Regression

BITS Pilani
Pilani Campus

Akanksha Bharadwaj
Asst. Professor, CS/IS

Significance of R-squared

In Graph 1:

All the points lie on the line and the R2 value is a perfect 1

In Graph 2:

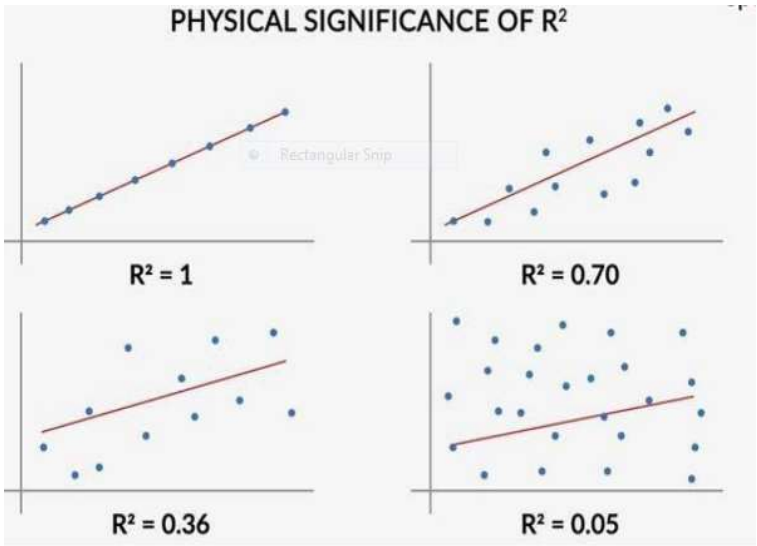
Some points deviate from the line and the error is represented by the lower R2 value of 0.70

In Graph 3:

The deviation further increases and the R2 value further goes down to 0.36

In Graph 4:

The deviation is further higher with a very low R2 value of 0.05



Testing the slope of the regression line

- For example, the slope of the regression line for the airline cost data is .**0407**. This value is obviously not zero.
- The problem is that this slope is obtained from a sample of 12 data points; and if another sample was taken, it is likely that a different slope would be obtained.
- For this reason, the population slope is statistically tested using the sample slope.
- The question is: If all the pairs of data points for the population were available, would the slope of that regression line be different from zero?

Hypothesis testing

- Here the sample slope, b_1, is used as evidence to test whether the population slope is different from zero. The hypotheses for this test follow.

$$\begin{cases} H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{cases}$$

- Note that this test is two tailed. The null hypothesis can be rejected if the slope is either negative or positive.

- A negative slope indicates an inverse relationship between x and y.

Hypothesis Testing

- To determine whether there is a significant *positive* relationship between two variables, the hypotheses would be one tailed, or

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 > 0$$

or inverse

- To test for a significant negative relationship between two variables, the hypotheses also would be one tailed, or

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 < 0$$

- In each case, testing the null hypothesis involves a ***t* test of the slope**.

t test of the slope

t TEST OF SLOPE

$$t = \frac{b_1 - \beta_1}{s_b}$$

where

$$s_b = \frac{s_\varepsilon}{\sqrt{SS_{xx}}}$$

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

why?

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

β_1 = the hypothesized slope

$df = n - 2$

t test for airline example

The test of the slope of the regression line for the airline cost regression model for $\alpha = .05$ follows. The regression line derived for the data is

$$\hat{y} = 1.57 + .0407x$$

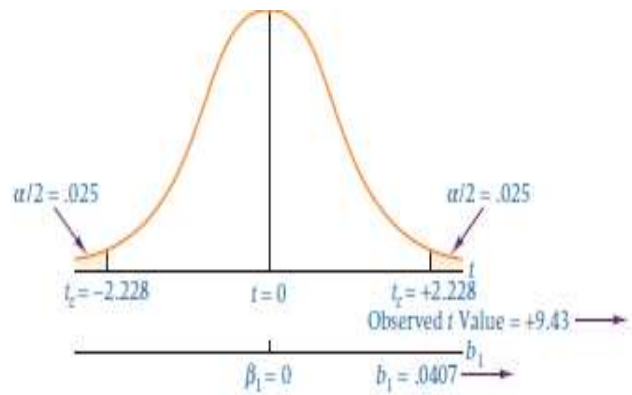
slope

The sample slope is $.0407 = b_1$. The value of $s_{\hat{y}}$ is .1773, $\Sigma x = 930$, $\Sigma x^2 = 73{,}764$, and $n = 12$. The hypotheses are

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

The df $= n - 2 = 12 - 2 = 10$. As this test is two tailed, $\alpha/2 = .025$. The table t value is $t_{.025,10} = \pm 2.228$. The observed t value for this sample slope is

$$t = \frac{.0407 - 0}{.1773\Big/\sqrt{73{,}764 - \frac{(930)^2}{12}}} = 9.43$$



The null hypothesis that the population slope is zero is rejected.

Accessing the model fit

1. **t statistic**: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not

2. **R-squared**: it tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

References

- Probability and Statistics for Engineering and Sciences,8th Edition, Jay L Devore, Cengage Learning
- Applied Business Statistics by Ken Black
- https://towardsdatascience.com/let-us-understand-the-correlation-matrix-and-covariance-matrix-d42e6b643c22
- https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765
- https://towardsdatascience.com/multicollinearity-why-is-it-a-problem-398b010b77ac
- https://pdfs.semanticscholar.org/d1ee/9331a2fe0fbb9c8ad27fbf378e3d4cb20163.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3412358/

“Correlation Is Not Causation”

- Whenever you work with regression analysis or any other analysis that tries to explain the impact of one factor on another, you need to remember the important adage:
- Correlation is not causation.
- This is critical and here’s why: It’s easy to say that there is a correlation between rain and monthly sales a product.
- Regression might show that they are indeed related.
- But it’s an entirely different thing to say that rain *caused* the sales. Unless you’re selling raincoats.

innovate achieve lead

BITS Pilani
Pilani Campus

SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS
Contact Session 9

Moving from SLR to MLR

The new aspects to consider when moving from simple to multiple linear regression are:

- **Overfitting**
  - As you keep adding the variables, the model may become far too complex
  - It may end up memorising the training data and will fail to generalize
  - A model is generally said to overfit when the training accuracy is high while the test accuracy is very low
- **Multicollinearity**
  - Associations between predictor variables
- **Feature selection**
  - Selecting the optimal set from a pool of given features, many of which might be redundant becomes an important task

Multicollinearity

- It refers to the phenomenon of having related predictor variables in the input dataset.
- In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated.
- You drop some of these related independent variables as a way of dealing with multicollinearity.

Identifying Multicollinearity

- **Looking at pairwise correlations**: Looking at the correlation between different pairs of independent variables

- **Variance Inflation Factor (VIF):** The VIF assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated.

- If there is no correlation the VIF will be 1. So the larger the number the more correlated the two variables are.

Variance Inflation Factor (VIF)

The VIF is given by:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R^2_j is the R^2-value obtained by regressing the jth predictor on the remaining predictors.

- A VIF of 1 means that there is no correlation among the jth predictor and the remaining predictor variables
- The general rule of thumb is that VIFs exceeding 4 warrant further investigation,
- while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

Fixing Multicollinearity

1. **Feature Engineer:** If you can find a way to aggregate or combine the two features and turn it into one variable

2. **Drop One:** It is common to drop one of the variables that are too highly correlated with another.

Equation for multiple linear regression

- Extending this notion to multiple regression gives the general equation for the probabilistic multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \in$$

where

y = the value of the dependent variable
β_0 = the regression constant
β_1 = the partial regression coefficient for independent variable 1
β_2 = the partial regression coefficient for independent variable 2
β_3 = the partial regression coefficient for independent variable 3
β_k = the partial regression coefficient for independent variable k
k = the number of independent variables

- In virtually all research, these values are estimated by using sample information.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k$$

where

$\hat{y}$ = the predicted value of y
b_0 = the estimate of the regression constant
b_1 = the estimate of regression coefficient 1
b_2 = the estimate of regression coefficient 2
b_3 = the estimate of regression coefficient 3
b_k = the estimate of regression coefficient k
k = the number of independent variables

Multiple regression Steps

1. Generate the list of possible dependent and independent variable
2. Collect data on variables
3. Check the relationship between each independent variable and the dependent variable using scatterplots and correlation
4. Check the relationship between independent variables using scatterplots and correlation – multicolinearity
5. Use the non redundant independent variables in the analysis to find the best fitting model
6. Use best fitting model to make predictions about the dependent variable

Some Problems with R-squared

- **Problem 1:** Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. Consequently, a model with more terms may appear to have a better fit simply because it has more terms.

- **Problem 2:** If a model has too many predictors and higher order polynomials, it begins to model the random **noise** in the data. This condition is known as overfitting the model and it produces misleadingly high R-squared values and a lessened ability to make predictions.

Adjusted R-squared

- The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.
- The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.
- It decreases when a predictor improves the model by less than expected by chance.
- Use the adjusted R-square to compare models with different numbers of predictors

Adjusted R^2

- Ranges from 0 to 1 with values closer to 1 indicating a stronger relationship
- Adjusted R^2 is the value of R^2 which has been penalized for the number of variables added to the model
- Therefore Adjusted R^2 is always smaller than R^2

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

$$R^2{}_{adj} = R^2(1 - R^2)\left(\frac{p}{n - p - 1}\right)$$

p = number of predictor variables (regressors, not including intercept)
n = sample size

Akaike Information Criterion

- AIC considers both the fit of the model and the number of parameters used. More parameters result in a penalty
- Allows us to balance over- and under-fitting in our modelled relationships
  - We want a model that is as simple as possible, but no simpler
  - A reasonable amount of explanatory power is traded off against model size
  - AIC measures the balance of this for us

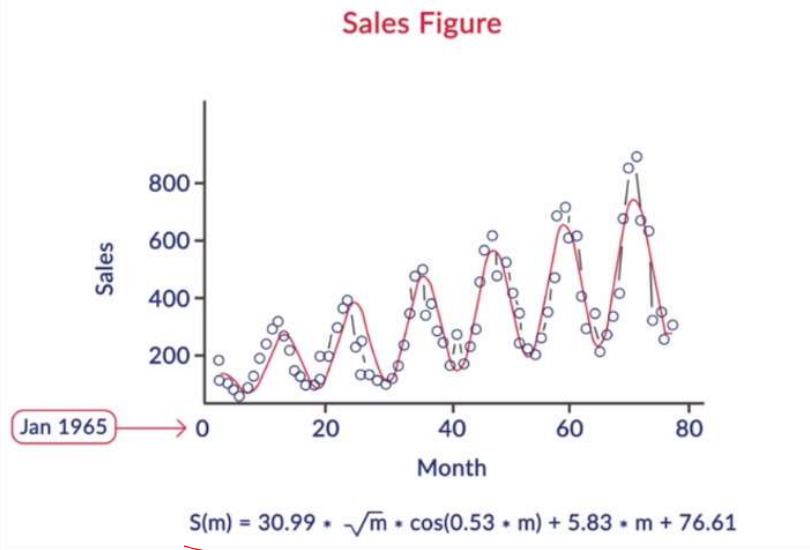
$$AIC = n \times log(\frac{RSS}{n}) + 2p$$

Residual sum of squares

- Here, n is the sample size meaning the number of rows you'd have in the dataset and p is the number of predictor variables.

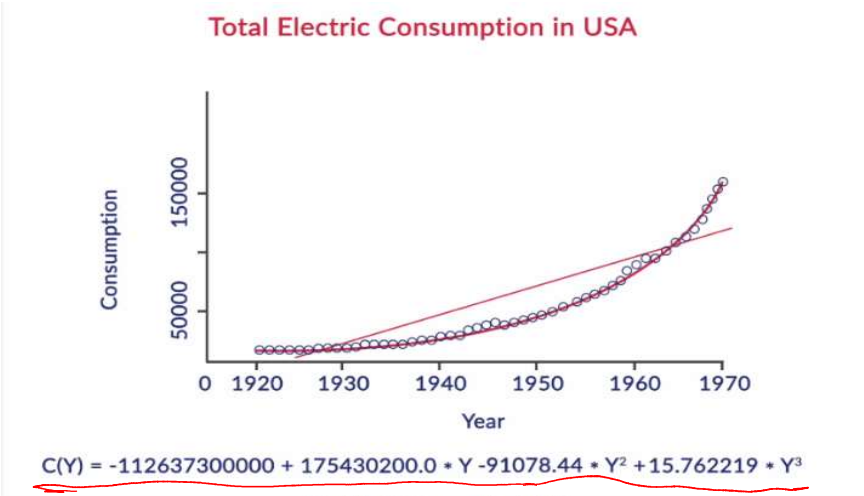
Non-linear Regression

In the first example, notice that the data points oscillate and follow a sine or cosine type of function.



Sales Figure

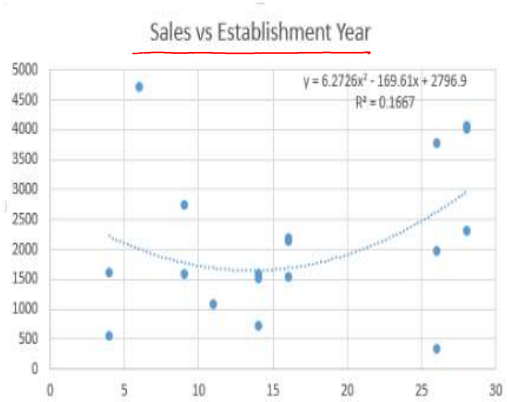
In the second example of electricity consumption, the data points gradually increase non-linearly, indicative of a polynomial or an exponential function:

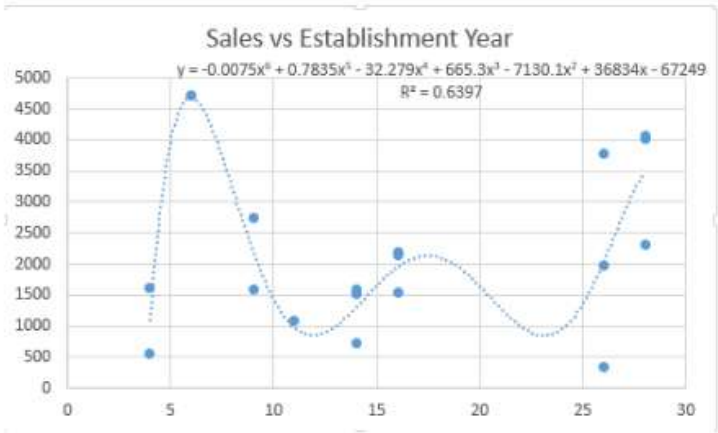


Total Electric Consumption in USA

Non-linear Regression

- Polynomial regression is another form of regression in which the maximum power of the independent variable is more than one.
- In this regression technique, the best fit line is not a straight line instead it is in the form of a curve.





What is bias?

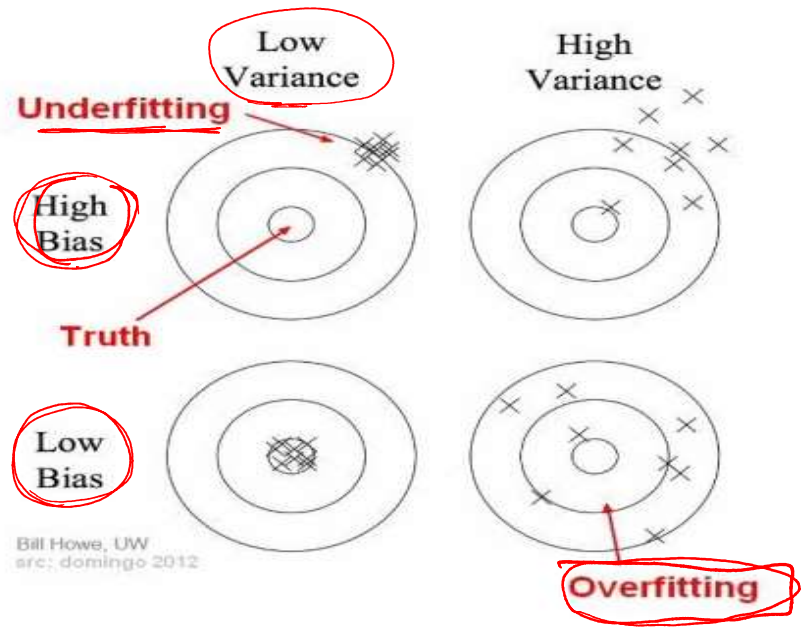
- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
- Model with high bias pays very little attention to the training data and oversimplifies the model.
- It always leads to high error on training and test data.

What is variance?

- Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.
- Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before.
- As a result, such models perform very well on training data but has high error rates on test data.

Bias and variance using bulls-eye diagram

- Center of the target is a model that perfectly predicts correct values.
- As we move away from the bulls-eye our predictions get worse and worse.

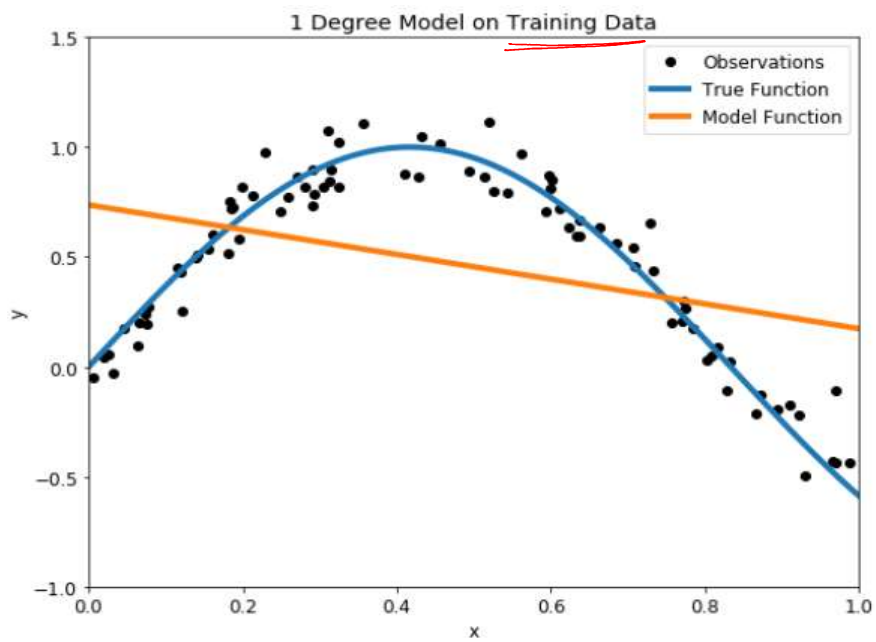


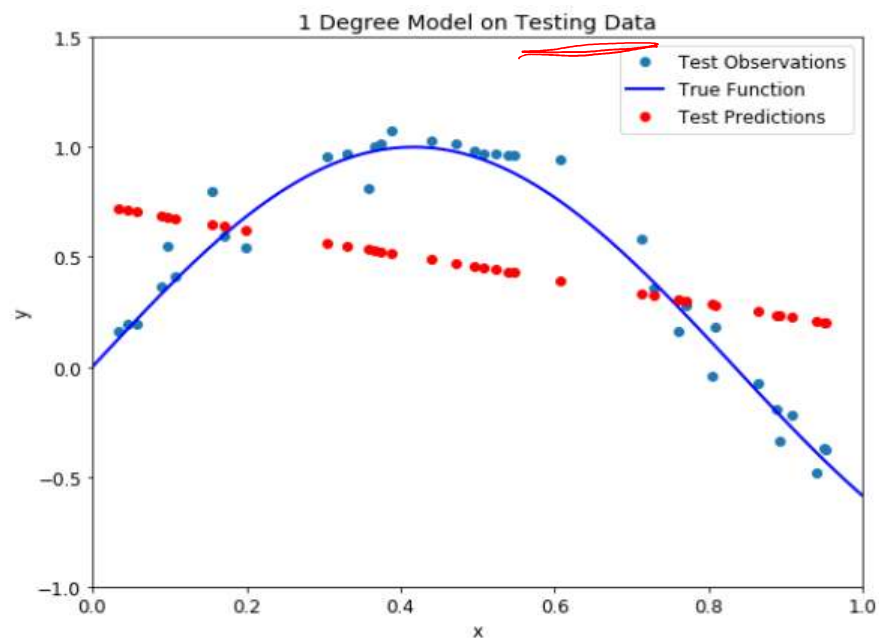
Overfitting

- A model learns relationships between the inputs, called **features**, and outputs, called **labels**, from a training dataset.
- During training the model is given both the features and the labels and learns how to map the former to the latter.
- A trained model is evaluated on a testing set, where we only give it the features and it makes predictions.
- We compare the predictions with the known labels for the testing set to calculate accuracy.
- Overfitting happens because your model is trying too hard to capture the noise in your training dataset

Overfitting vs. Underfitting

- The problem of Overfitting vs Underfitting finally appears when we talk about the polynomial degree.
- The degree represents how much flexibility is in the model, with a higher power allowing the model freedom to hit as many data points as possible.

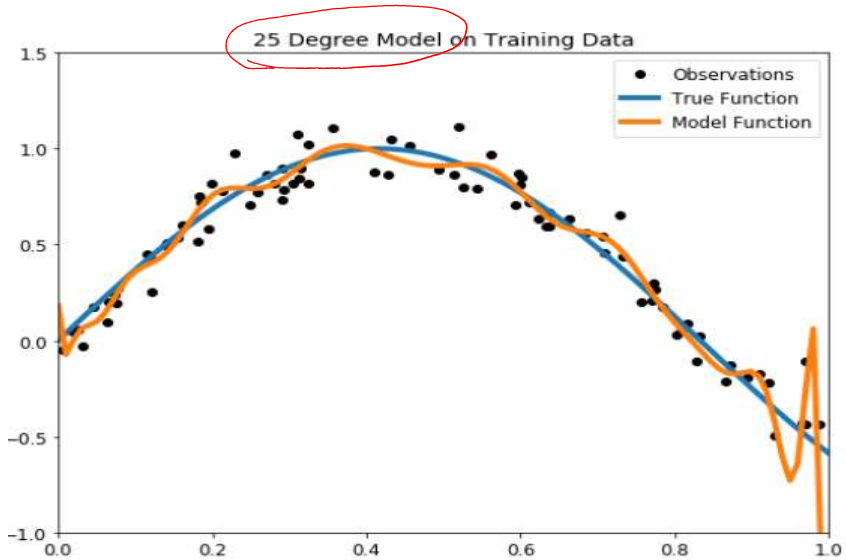


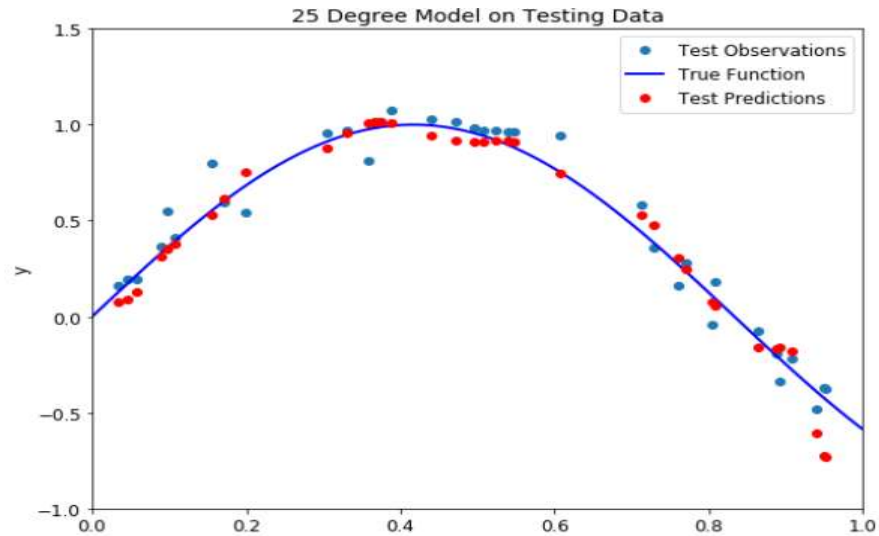


Underfitting Example

- Our model passes straight through the training set with no regard for the data! This is because an underfit model has low variance and high bias.
- Variance refers to how much the model is dependent on the training data.
- For the case of a 1 degree polynomial, the model depends very little on the training data because it barely pays any attention to the points!
- Instead, the model has high bias, which means it makes a strong assumption about the data.
- For this example, the assumption is that the data is linear, which is evidently quite wrong. When the model makes test predictions, the bias leads it to make inaccurate estimates.
- The model failed to learn the relationship between x and y because of this bias, a clear example of **underfitting**.

Overfitting Example





- This is a model with a high variance, because it will change significantly depending on the training data.
- The predictions on the test set are better than the one degree model, but the twenty five degree model still does not learn the relationship because it essentially memorizes the training data and the noise.

Solution

- Our problem is that we want a model that does not “memorize” the training data, but learns the actual relationship!
- How can we find a balanced model with the right polynomial degree?
- If we choose the model with the best score on the training set, we will just select the overfitting model but this cannot generalize well to testing data.
- Fortunately, there is a well-established data science technique for developing the optimal model: **validation**.

Validation

- We need some sort of pre-test to use for model optimization and evaluate. This pre-test is known as a validation set.

- A basic approach would be to use a validation set in addition to the training and testing set.

- This presents a few problems though: we could just end up overfitting to the validation set and we would have less training data.

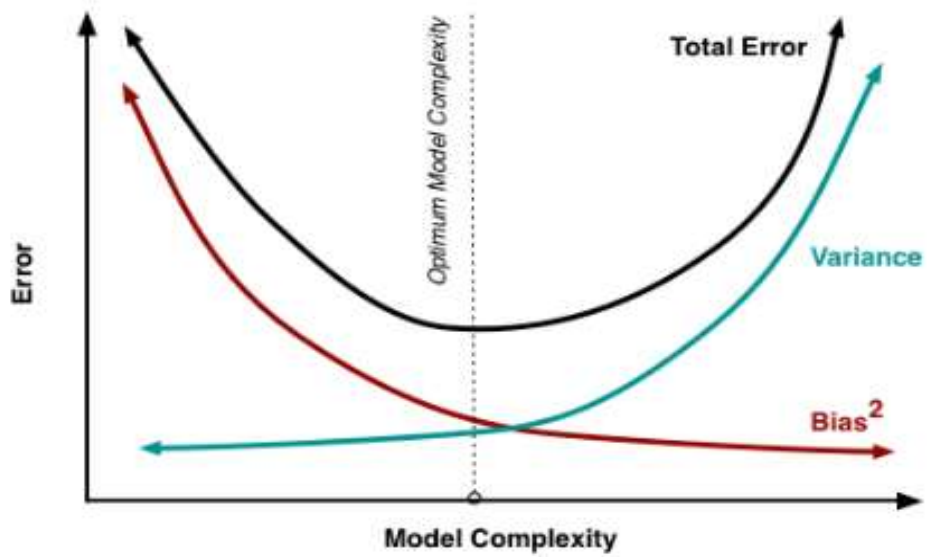
k-fold cross-validation

- A smarter implementation of the validation concept is k-fold cross-validation
- Let's use five folds as an example. We perform a series of train and evaluate cycles where each time we train on 4 of the folds and test on the 5th, called the hold-out set.
- We repeat this cycle 5 times, each time using a different fold for evaluation.
- At the end, we average the scores for each of the folds to determine the overall performance of a given model.
- This allows us to optimize the model before deployment without having to use additional data.

Bias Variance Tradeoff

- To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.
- An optimal balance of bias and variance would never overfit or underfit the model.

Total Error = Bias^2 + Variance + Irreducible Error



Regularization

- This technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.
- A simple relation for linear regression looks like this.

$$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$$

- The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Regularization

- If there is noise in the training data, then the estimated coefficients won't generalize well to the future data.

- This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.

Ridge Regression

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

- *RSS is modified by adding the shrinkage quantity.*
- Now, the coefficients are estimated by minimizing this function.
- The increase in flexibility of a model is represented by increase in its coefficients, and if we want to minimize the above function, then these coefficients need to be small.
- This is how the Ridge regression technique prevents coefficients from rising too high.
- Also, notice that we shrink the estimated association of each variable with the response, except the intercept β0, This intercept is a measure of the mean value of the response when xi1 = xi2 = …= xip = 0.

Lasso/L1 Regression

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

LASSO (Least Absolute Shrinkage Selector Operator), is quite similar to ridge

What does Regularization achieve?

- Regularization, significantly reduces the variance of the model, without substantial increase in its bias.
- So the tuning parameter λ, used in the regularization techniques described above, controls the impact on bias and variance.
- As the value of λ rises, it reduces the value of coefficients and thus reducing the variance.
- Till a point, this increase in λ is beneficial as it is only reducing the variance(hence avoiding overfitting), without loosing any important properties in the data.
- But after certain value, the model starts loosing important properties, giving rise to bias in the model and thus underfitting.
- Therefore, the value of λ should be carefully selected.

Logistic Regression

- Logistic Regression is used when the dependent variable(target) is categorical (binary in nature).
- For example,

  To predict whether an email is spam (1) or (0)

  Whether the tumor is malignant (1) or not (0)

Example: Logistic Regression

- ***E.g.*** When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

- This type of a problem is referred to as **Binomial Logistic Regression**, where the response variable has two values 0 and 1 or pass and fail or true and false.

- **Multinomial Logistic Regression** deals with situations where the response variable can have three or more possible values.

Logistic Regression

Univariate logistic regression equation :

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$$

The left-hand side of this equation is what is called log odds.

Basically, the odds of having an event (P/1-P).

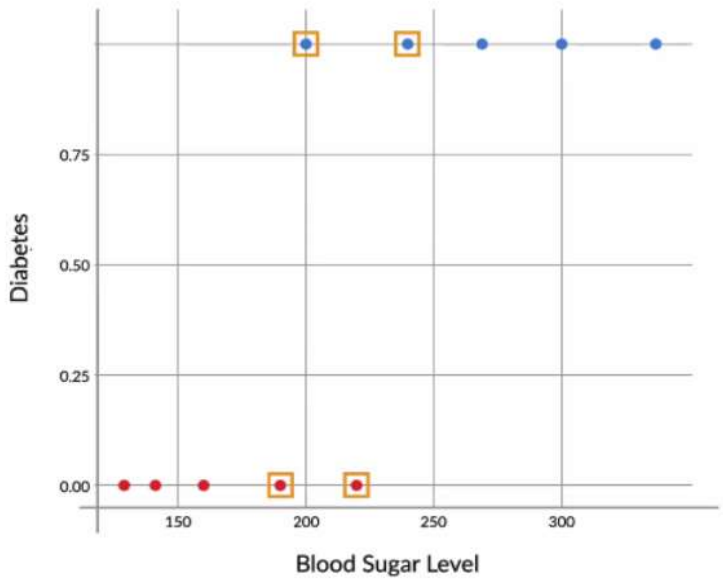
Multivariate logistic regression equation :

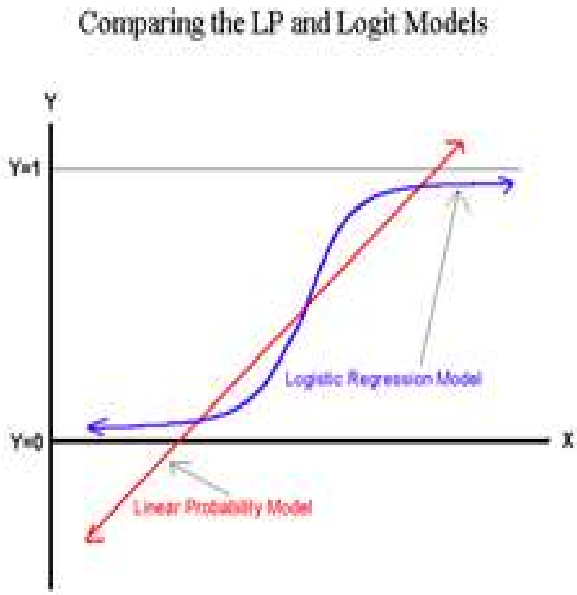
$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots)}}$$

Where P denotes the probability of the event we are trying to predict with multiple independent variables.

Logistic regression example

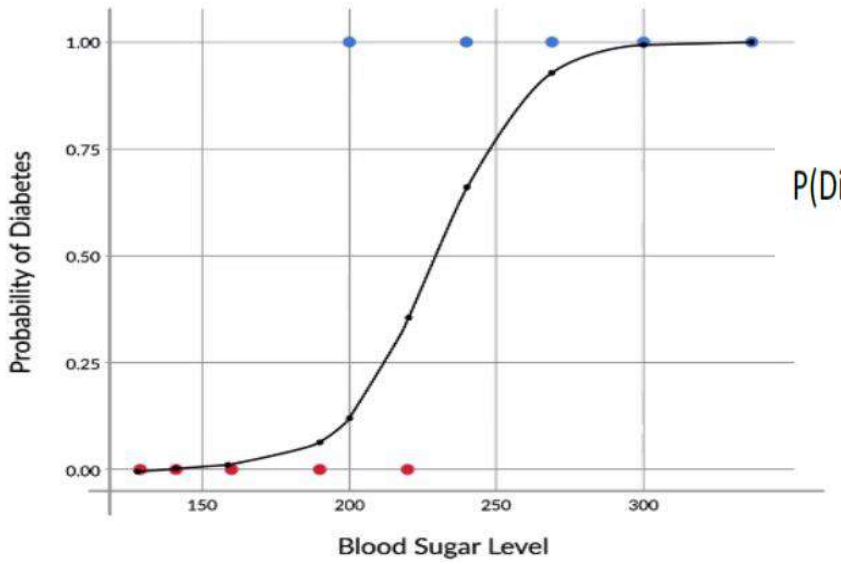
- Consider a case where in we have blood sugar level of 10 patients along-with the status of them being diabetic.





Logistic regression example

One such curve which can model the probability of diabetes very well, is the **sigmoid curve.**



$$P(\text{Diabetes}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

References

- Probability and Statistics for Engineering and Sciences,8th Edition, Jay L Devore, Cengage Learning
- Applied Business Statistics by Ken Black
- https://towardsdatascience.com/let-us-understand-the-correlation-matrix-and-covariance-matrix-d42e6b643c22
- https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765
- https://towardsdatascience.com/multicollinearity-why-is-it-a-problem-398b010b77ac
- https://pdfs.semanticscholar.org/d1ee/9331a2fe0fbb9c8ad27fbf378e3d4cb20163.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3412358/