# Least Squares Regerssion:

Slope

Y intercept

# ANOVA

|  |  |  |
| --- | --- | --- |
| Equation 1: SSR | Regression Sum of Squares: Variance explained by model. |  |
| Equation 2: SSE | Error Sum of Squares: Unexplained variance. |  |
| Equation 3: SST | Total Sum of Squares: Total variance of sample. | = SSR + SSE |

#### Model DF:

Total Number of Variables = **# of model parameters** (Does not include intercept)

Having too many variables can lead to overfitting which in turn leads to unreliable results. The model DF adds a penalty for each predictor variable used in the model. For **Error! Reference source not found.** we have one predictor variable, *x*, so the model DF is 1.

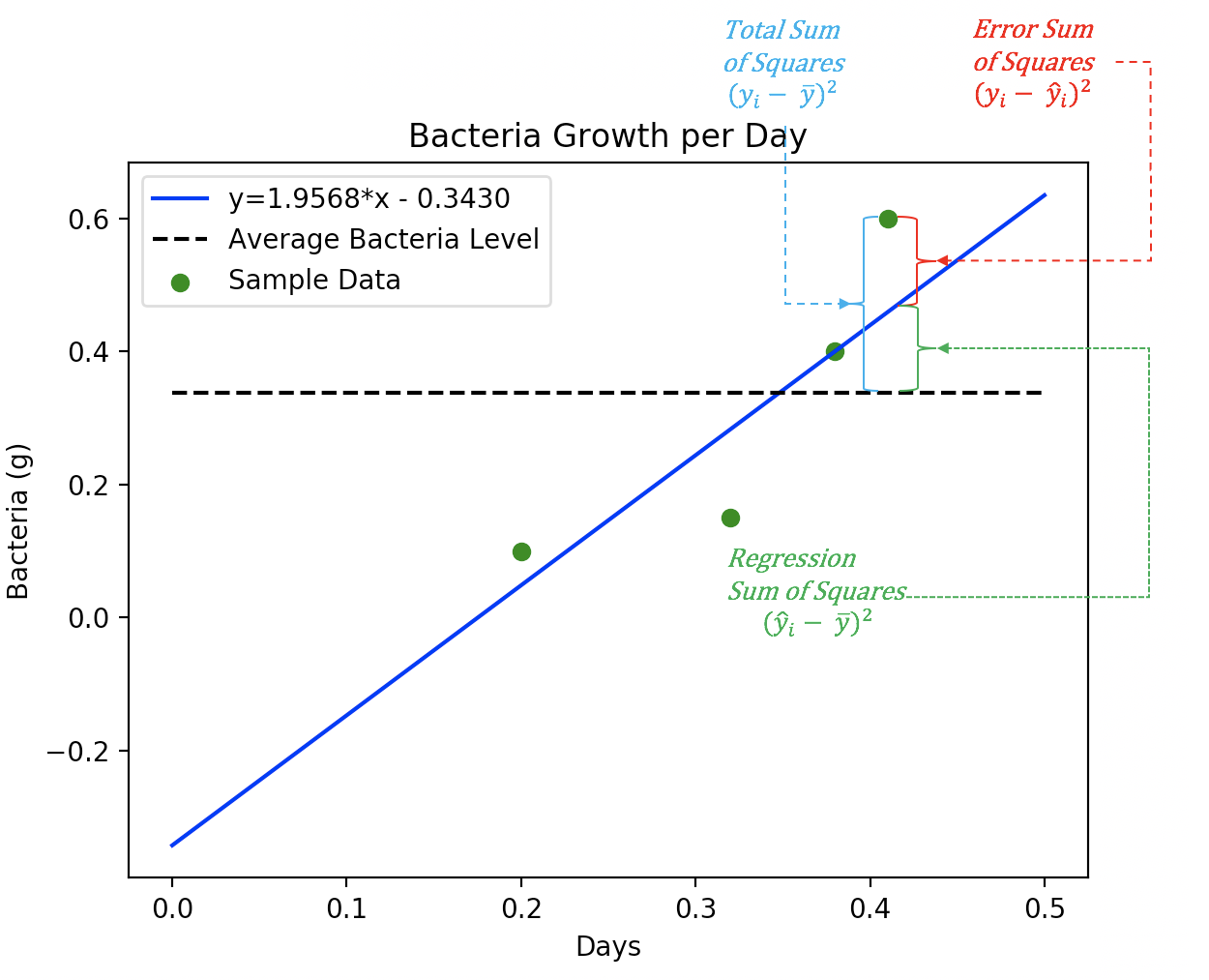
#### Error DF

= **N – Model DF – 1**(Includes the intercept)  
The error DF represents the variance which is not explained by the model.

#### F Statistic

Ratio of error explained by the model, higher is better. Must be greater than value in [table](http://www.socr.ucla.edu/Applets.dir/F_Table.html) for confidence interval to disprove null hypothesis

ANOVA diagram explanation



The difference in y value between the sample data (points on the graph) and the prediction line for any x is due to errors not acounted by our model.

The difference in y value between our prediction line and the mean value for samples (would be a horizontal line, same y value for all x values) for any x is the error accounted by our model.

Ideally we want our model to account for the most of the errors, that would mean our SST, would be mostly due to SSR. If this were the case, our SSE would be very small and for any x value, the difference in y between a sample and our prediction line would be small.

# Model Evaluation(lec2)

#### Coefficient of Determination

Proportion of variablility accounted for by the model.

=

Adjusted coefficient of determination adds penalty for predictor variables. Reduced bias(less overfitting)

p = Number of predictors.  
N = Sample size.

= 1 -

#### RMSE root mean square error

Standard deviation of residuals. Average variation between actual and predicted values. Lower = better.

RMSE = =

### AIC

To help fight bias from over-fitting, the AIC (Akaike Information Criterion) is a measure of validity which combines the log-likelihood score with a penalty for larger numbers of model variables. When comparing models, the desirable model will have a lower AIC score. The AIC does not use an absolute score.

**AIC = -2\*Log-likelihood(LL) + 2\*k**

k is number of predictor variables + 1 for the intercept

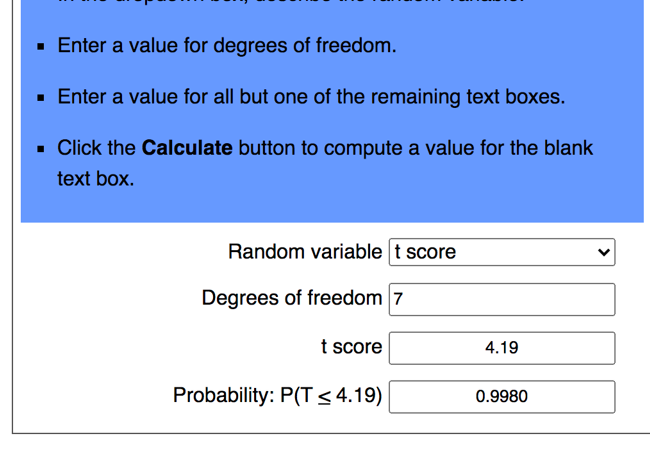
### BIC

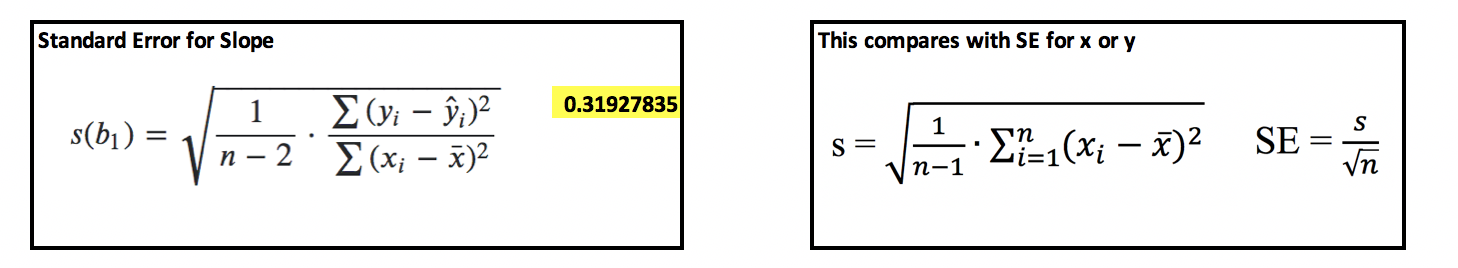
The Bayesian Information Criterion is used to evaluate models based on the likelihood function. The model with the lowest BIC is preferred. Similar to the AIC score, the BIC score also does not use an absolute value. The BIC score is based on the likelihood function and the AIC. BIC has a higher penalty for total number of predictor variables than the AIC.

k is number of predictor variables + 1 for the intercept.

**BIC = -2 \* LL + log(N) \* k**

# Hypothesis testing

t = = = 4.185266457680251  
  
p > |t| = 2\*(1-0.998) = 2\*0.002 = 0.004



# Logistic Regression

validating logistic regression:

precision  
 relevant results vs all results received

recall  
 ratio of positives found vs all positives  
=

### F1-Score

A high F-score is as a weighted mean of the precision and recall scores. The F-score is at its best value at 1 and worst at 0.

*F1*

# Scaling

Without scaling the weighting of various features during model building and use may be disproportionate to other features. You will sometimes notice that improvements can be made to your logistic regression and linear regression models if the data is scaled.

### Comparing MinMax, Standard and Robust Scalers

* Use MinMaxScaler as the default if you are transforming a feature. It is non-distorting.
* Use RobustScaler if you have outliers and want to reduce their influence. However, you might be better off removing the outliers, instead.
* Use StandardScaler if you need a relatively normal distribution.

#### MinMax Scaler

MinMaxScaler preserves the shape of the original distribution. It does not meaningfully change the information embedded in the original data. MinMaxScaler essentially restricts the range between 0 and 1 (or -1 to 1 if there are negative values). If the distribution is not Gaussian (normal) or the standard deviation is very small, the min-max scaler is usually recommended.

The equation for MinMaxScaler is:

#### StandardScaler

Deep learning algorithms often call for zero mean and unit variance. Regression-type algorithms also benefit from normally distributed data with small sample sizes. The StandardScaler algorithm outputs something very close to a normal distribution by changing the values so the distribution standard deviation from the mean equals one. StandardScaler does distort the relative distances between the feature values which may not be desirable in many cases.

Example 3: StandardScaler

The StandardScaler() class calculates the standard deviation with zero degrees of freedom instead of using the usual value of 1. The sd equation becomes:

=

Scaled value =

#### RobustScaler

RobustScaler transforms the feature vector by subtracting the median and then dividing by the interquartile range (75% value — 25% value). RobustScaler users less data so it reduces the influence of outliers.

RobustScaler = =

# Binning(lec3):

Binning can make small variables more significant. It can also help understand patterns in these variables, making it easier to anaylyze and interpret.

# Dummy Variables(lec3):

Dummy variables allow the regression algorithm to correctly analyze attribute variables.  
Without dummy variables, a regeression model would not be able to correctly anaylyze categorical  
since any value given to a category is not directly comparable to a value given to another category.  
category 1 is not half as much as category 2.

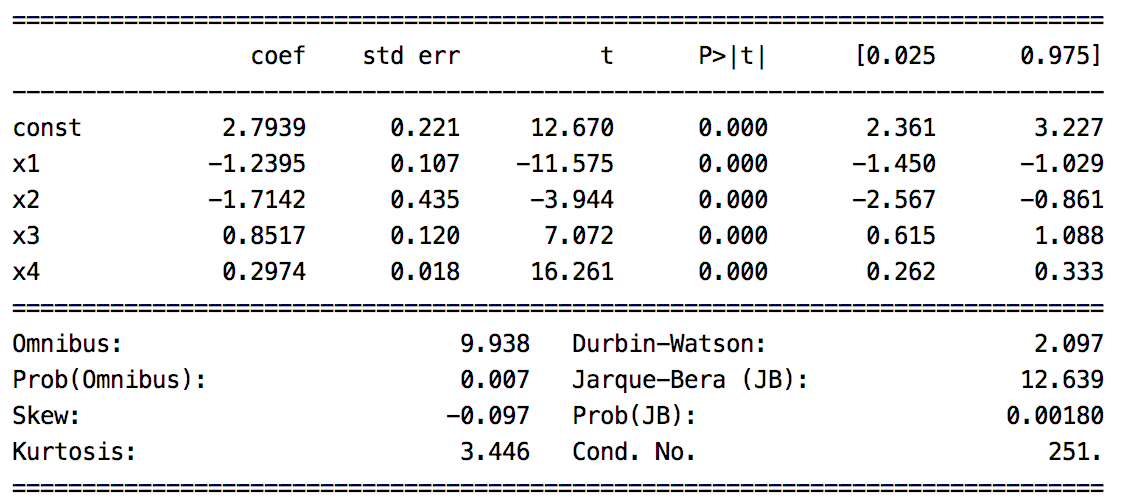
Dummy variables ussually take the values 0, or 1 (like a switch) indicating the presence   
of an attribute in a sample.   
  
You need k-1 dummy variables, where k is the number of types of category.  
Example: working, in-school, unemployed. You would only need dummy variables for working and in-school and if a sample is neither, it is then unemployed.

# Inputing:

Imputing can increase the performance of a model, but it can also have risks. There are many ways to impute data and each technique has positives and negatives. Imputing can potentially reduce the vairance in a model and increase its bias, it can also potentially distort the relationships between variables. Using more than one method to impute data can help with the issues of impyuting, by also adding some variance in the model. Additonally, before imputing one should know why there is data, often missing values might be hiding some additonal information about the data (values may be missing for a reason).

# Outliers

Removing outliers may introduce bias into our model. But keeping them may also skew the results of the model for more normal values

Regression Output:  
  


#### Durbin Watson Test (Error residual independence)

The Durbin Watson test indicates whether or not the error residuals are independent of each other. Ideally the model has accounted for as much explainable variance as possible so the residual errors are independent of each other. Table 1 explains how to interpret the different Durbin-Watson ranges. A Durbin-Watson value around 2 is desirable.

Table 1: Ranges for Durbin-Watson

|  |  |
| --- | --- |
| **Range** | **Interpretation** |
| 0 to 2 | A positive correlation exists among residuals. |
| Around 2  *(Around 2 is ideal)* | Error residuals are independent (which is preferred). |
| 2 to 4 | A negative autocorrelation |

#### Jarque-Bera (Model normality)

The Jarque–Bera test is a goodness-of-fit test of whether a model has the skewness and kurtosis matching a normal distribution. The null hypothesis, , suggests that the distribution is normal where the Kurtosis, or height, is approximately 3 and skewness is 0. When the p-value > 0.05 accept the null hypothesis. With samples below 2000, JB is prone to rejecting the null hypothesis. For the case presented in **Error! Reference source not found.**, normality is rejected.

### Conditional Number (Collinearity)

The conditional number measures likelihood for collinearity or unwanted interdependence between predictor variables. Ideally predictor variables are independent. High interdependence means higher variance in results from our model.

Linear regression models are prone to higher amounts of collinearity. If the models have high collinearity this should be documented and explained to a subject matter expert who is familiar with the model variables and the goals of the model. Sometimes high collinearity is acceptable. For example, SAT and LSAT scores on college entrance exams are highly correlated. Both indicate that the applicant is prepared for a rigorous academic program. Collinearity, could be high for these predictor variables yet we may still want both variables in the model.

Other times, high collinearity may be unacceptable. For example, a stock portfolio may rely on a complex relationship between sectors. A portfolio model may be incorrectly constructed under the assumption that oil and housing prices rise together yet an unexpected change in one of these sectors could have a detrimental effect on the prediction.

### Omnibus Test (Normality Goodness of Fit)

The Omnibus tests examines the model for goodness of fit with a normal distribution.

It is normally distributed.

Since the p-value is small in **Error! Reference source not found.** we would reject the null hypothesis.

This section suggests ways that you may impute content. You may create other methods to impute values as well or you may build a method for imputing that uses a hybrid of techniques.

For now, you may keep your imputing methods simple but they could become very advanced if needed.

Often imputing techniques do not offer much lift in the regression results. On the other hand, I have occasionally experienced significant performance gains through imputing. In any case, the topic is a necessary one and it can make a positive difference when done properly.

### Imputing by Mean, Median or Mode

For linear regression algorithms you can impute missing values by providing the sample mean, median or mode. This technique can be implemented quickly but it also reduces the model variance. Be wary that bias and inaccuracy may result.

### Regression imputation

Missing values can be predicted with simple OLS models based on other variables. This preserves relationships among variables in the imputation model. This technique can become complex if predictor variables are also missing data.

### Hot deck imputation

Hot deck imputation involves selecting the missing value from a sample has that similar values for other variables. Or, compute the average value from a random selection of similar samples.

### Cold deck imputation

Cold deck imputation is a systematically chosen value from a sample that has similar values on other variables. Cold deck imputation is similar to hot deck imputation but it removes the random variation. For example, you may always choose the third individual in the same experimental condition and block.

### Multiple imputation

You can use more than one method to impute values. Because these methods have a random component, the multiple estimates are slightly different. This re-introduces some variation that your software can incorporate in order to give your model accurate estimates of standard error. Multiple imputation solves a lot of problems with missing data (though, unfortunately not all) and if done well, leads to unbiased parameter estimates and accurate standard errors. You may choose to average or weight the results to generate a single result based on multiple methods.

# Analyzing Data