Week 2 :An Overview of Regression Methods

June 21, 2020

This week we cover the main and most commonly used regression models. Regression models fall under supervised learning and are applied to tasks where the target variable has a continuous value instead of a group of classes. We start by refreshing our knowledge with a simple uni-variate linear regression and build our way up to end with Logistic regression.This week's material will be largely based on the book "Introduction to Statistical Learning"[1]. It is freely available here: http://faculty.marshall.usc.edu/gareth-james/ISL/. The book is a also a rich source of material for more in depth knowledge.

Outline: Learning Objectives

1. Linear Regression
2. Multivariate Regression
3. Ridge Regression
4. Lasso Regression
5. Polynomial Regression
6. Logistic Regression

Linear Regression

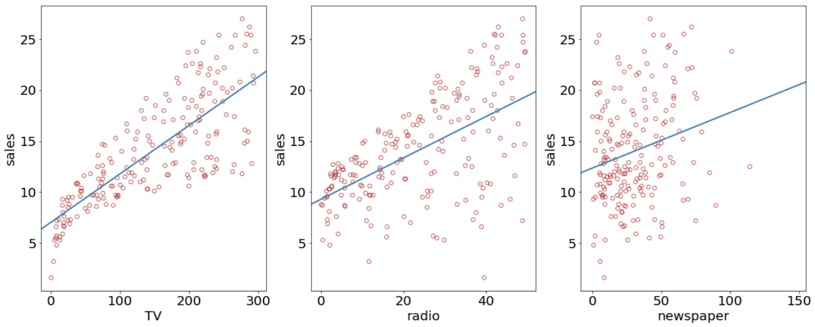


Figure 1: Advertising Data. Can we predict future sales? is the relationship linear?. *Linear Regression is a simple approach of supervised learning. It assumes that the relationship between the feature and the target is linear.*

We assume that the linear regression model, which is a straight line, is mathematically expressed by the following equation[2]:

[1] Trevor Hastie Gareth James, Daniela Witten and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 8th edition, 2017. ISBN 978-1-4614-7138-7

[2] The betas are called coefficients or parameters. Epsilon is the error term, which is the difference between the predicted target given X and the observed y

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

The element that interest the most is the error term epsilon which could be written as below:

$$\epsilon_i = y_i - \hat{y}_i \tag{2}$$

The errors also known as the residuals are the difference between the observed y_i and the estimated $\hat{y}_i$ by our model[3]. That is the grey distance in figure 2.

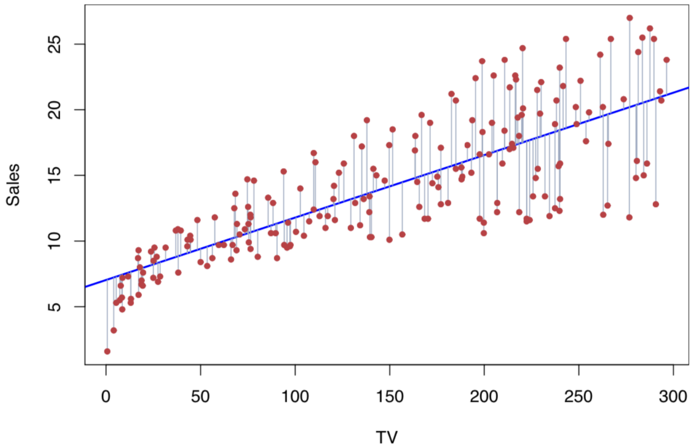


Figure 2: Based on the *Ordinary Least Squares model (OLS)*, the best fitting straight line is the ones that minimizes the error terms. In other word the one that minimizes the distance between the projected target and the observed target.

The simplest algorithm used to fit the straight line is Ordinary Least Squares (OLS). OLS estimates the betas by minimizing the below equation, also known as residual sum of square:

$$RSS = \epsilon_1^2 + \epsilon_2^2 + ... + \epsilon_n^2 \tag{3}$$

The minimizing values can be shown to be:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{5}$$

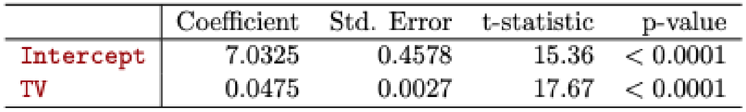
Assessing the Overall Accuracy of the Model

$$R^2 = \frac{TSS - RSS}{TSS}, \tag{6}$$

Where,

$$TSS = \sum_{i=0}^{n}(y_i - \overline{y})^2, \tag{7}$$

[3] The superscript ˆ denotes an estimated value. The superscript overline (-) denotes the sample mean of the underlying variable



	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Figure 3: *Results for the advertising data*

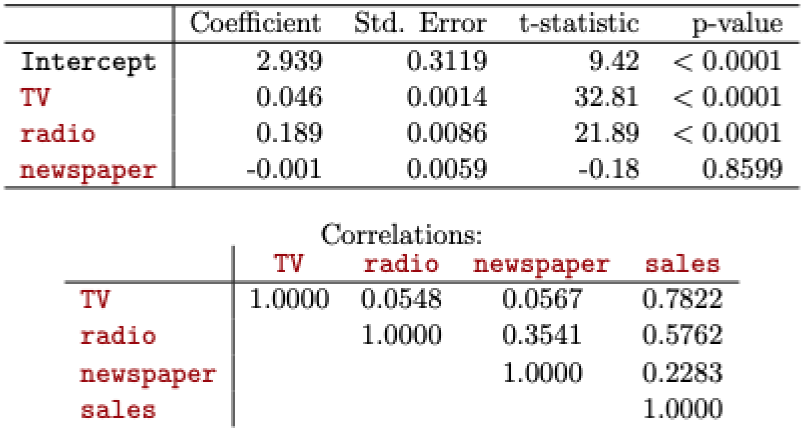
The R^2 is the most important indicator on how good the fit is. It is easy to show that the value of R^2 falls between 0 and 1. As the residuals go to zero, R^2 will converge to one. The closer it is to 1 the smaller are residuals the better is the fit. In machine learning and for the sake of this course we will use R^2 as the main indicator of the quality of the model.

Multiple Linear Regression

With a multi-dimentional regression the model becomes more complex:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{8}$$

In our particular marketing case we will be adding to new features, radio and newspaper adds, hoping we can learn more for a slightly more complex model. While the math for estimating the parameters for a multiple regression gets more complicated, we are not concerned with that, given that python will do the work for us. In principle everything that applies to the one dimensional regression



	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Figure 4: *Results for the advertising data adding two new features Radio and Newspaper advertising*

applies to the multiple regression model. In other world the R^2 remains our most important indicator when it comes to the quality of model. In an ideal world the correlation between different features is close to if not zero. With substantial correlation the variance of the parameters of each feature tend to increase, sometimes dramatically. Interpretations become hazardous since when one of the features changes everything else changes.

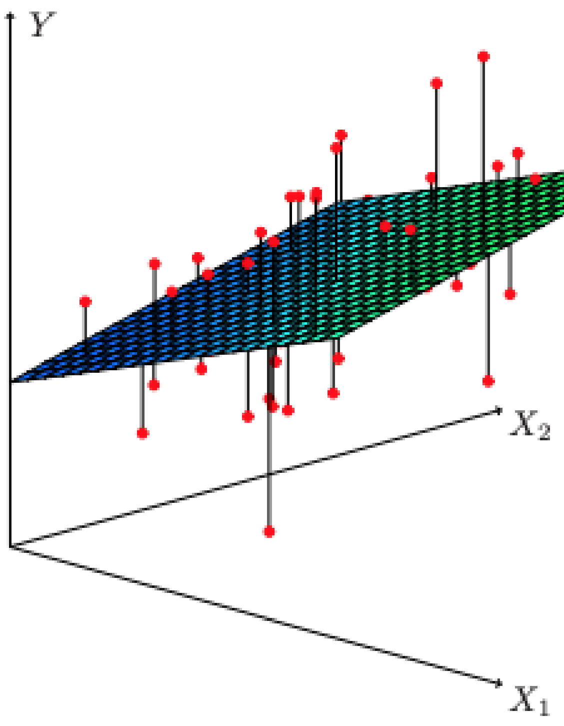


Figure 5: *In a two dimensional linear regression model we aim to fit a plane instead of a straight line*

Deciding on the important features

1. All Subset Approach The most direct approach is called all subsets: we compute the OLS fit for all possible subsets of features and then choose between them based on some criterion that balances training error R^2 with model size.

2. Forward Selection Begin with the null model — a model that contains an intercept but no predictors. Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. Add to that model the variable that results in

the lowest RSS amongst all two-variable models. Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

3. **Backward selection** Start with all variables in the model. Remove the variable with the largest p-value — that is, the variable that is the least statistically significant. The new (p 1)-variable model is fit, and the variable with the largest p-value is removed. Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

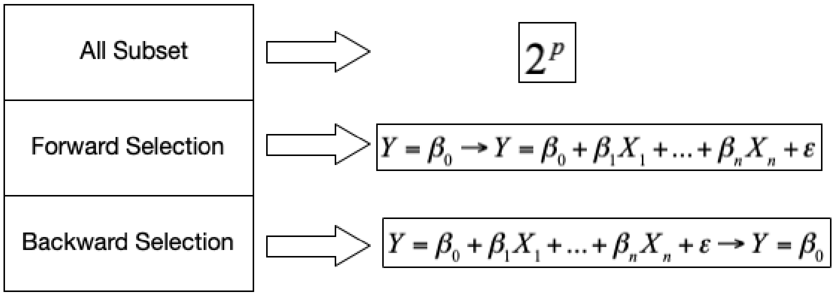

All Subset
2^p
Forward Selection
$Y = \beta_0 \rightarrow Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$
Backward Selection
$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon \rightarrow Y = \beta_0$

Figure 6: Selecting the optimal combination of features

Ridge Regression

As an alternative to the previous feature selection methodology, we can fit a model containing all p predictors using a technique that constrains or regularizes[4] the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero. It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

Instead of minimizing the RSS as we did earlier, the ridge regression solve for the $\hat{\beta}$ by minimizing the below equation:

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{9}$$

The objective function we are minimizing is similar to the one we saw with simple regression but we are adding a new term penalty term where λ is a *tuning parameter*[5].

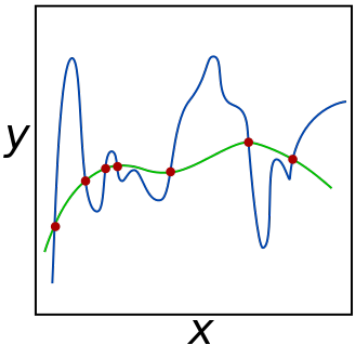


Figure 7: The green and blue functions both incur zero loss on the given data points. A learned model can be induced to prefer the green function, which may generalize better to more points drawn from the underlying unknown distribution, by adjusting lambda.*source* : Wikipedia

[4] In mathematics, statistics, and computer science, particularly in machine learning and inverse problems, regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting.*source* : Wikipedia

[5] The ridge regression imposes constraint on the parameters betas. The tuning term λ regularizes the coefficients such that if the coefficients take

As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small. However, the second term, a shrinkage penalty, is small when $\beta_1, \ldots, \beta_p$ are close to zero, and so it has the effect of shrinking the estimates of β_j towards zero. The tuning parameter serves to control the relative impact of these two terms on the regression coefficient estimates. Selecting a good value for is critical; cross-validation is used for this[6].

Scaling the Features

1. The standard least squares coefficient estimates are scale equivariant: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of 1/c. In other words, regardless of how the jth feature is scaled,$X_j\hat{\beta}_j$ will remain the same.

2. In contrast, the ridge regression coefficient estimates can change substantially when multiplying a given feature by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.

3. Therefore, it is best to apply ridge regression after standardizing the predictors, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}} \tag{10}$$

Lasso Regression

Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.

The Lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients $\hat{\beta}^L_\lambda$, minimize the quantity:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{11}$$

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.

- However, in the case of the lasso, the ‘1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.

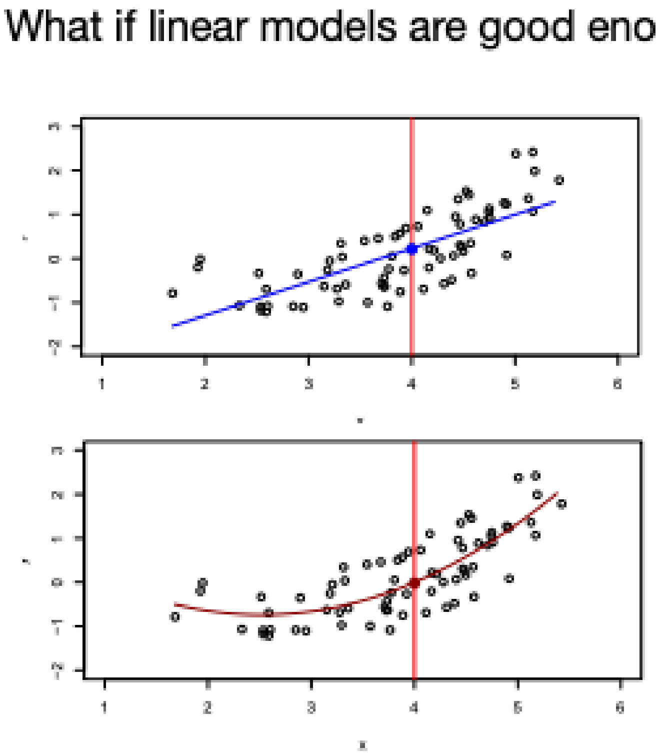
[6] We will look more into cross-validation techniques in *Week 4*

- Similarly to the subset selection, the lasso performs variable selection.

- Lasso regression falls under the sparse models[7], that is models that involve only a subset of the variables.

- similarly to the Ridge regression choosing for the right value of λ is key and it is achieved by doing cross-validations

Polynomial Regression



What if linear models are good enough?

Figure 8: linear Vs Non-linear fitting

This is how a degree-n polynomial regression model looks like:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \beta_d x_i^d + \epsilon_i \tag{12}$$

Obviously, the formula 12 applies to one feature X, where we have created new variables $X_1 = X, X_2 = X^2, ..., X_n = X^n$. This could also be applied to multiple features regressions like the below example where we try to predict income by looking at two features: years of education as well as seniority.

[7] Sparse Models are models where only a small fraction of parameters are non-zero – arise frequently in machine learning. Sparsity is beneficial in several ways: sparse models are more easily interpretable by humans, and sparsity can yield statistical benefits – such as reducing the number of examples that have to be observed to learn the model. In a sense, we can think of sparsity as an antidote to curse of dimensionality

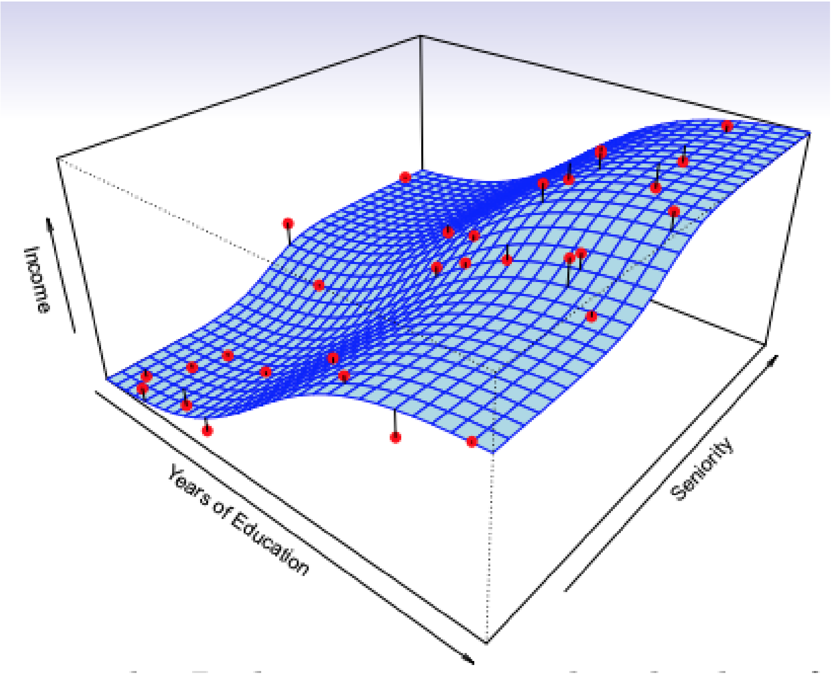


Figure 9: 2-dimentional non-linear regression

Logistic Regression

Even though, it falls under regression, most of the applications of logistic regressions are on classification problems, a topic that we will only cover next week. But assume that we solving for a task where the target variable takes a binary form. Like in the case we saw last week where we are trying to predict whether an email is spam or not based on some quantitative features. Or lets assume we are trying to predict whether a client is going to default on his credit card. The target variable in this case would be binary, default or no-default.

$$Y = \begin{cases} 0, & \text{if default} \\ 1, & \text{if no-default} \end{cases} \tag{13}$$

Can we simply perform a linear regression of Y(default) on X(income and balance) and classify as Yes if Y >^ 0.5?

- We probably could run a linear regression and the predicted value of Y could be interpreted as the probability of default.
- The problem with linear regressions nothing is stopping $\hat{Y}$ from getting values higher that one or negative.

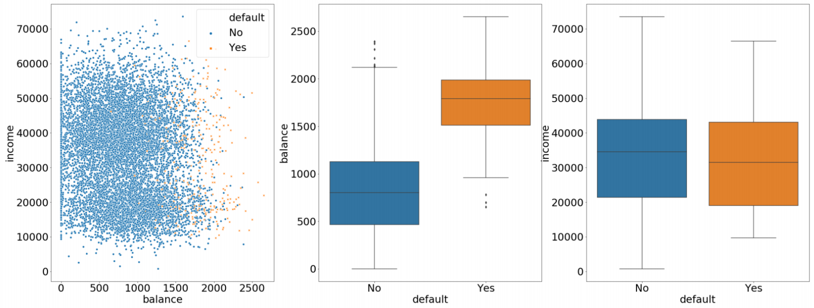


Figure 10: Credit Card Default based on income and balance. Default observations are denoted in orange.

Logistic regression uses the form:

$$Y = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{14}$$

It is easy to see that as X goes to negative infinity, Y converge to zero and when X go to infinity Y converges to 1.

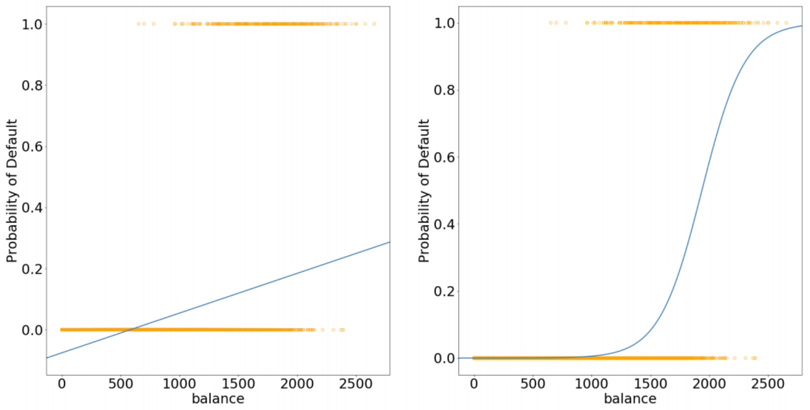


Figure 11: Logistic regression ensures that our estimate for Y lies between 0 and 1

Unlike the previous regressions, the logistic regression does not use ordinary least square method to solve for the parameters. Instead we use the Maximum Likelihood Estimation[8] method which beyond the scope of this course.

Learning Summary

- We have learned the concepts behind six different regression techniques
- We know how to interpret p-values, RSS and R^2
- we learned about different techniques to select the important features: subset, forward and backward selections

[8] In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable. *source* Wikipedia

- We introduced the concept of regularization and the tuning parameter λ
- We introduced one scaling method for features
- We understand the advantage of sparse models

References

Trevor Hastie Gareth James, Daniela Witten and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 8th edition, 2017. ISBN 978-1-4614-7138-7.