Predictive Modeling
Multiple Linear Regression

Objectives

01 Multiple Linear Regression

02 Testing overall model

03 Accessing impact of individual variables

04 Adjusted R-Squared

Objectives

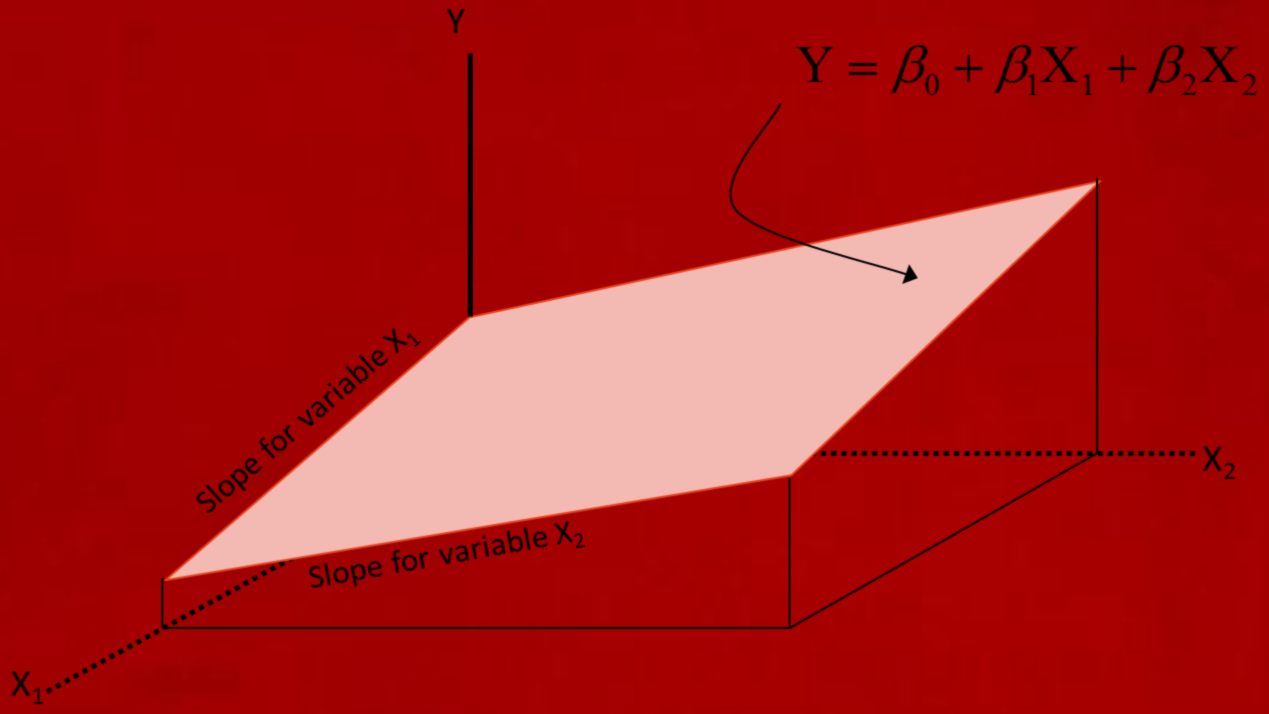
05 | Multicollinearity

06 | Model diagnostics

07 | Conclusion

Multiple Regression

- Using multiple predictor variables instead of single variable
- We need to find a perfect plane here



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Testing Overall Model

- For testing overall model, we perform following hypothesis test (ANOVA):

To test : H_0: $\beta_1 = \beta_2 \; \ldots \ldots \ldots = \beta_p = 0$
H_a: $Atleast\ on\ of\ the\ \beta_i s\ is\ non\ zero.$; for all i = 1, 2, …..,p

- Test statistic follows F- distribution.

Test statistic: $F = \frac{SS_{Reg}/dfReg}{SS_{Err}/dfErr} = \frac{SS_{Reg}/k}{SS_{Err}/(n-k-1)_r} = \frac{MS_{Reg}}{MS_{Err}}$

- Observe the P-value.
- Reject H_0 if P-value is less than level of significance.

Note: When H_0 is true, there is no relationship between target and predictor variables. So MLR model is not a valid model.

Individual Impact of variables

- If ANOVA tells us the MLR model is possible, we can identify which all variables are important in order to predict target variable. For this we perform t – test for all the individual predictor variables.

To test : $H_0\colon\ \beta_i = 0$

$H_a\colon\ \beta_i \neq 0$; for all i = 1, 2, …..,p

- Individual variable coefficient is tested for significance
- Beta coefficients follow t distribution.

Test statistic: $t = \dfrac{\beta_i}{s(\beta_i)}$

- Individual P values tell us about the significance of each variable
- A variable is significant if P value is less than 5%. Lesser the P-value, better the variable

Reject H$_0$ if: $t > t(\frac{\alpha}{2}; n-k-1) \quad or$

$t < -t(\frac{\alpha}{2}; n-k-1)$

- Note it is possible all the variables in a regression to produce great individual fits, and yet very few of the variables be individually significant.

Individual Impact of variables

- A variable is significant if P value is less than 5%.
- Lesser the P-value, better the variable
- If a variable has p-value less than 5%, if we drop that variable then we may see a drop in R-Squared value
- If a variable has p-value greater than 5%, if we drop that variable then we may not see any significant change in R-Squared value

Adjusted R-Squared

- The training set MSE is generally an underestimate of the test MSE. (Recall that MSE = RSS/*n*.)
- This is because when we fit a model to the training data using least squares, we specifically estimate the regression coefficients such that the training SSE (but not the test SSE) is as small as possible.
- In particular, the training error will decrease as more variables are included in the model, but the test error may not.
- Therefore, training set SSE and training set *R*_square can not be used to select from among a set of models with different numbers of variables.
- However, a number of techniques for *adjusting* the training error for the model size are available. These approaches can be used to select among a set of models with different numbers of variables.
- We will discuss Adjusted R_squared value.

Adjusted R-Squared

- Is it good to have as many independent variables as possible? Nope
- R-square is deceptive. R-squared never decreases when a new X variable is added to the model – True?
- We need a better measure or an adjustment to the original R-squared formula.

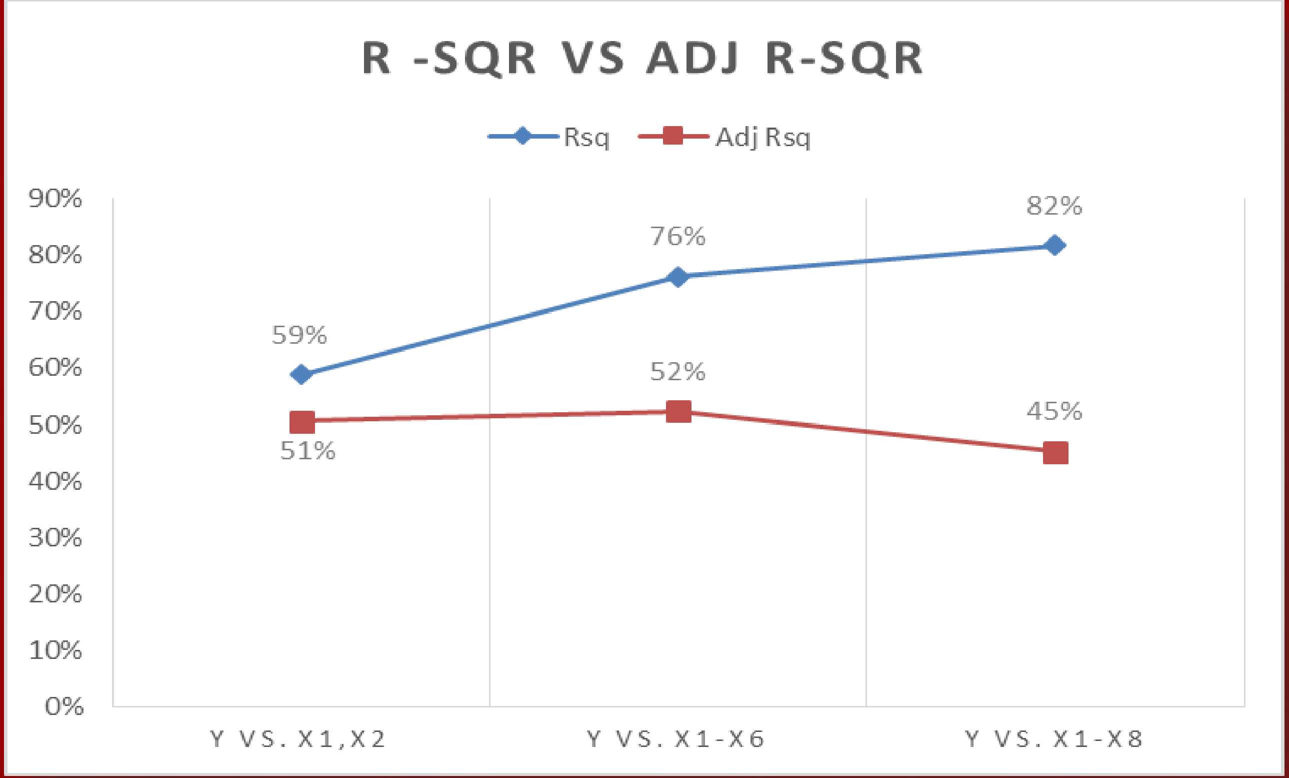
Adjusted R-Squared

- Its value depends on the number of explanatory variables
- Imposes a penalty for adding additional explanatory variables
- It is usually written as (R-bar squared)
- Very different from R when there are too many predictors and n is less

$$\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

n-number of observations, k-number of parameters

R-Squared vs Adjusted R-Squared



Validation and Cross-Validation

- As an alternative to the approaches just discussed, we can directly estimate the test error using the validation set and cross-validation methods. We can compute the validation set error or the cross-validation error for each model under consideration, and then select the model for which the resulting estimated test error is smallest.

- This procedure has an advantage relative to metrics like adjusted *R_squared*, in that it provides a direct estimate of the test error and makes fewer assumptions about the true underlying model. It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance.

Multiple Regression- issues

Multiple regression is wonderful - In that it allows you to consider the effect of multiple variables simultaneously. But when we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

1. *Non-linearity of the response-predictor relationships.*
2. *Correlation of error terms.*
3. *Non-constant variance of error terms.*
4. *Outliers.*
5. *High-leverage points.*
6. *Collinearity.*

Multicollinearity

- **Multicollinearity** is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.
- It is caused by an inaccurate use of dummy variables, by the inclusion of a variable which is computed from other variables in the data set.
- Multicollinearity can also result from the repetition of the same kind of variable.
- Generally, occurs when the variables are highly correlated to each other.
- The parameter estimates will have inflated variance in presence of multicollinearity.
- Sometimes the signs of the parameter estimates tend to change.
- If the relation between the independent variables grows really strong then the variance of parameter estimates tends to be infinity.

Multicollinearity Detection

- Several methods offer an approach to this problem.
- One way is to examine a correlation matrix to search for possible inter-correlations among potential predictor variables.
- If several variables are highly correlated, the researcher can select the variable that is most correlated to the dependent variable and use that variable to represent the others in the analysis.
- One problem with this idea is that correlations can be more complex than simple correlation among variables.
- In other words, simple correlation values do not always reveal multiple correlation between variables. In some instances, variables may not appear to be correlated as pairs, but one variable is a linear combination of several other variables.
- This situation is also an example of multicollinearity, and a cursory observation of the correlation matrix will probably not reveal the problem.

Multicollinearity Detection

- Stepwise regression is another way to prevent the problem of multicollinearity. The search process enters the variables one at a time and compares the new variable to those in solution.
- If a new variable is entered and the *p* values on old variables become nonsignificant, the old variables are dropped out of solution.
- In this manner, it is more difficult for the problem of multicollinearity to affect the regression analysis. Of course, because of multicollinearity, some important predictors may not enter into the analysis.

Multicollinearity Detection

- Other techniques are available to attempt to control for the problem of multicollinearity. One is called a **variance inflation factor (VIF)**, in which a regression analysis is conducted to predict an independent variable by the other independent variables.
- In this case, the independent variable being predicted becomes the dependent variable. As this process is done for each of the independent variables, it is possible to determine whether any of the independent variables are a function of the other independent variables, yielding evidence of multicollinearity.

Multicollinearity Detection

- $1/(1-R^2)$ is called VIF.
- We can calculate VIF values for each of the predictor variables. (using software packages)

Multicollinearity Detection

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

- Build a model X1 vs X2 + X3 + X4 find R square, say R_1^2
- Build a model X2 vs X1 + X3 + X4 find R square, say R_2^2
- Build a model X3 vs X1 + X2 + X4 find R square, say R_3^2
- Build a model X4 vs X1 + X2 + X3 find R square, say R_4^2
- Finally, $VIF(X_j) = 1 / (1 - R_j^2)$ for j = 1,2,3,4

- For example, if R_3^2 is 95% then we don't really need X3 in the model
- Since it can be explained as liner combination of other three variables
- For each variable we find individual R square.

Model Diagnostics

Residual Analysis:

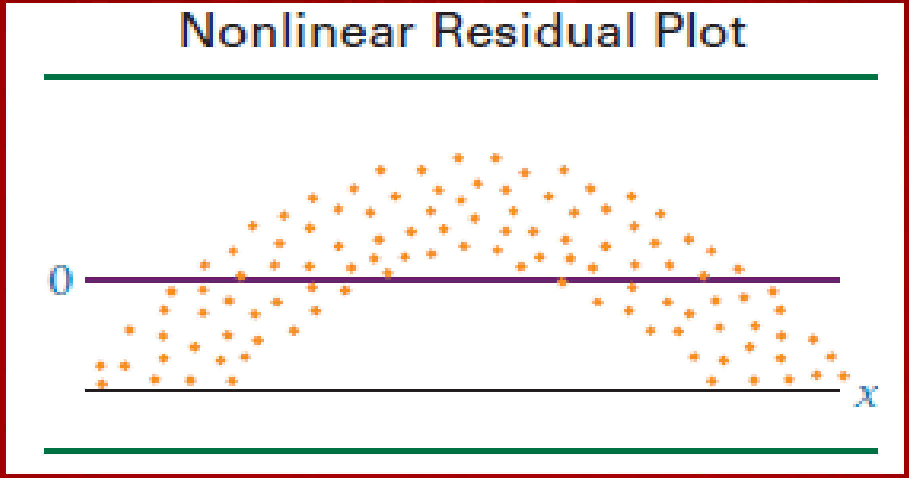
It uses Residual Plot in which Residuals are usually plotted against the fitted values (y_hat).

We can use residuals to study whether:

- The regression function is nonlinear.
- The error terms have non constant variance.
- The error terms are not independent.
- There are outliers.
- The error terms are not normally distributed.

Model Diagnostics

- If a residual plot such as the one in Figure attached appears, the assumption that the model is linear does not hold.

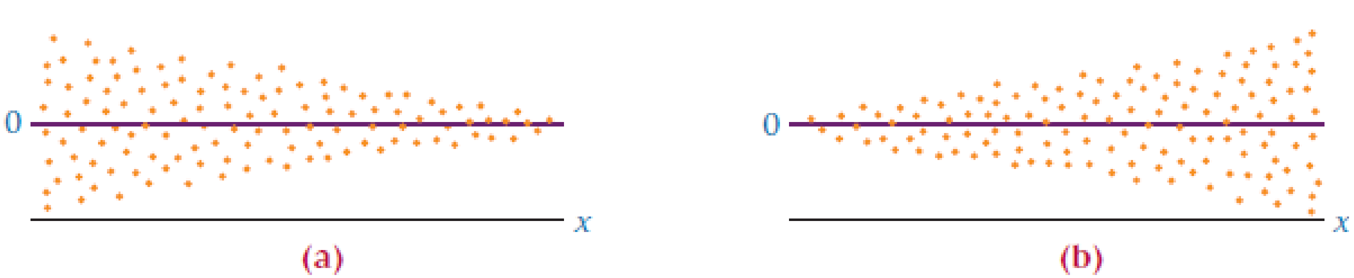


- Note that the residuals are negative for low and high values of *x* and are positive for middle values of *x*. The graph of these residuals is parabolic, not linear.
- Any significant deviation from an approximately linear residual plot may mean that a nonlinear relationship exists between the two variables.

Model Diagnostics

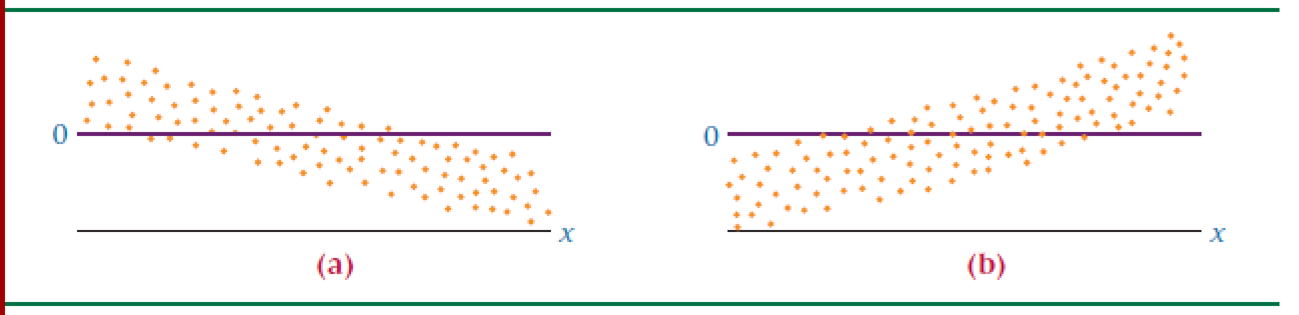
- The assumption of *constant error variance* sometimes is called **homoscedasticity**.
- If *the error variances are not constant* (called **heteroscedasticity**), the residual plots might look like one of the two plots in Figure below.
- Note in Figure (a) that the error variance is greater for small values of *x* and smaller for large values of *x*. The situation is reversed in Figure (b).

Nonconstant Error Variance



Model Diagnostics

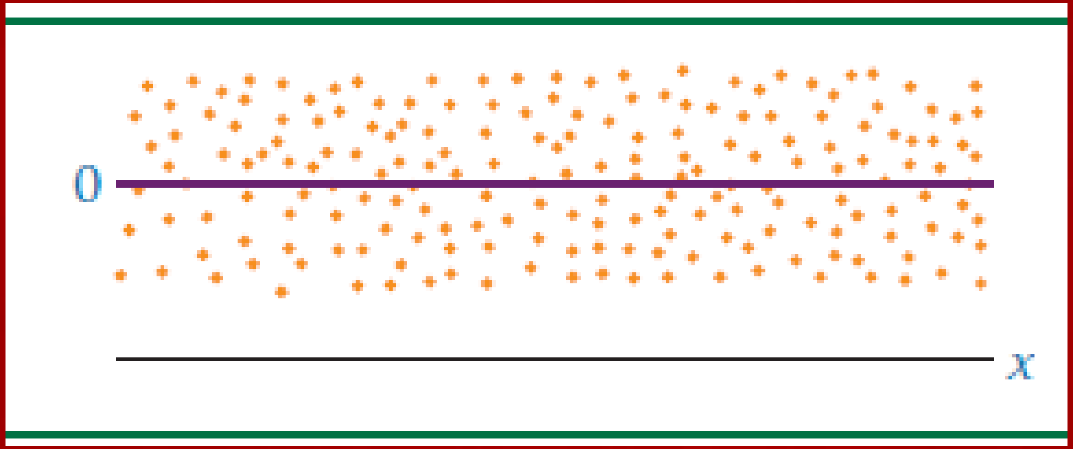
- If the error terms are not independent, the residual plots could look like one of the graphs given below.



- According to these graphs, instead of each error term being independent of the one next to it, the value of the residual is a function of the residual value next to it. For example, a large positive residual is next to a large positive residual and a small negative residual is next to a small negative residual

Model Diagnostics

- The graph of the residuals from a regression analysis that meets the assumptions—a *healthy residual graph*—might look like the graph as given in figure below.
- The plot is relatively linear; the variances of the errors are about equal for each value of *x*, and the error terms do not appear to be related to adjacent terms.



Model Diagnostics

For validating normality condition:

we can use Q-Q plot and/or any of the following non-parametric tests:

- Kolmogorov Smirnov Test
- Shapiro Wilk Test
- Anderson Darling Test

Some tips

1. Plot and examine the data.
2. If necessary, transform the X and/or Y variables so that:
   - the relationship between X and Y is linear, and
   - Y is homoskedastic (that is, the scatter in Y is constant from one end of the X data to the other).
3. Use VIF method to remove multicollinearity.
4. Plot the residuals plot.
   - If the residuals increase or decrease with X, they are heteroscedastic. Transform Y to cure this.
   - If the residuals are curved with X, the relationship between X and Y is nonlinear. Either transform X or fit a nonlinear curve to the data.
   - If there are outliers, check their validity, and/or use robust regression techniques.
5. Plot the distribution of the residuals (either as a histogram, or a normal quantile plot).
   - If the residuals are not normally distributed, your estimates of statistical significance and confidence intervals will not be accurate.

Conclusion - Regression

- Try adding the polynomial & interaction terms to your regression line. Sometimes they work like a charm.
- Adjusted R-squared is a good measure of training/in time sample error. We can’t be sure about the final model performance based on this. We may have to perform cross-validation to get an idea on testing error.
- Outliers can influence the regression line; we need to take care of data sanitization before building the regression line.
- We need to detect and treat multicollinearity issue.

Thank You

Hero Vired