are uncorrelated

2.2 Linear Regression

Linear regression focuses on the conditional probability distribution of the response variable based on the predictors

Model Structure

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

Least-Squares

The goal of least-squares regression is to minimize the sum of mean squared loss.

$$\hat{\beta} = argmin \sum_{i=1}^{n} (\beta \cdot x_i - y_i)^2$$

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

Model Assumptions

- Weak Exogeneity: Predictor variables are constants and are error free
- Linearity: Response is a linear combination of predictors
- Constant Variance: variance of errors is independent of predictor variables
- Independence of Errors
- No Perfect Multicolinearity: no linear relationships between predictor variables

Unique vs Marginal Effects

Regression coeffecients can be understood as the "unique effect" or the expected value for the partial derivative of the predictor on the regressor, in other words, the resulting response change for a unit increase in the predictor. On the other hand, correlation coeffecients or simple linear regression coeffecients can be understood as the "marginal effect" or the total derivative of the predictor on the regressor.

If other predictors capture the impact of a covariate on the response, the unique effect can be 0 despite the marginal effects being large. Conversely, if other predictors capture the variabiilty of the response in the same way the existing predictor does. Therefore, the total derivative is near 0, but the partial would be large.

Group Effects

When certain predictors move together, the interpretation of a regression coefficient loses its meaning. Instead, we introduce the idea of a group effect or the movement in the response variable given that the subset predictors changes according to defined weight parameters. Group effects generalize the idea of unique effects to groups of variables.

2.3 Logistic Regression

Logistic regression models the probability of an event occuring by modeling the log-odds of the event as a linear combination of the explanatory variables. The method can be extended beyond binary classifications to multinomial logistic regression and to ordered classes via ordinal logistic regression. Although the probabilities can be used to construct a binary classifier, it is important to note that logistic regression itself is not a classifier.

Model Structure

$$p = \frac{1}{1 + e^{-X\beta}}$$

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

For mulitnomial logistic regression, we can generalize to N + 1 categories where each category is given its own probability distribution. Furthermore, the sum of the probabilities over all categories is equal to unity.

We define our probailities as follows:

$$p_n(x) = \frac{e^{\beta_n x}}{1 + \sum_{u=1}^{N} e^{\beta_u x}}$$

$$p_0(x) = 1 - \sum_{u=1}^{N} p_n(x) = \frac{1}{1 + \sum_{u=1}^{N} e^{\beta_u x}}$$

Estimation Method

Logistic regression coefficients are generally solved via MLE methods since there is no closed form. If the solution does not converge, there are a handful of possible interpretations:

- Large variable to response ratio (conservative Wald Estimate, a hypothesis testing statistics)
- Multicolinearity
- Sparse data
- Perfect predictors for classification (yields infiite coefficients in MLE)

Rule of ten states that stable solutions generally arise when there are at least 10 responses per predictor

2.4 Poisson Regression

Poission regression or log-linear model is a generalized linear model that is used to model count data. The model assumes that the response follows a poisson distribution and can that the log expected value can be modeled via a linear combination of predictors.

Model Structure

$$\log(E(Y|x)) = \alpha + \beta X$$

The corresponding probability mass function of Y is given by:

$$p(y|x, \alpha, \beta) = \frac{e^{y(\alpha+\beta X)e^{-e^{\alpha+\beta X}}}}{y!}$$

Maximum Likelihood Estimation

Given a dataset of response, we can construct a log-likelihood function:

$$\ell = \sum_{i=1}^{m} \left(y_i(\alpha + \beta X)x_i - e^{\alpha+\beta X} \right)$$

This likelihood function can then be optimized via traditional convex optimization methods.

Additions of regularizations parameters akin to ridge regression can reduce overfitting.

Model Assumptions

- Arrival of each event is independent
- Mean is equal to variance (assumption from Poisson distribution)

2.5 Principal Component Regression

PCR is a regression where the predictors are principal components of the explanatory variables rather than the variables themselves. PCR can be thought of as regularized regression method since yypically only the subset of principal components with higher variances are used in the regression. Consequently, PCR provides a work around for multicolinearity and high dimensionality in explanatory data.

Model Structure

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

Since PCR is simply linear regression with principal components as the new predictors, they follow the same model structure. Consequently, if every principal component is used in the regression, the model is equivalent to a simple OLS regression.

Variance of a linear form of a PCR estimator is less than the OLS counterpart. Similarly, k principal components give the best linear approximation of a rank k of the observed data.

The general process is outlined as follows:

- Perform Principal Component Analysis on the data matrix to select new regressors
- Regress the chosen principal components on the response via OLS
- Transform the corresponding principal components back to their explanatory variables via

Principal Component Analysis

PCA can be thought of as setting a p-dimensional elipsoid to the data where each axis of the elipsoid is a principal component. We can calculate principal components by finding the eigenvectors of the covariance matrix. The eigenvectors can then be used to diagonalize the covariance matrix in which we can understand the corresponding eigenvalues as the relative variance explained by the associated principal component (eigenvector).

2.6 Least Angle Regression

Least Angle Regression is a regression method designed for high-dimensional data that estimates which predictors are relevant and their associated coefficients. Least Angle Regression provdies a curve with a solution for each the values of the L1 norm of the parameter vector. Similar to step-wise regression (iterative addition of predictors) in its addition of predictors, this method increases parameters based on their correlation with the residual.

Algorithm

1. Begin with all coefficients of β set to 0
2. Find the predictor x_j with highest correlation to Y
3. Increase the coeffiecient β_j in the direction of the correlation and calculate the residual
4. Continue until another predictor x_i has equivalent correlation with the residual as x_j
5. Increase both predictors x_i, x_j in their joint least squares direction until another predictor matches their correlation with the residual
6. Repeat the above process until all predictors are in the model

Pros

- Computationally equivalent to forward selection
- Solution is useful for cross-validation
- Algorithm aptly deals with correlated variables
- Easily modifable for other similar methods
- Effective when $p \gg n$

Cons

- Sensitive to noise in the response variable since the regression is hyper dependent on residuals
- Multicolinearity in predictors can cause them to be misinterpretted as causal variables – particularly in high-dimensional data

3 Non-parametric Regression

Non-parametric regression does not assume a form of relation between the response and predictors.

Given the relation,

$$E(Y|X) = m(X)$$

the goal of non-parametric regression is to estimate the function m.

3.1 Nearest-Neighbor Interpolation

Nearest neighbor interpolation simply estimates points their closest neighbors. For a fixed k, we define

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$

where $N_k(x)$ are the k nearest neighbors of x in the dataset.

At the cost of its simplicity and flexibility, this method often results in jagged estimate.

3.2 Kernel Regression

The key idea behind kernel regression is to estimate the function m by using a locally weighted average that is defined by a kernel weighing function.

$$\widehat{m_h}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}$$

where $K_h = \frac{1}{h} K(\frac{t}{h})$

Kernel regression is very flexible and is advantageous in situations with high dimensionality, but is slower to train, requires a large dataset, and dependent on the kernel selection.

3.3 Local Regression

Local regression is a generalization of moving averages and polynomial regression that combines multiple regression models via a k-nearest neighbors based model. The model works by fitting a low-degree polynomial to each point in the data set via weighted least squares regression. Subsets for these low-degree polynomial fit are determined via k-nearest neighbors. These subsets are tuned via a smoothing parameter *alpha*. The degree of the locally fit polynomial is flexible, but when a 0 degree polynomial is used, we find a weighted moving average.

Pros

- No specifications on the function structure
- Very flexible and can model complex processes

Cons

- Inefficient data use
- Requires large dense data sets
- Output is not easily interpretable
- Computationally expensive

3.4 Gaussian Process Regression

This regression assumes that the function is a sample from the infinite dimension multivariate gaussian. Conditional on the data provided, GPR calculates the probability that the true function is the f chosen. GPR further quantifies the mean and uncertainity of these functions via a multivariate gaussian.

Pros

- Interpolates observations
- Probabilistic prediction provides easy uncertainity interpretability

Cons

- Require full data to make a prediction
- Curse of dimensionality

4 Estimation Methods

4.1 Least Squares

This is the standard approach to approximate the solution of an overdetermined system in which we aim to minimize the sum of squares of the residuals. Linear least squares have a closed form solution whereas non-linear least squares require iterative solutions.

Limitations

This method can only account for observational error in the dependent variable.

4.2 Total Least Squares

Total Least Squares is an adjustment to the least squares method that allows for errors in both the dependent and independent variables. In this method the objective function includes terms corresponding to the variance-covariance matricies of both the independent and dependent variables.

$$S = r_x^T M_x^{-1} r_x + r_y^T M_y^{-1} r_y$$

4.3 Weighted Least Squares

Weighted least squares is an estimation method that accounts for heteroscedasticity in the data. In this situation, rather than the weight matrix being diagonal, is the inverse of the covariance matrix of y.

The objective function is given by:

$$S = r^T W r$$

where r is the residuals and W is the weighting matrix.

4.4 Partial Least Squares

Partial Least Squares regression is used to find relations between X and Y by finding the multidimensional space in X that explains the maximal multidimensional variance in Y. It is particularly useful when X has multicolinearity and when there are more variables than observations.

Partial Least Squares Algorithm

For k components, repeat the following procedure k times:

1. Find directions of maximal covariance in X and Y space
2. Perform least squares regression on input score
3. Deflate X and/or Y

4.5 Regularized Regression

Regularized regression is the process of solving least squares regression with additional constraints. Regularized regression methods are useful when there are more variables than observations since under these conditions, the least squares regression is indetermined. It also serves as a method of improving generalization of the model and reducing overfitting. In the bayesian framework regularized regression can be understood as priors on the least squared solution.

4.5.1 Ridge Regression

Ridge regression adds a penalty based on the ℓ_2 norm based on a penalization parameter λ. The regression yields the following closed form:

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

Ridge regression accepts bias to reduce variance in the model. The addition of the λ term acts to ensure the eigenvalues of the covariance matrix are strictly greater than 0. Ridge regression acts to minimize coeffecients and is useful in situations where the predictors exhibit multicolinearity. Ridge regression is more accurate than other regularized regression methods in the $n > d$ regime for highly correlated predictors.

4.5.2 Lasso Regression

Lasso regression adds an ℓ_1 norm penalty to the loss function. While the function is convex, it is not strictly convex and does not have a close formed solution. Unlike ridge regression the lasso pushes coefficients towards 0 which allows it act as both a regularization and feature selection method. Unfortuantely, in situatiosn with correlated predictors, lasso regression selects an arbitrary subset of the predictors and does not account for group effects. Lasso regression is also the minimal possible relaxation of the ℓ_0 penalty that yields a weakly convex optimization problem.

4.5.3 Elastic Net Regression

Elastic net regression combines both the ℓ_1 penalty from Lasso and the ℓ_2 norm from ridge regression to create a regression that selects n covariates when $p > n$ and selects a singular covariate from a set of correlated covariates. Furthermore, this regrsesion performs well when $n > p$ and there is multicolinearity in the dataset.

5 Timeseries Methods

Timeseries are series of data points that are indexed in time order. They are generally a sequence of discrete data points. Many timeseries models are built on autocorrelation, correlation with the signal and lagged versions of itself. It is particularly useful in identifying periodic signal that is obscured by noise in the data.

5.1 Autoregressive Model

The Autoregressive model assumes that the response signal is a linear combination of its previous values and a stochastic noise term.

$$X_t = \sum_{i=1}^{p} \psi_i X_{t-i} + \epsilon_t$$

The ideal parameter p is one afterwhich the partial autocorrelations are 0. A consequence of this model is that spikes in the data impact all future values.

5.2 Moving-Average Model

Conceptually, the moving-average model is a linear regression of the current value against current and previous noise terms. You can think of it as regressing on differences from the mean.

$$X_t = \mu + \sum_{i=1}^{1} \theta_i \epsilon_{t-i} + \epsilon_t$$

Unlike in the AR model, shocks in the MA model only impact the future up to q timesteps forward.

5.3 Autoregressive Integrated Moving-Average Model

ARIMA models combine AR and MA models with time differencing to fit the data as well as possible.

$$X_t - \alpha_1 X_{t-1} - ...\alpha_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$$

Differencing the timeseries allows you to work with non-stationary timeseries and stabilize the mean of the timeseries.

5.4 Autoregressive Conditional Heteroscedasticity Models

ARCH models predict the variance of errors in current values as a function of errors in the previous timesteps. Unlike the ARIMA family of models, these familes focus on correlating the variance of errors rather than the means. ARCH model is appropriate with an AR model is assumed for the error variance. We similarly refer to this model as the GARCH model when ARMA is assumed for the error variance.

We assume that error terms take the form $\epsilon_t = z_t \sigma_t$ where z_t is a stochastic variable and σ_t is a time-dependent standard deviation.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2$$

In the GARCH model we add an additional lagged relationship:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2$$

5.5 Decomposition

Decomposition of a timeseries aims to split a timeseries into the following sections:

- Trend: Representation of long-term behaviour of the timeseries
- Cyclic: Repeated but not periodic fluctuations in the timeseries
- Seasonal: Periodic of seasonal behaviours of the timeseries
- Irregular: Noise or idiosyncratic component

These 4 sections can be split into additive and multiplicative models.

$$y_t = T_t + C_t + S_t + I_t \qquad y_t = T_t \cdot C_t \cdot S_t \cdot I_t$$

Note that trend and cyclic components can occassionaly be treated as a singular component. Analysis of the cyclic component can also be done via spectral analysis.

6 Model Fit and Improvements

Building a model is only half the battle. Now, we focus our efforts on validating our models and measuring the goodness of fit.

6.1 Residual Analysis

Residuals are the differences between the observed and predicted values. Patterns in residuals indicate poor model fit. For example, a fanning effect seen in the residuals can be indicative of heteroscedasticity in the model.
Below is a short list of plots useful for visually analyzing residuals:

- Scatter plot of residuals vs predictors
- Scatter plot of residuals vs time
- Run charts of response and errors vs time
- Lag plots
- Histogram and qunatile plots

6.2 Outliers

Regression models can be susceptible to outliers – particularly when dealing with weak signals. Cook's Distance is a metric used to measure the influence of a point on a least squares regression. It can be used to identify data points that require special attention or to select regions to collect more data.

At a high level Cook's Distance measures the impact of deleting a given observation.

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y_j} - \hat{y_{j(i)}})^2}{p \cdot MSE}$$

where p is the rank of the model and $y_{j(i)}$ is the model fit without point i.

Generally a Cook's distance greater than 1 can be used as a cutoff to remove overly influential data points.

6.3 Heteroscedasticity Corrections

A common error found in the data is heteroscedasticity, inconsistent variance in errors across the model. This can be addressed through different regression estimation methods such as weighted least squares or via data transformations. Below is a quick outline of ways to handle heteroscedasticity in your model.

- Stabilizing transformations of the data such as the log transform
- Weighted least squares estimation methods
- Using heteroscedasticity-consistent standard errors

6.4 Coefficient of Determination

Also known as R^2 the coefficient of determination is a measure of how much of the response is explaiend by the predictors. It tends to range between 0 and 1, but R^2 can be negative when the test data is not equivalent to the training data. This situation can be interpreted as the mean being a better fit to the data than your model.

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 is the squared correlation coefficient between the predicted and observed values when the model is a linear least squares regression with an intercept. If the model is a simple linear regression, then the R^2 is the squared correlation coefficient between the predictor and the response.

Adjusted R^2 adds an additional penalty for the number of predictors in the model to reduce chances of extraneous variables inflating the goodness of fit.

$$R^2_{adj} = 1 - \frac{RSS/df_{res}}{TSS/df_{tot}}$$

where $df_{res} = n - p$, the degrees of freedom of the model and $df_{tot} = n - 1$, the degrees of freedom of the mean.

Important Caveats of R^2

- Does not indicate causal relations
- Can't tell you if the variance explanation is misatributed to the predictors present
- Doesn't tell you if the regression model is correct
- Doesn't inform you if the predictor selection is correct
- Blind to colinearity in the data
- May not have sufficient data for a resonable conclusion

6.5 Mean Squared Prediction Error

Mean squared prediction error is the expected squared difference between fitted values and observations.

$$MSPE = \frac{1}{n}\sum_{i=1}^{N}(f(x_i) - \hat{f}(x_i)) = ME^2 + VAR$$

Since the true function f is unknown, we resort to estimating this error. MSPE can be calculated exactly the model is tested on new data. This can occur if new data is collected or if data is artificially removed for an out-of-sample test. If MSPE is similar in in-sample and out-of-sample data, it indicates that the model is well-fit, but if the in-sample greatly outperforms the out-of-sample test, the model is likely overfit.

6.6 Bootstrap

One way of improving your model is via residual bootstrapping.

1. Fit a model and retain values for $\hat{y}_i$ and corresponding residuals $\epsilon_i = y_i - \hat{y}_i$
2. For each data point (x_i, y_i), add a randomly sampled residual ϵ_j to the response y_i
3. Refit the model with these newly adjusted response variables
4. Repeat steps 2 and 3 for large n

This model better attributes the random error of the model across all data points.

6.7 Cross-Validation

Cross-Validation is a method to measure how a model generalizes to an independent data set. CV partitions the data into subsets for model training and testing. The results of the test sets are averaged across many partitions to more accurately predict model performance. There exist a handful of CV methods including leave-one-out CV, k-fold CV, and nested CV. While a powerful method to estimate predictive performance, CV is computational expensive and requires data to be drawn from the same population.

7 Model Selection

This section is focused on metrics to select a model given a subset of models.

7.1 Mallows's Cp

Mallows's Cp is a goodness of fit measurement used to evaluate models fit with ordinary least squares.

For P regressors selected from a set of K total predictors, we define the statistics as:

$$C_p = \frac{SSE_p}{S^2} - N + 2(P+1)$$

- Error Sum of Squares: $SSE_p = \sum_{i=1}^{N}(Y_i - Y_{pi})^2$
- Estimate of Residual Variance on Full predictor Set: $S^2 = \frac{1}{N-K}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$

While a useful statistic to measure model fit, Mallows's Cp requires a large sample size and is not suited for large complex model selection.

7.2 F-Test

The F-test is useful in determining whether a given model is a significantly better fit to the data than another.

$$F = \frac{\frac{RSS_1 - RSS_2}{p_2 - p_1}}{\frac{RSS_2}{n - p_2}}$$

If the model is fit via weighted least squares, you can substitute the RSS with χ^2, the weighted sum of squared residuals. If model 2 does not have significant improvement over model 1, then F will follow an F distribution with $(p_2 - p_1, n - p_2)$ degrees of freedom. We reject the null hypothesis if F is greater than a critical value of the F distribution.

7.3 Information Criteria

Information Criteria quantifies the trade-off between model fit and simplicity (overfitting and underfitting) and is thus useful in model selection. The ideal model in a set of candidate models is that with the minimum value.

Aikake Information Criteria

For k paramaters in the model and the corresponding maximized likelihood function $\hat{L}$, the AIC is calculated by

$$AIC = 2k - 2\ln(\hat{L})$$

AIC does not tell us any information about the absolute quality of a model, but rather its relative quality to other models.

Bayesian Information Criteria

Across n observations, a model with k paramaters corresponding maximized likelihood function $\hat{L}$, the BIC is calculated by

$$BIC = k\ln(n) - 2\ln(\hat{L})$$

The BIC has a higher probability of selecting the right model as n approaches infinity, but the AIC has a lower likelihood of selecing a bad model.

8 Future Additions

- Regression Trees
- Regression Splines
- Classification Methods Section
- SVMs
- Exploratory Data Analysis Section
- Correlation Methods
- Feature Selection Section
- ROC + Precision Recall Curves
- Probit regression
- Optimal Lambda Choice for regularized regression
- Solution to weighted least squares