

Lesson 10: The Granddaddy of All Models: Regression Analysis

Version 2024-25T1.1

Instructor: Dr. Kam Tin Seong
Associate Professor of Information Systems (Practice)
School of Computing and Information Systems
Singapore Management University

Welcome to Lesson 10: The Granddaddy of All Models: Regression Analysis. In this lesson, I am going to share with you how to build explanatory models by using Linear Regression methods. I will focus my discussion in explaining the basic concepts and methods of linear regression models. I will also explain to you various methods for finding a meaningful explanatory model.

What will I learn in
this lesson?

Learning Outcomes

Upon completion of this lesson, student will be able to:

- understand what regression analysis is and what it can be used for.
- understand the basic concepts, assumptions and methods of linear regression methods
- interpret the analysis results accurately
- validate regression analysis results
- compare and evaluate explanatory analytics models generated by different MLR methods

Why Regression Analysis?



I've used regression extensively and love it for all of its flexibility. You can use:

- multiple independent variables
- continuous and categorical variables
- higher-order terms to model curvature
- interaction terms to see if the effect of one predictor depends upon the value of another

That's all cool stuff. But the list leaves out an almost magical property of regression analysis. Regression has the ability to disentangle some very convoluted problems. Problems where the explanatory variables seem enmeshed together like spaghetti. Suppose you're a researcher and you are studying a question that involves intertwined explanatory variables. For example, you want to determine:

- whether socio-economic status or race has a larger effect on educational achievement
- the importance of education versus IQ on earnings
- how exercise habits and diet effect weight
- how drinking coffee and smoking cigarettes are related to heart disease
- if a specific exercise intervention (separate from overall activity levels) increases bone density

These are all research questions where the explanatory variables are likely to be correlated with each other and they could all influence the dependent variable. How

do you untangle this web and separate out the effects? How do you determine which variables are significant and how large of a role does each one play?
Regression comes to the rescue!

Condominium resale case

The Analytics task

- Build a model to discover factors affecting resale prices of condominium by taking into consideration a set of geographically referenced explanatory variables.

Condominium Resale Prices Case Study

Setting the Scene

In this case study, you are tasked to build a hedonic pricing model to determine factors affecting resale prices of condominiums.

An introduction to regression

- A method of predicting the value of a dependent variable (y) based on one or more independent variables (x)
 - Does X affect Y? If so, how?
 - What is the change in Y given a one unit change in X?
- Assume a linear relationship between x and y, fit a straight line to data points that minimizes total error

In order to provide an answer to the question, regression analysis will be used.

A regression model is a mathematical model that explains and predicts a continuous response variable. In our analysis, a regression model will be developed to explain the resale value of a condominium based on the floor area of the unit. In this case, the **response variable** or more popularly known as the **dependent variable** is the resale value. The floor area of the condominium unit, on the other hand, is called **independent variable** or **explanatory variable**.

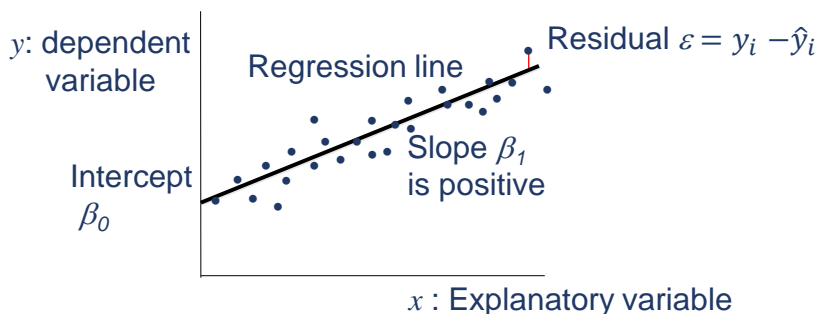
Regression analysis aims to determine the best fit line that explains the relationship between the dependent variable and independent variable(s). This is the **explanatory** role of a regression model.

The regression model developed can then be used as a **prescriptive model**. For example, when a potential seller walks into our real estate company, we can use the regression model to estimate the possible re-sale value if we know the floor area of the resale condominium unit owned by the potential seller. In this lesson, however, we will only focus on building explanatory models.

A Simple Linear Regression Model



- The formula $y_i = \beta_0 + \beta_1 x_i + \varepsilon$



In statistics, a **simple linear regression** is the **best fit** line of a linear regression model with a **single explanatory variable**.

The adjective **simple** refers to the fact that this regression is one of the simplest in statistics. A simple regression consists of one dependent variable and one independent variable.

The y is the **dependent variable**. It is the variable representing the process you are trying to understand or predict (i.e. re-sale value). While you can use regression to explain the dependent variable, you always start with a set of known y values and use these to build (or to calibrate) the regression model. The known y values are often referred to as **observed values**.

The x is the **independent variable**. It is the variable used to model or predict the dependent variable values. In the Toyota Corolla case study, the independent variable can be the age of the car.

For a simple linear regression, the dependent and independent variables must be in **continuous** data type.

The **intercept** (β_0) of the fitted line is such that it crosses the vertical axis of the graph. It represents the expected value for the dependent variable if the independent variable is zero.

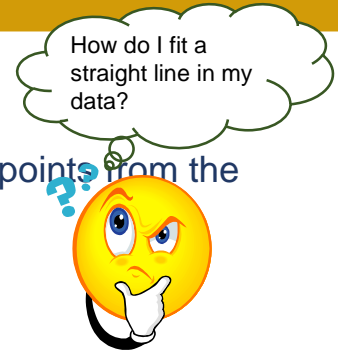
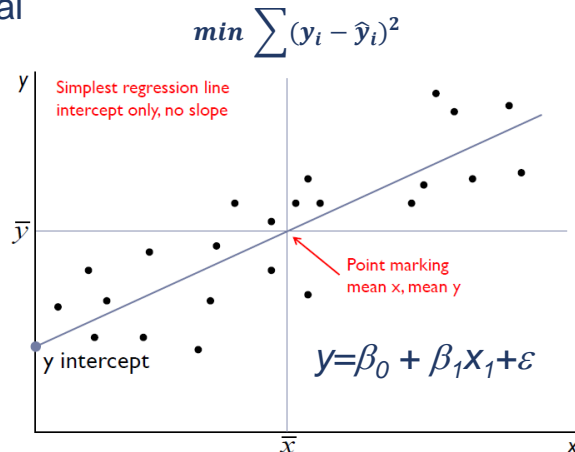
The **slope** (β_1) of the fitted line is equal to the correlation between y and x corrected by the ratio of standard deviations of these variables. It represents the strength and type of relationship the explanatory variable has to the dependent variable. When the relationship is positive, the sign for the parameter is also positive. On the other hand, a β_1 with negative sign represents negative relationship. When the relationship is a strong one, the value of β_1 is large. When relationship is weak, value of β_1 is near zero.

These parameters: β_0 and β_1 are also known as the **regression coefficients**.

The **residual** ε represents the differences between the values of the dependent variable predicted by the simple regression line and their corresponding observed values. It is also popularly known as the **random error term**. It is worth knowing that this term represents the fact that a regression model will not fit the data collected perfectly.

The Least Squares Method

- The sum of the vertical deviations (y axis) of the points from the line is minimal



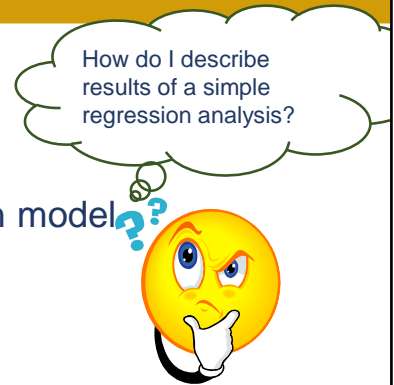
A simple linear regression fits a straight line through the set of n points in such a way that makes the **sum of squared residuals** of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible. This method is known as the **least square method**.

The way we determine which line (there are an infinite number of potential lines) is the best fit is easy.


- We define a line that passes through the point determined by the mean x value and the mean y value
- The slope of this line needs to minimize the residual error

Interpreting a simple regression

- Describe the regression model
- Describe the overall fit of the simple regression model
 - Summary of Fit report
 - Analysis of Variance
- Describe the model parameters
 - Parameter estimates report



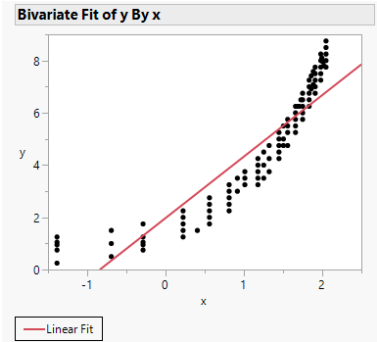
In this section, you will learn how to interpret a simple linear regression. This includes the overall simple linear regression model, the overall fit and the parameters of the model.



School of
Computing and
Information Systems

Visualising a simple regression model

- Statistical significance does not mean **linear relationship** between an explanatory variable (i.e. x) and a dependent variable (i.e. y).



Linear Fit

$y = 1.9759645 + 2.3523873x$

Summary of Fit

RSquare	0.819006
RSquare Adj	0.817159
Root Mean Square Error	1.015062
Mean of Response	4.621
Observations (or Sum Wgts)	100

Analysis of Variance

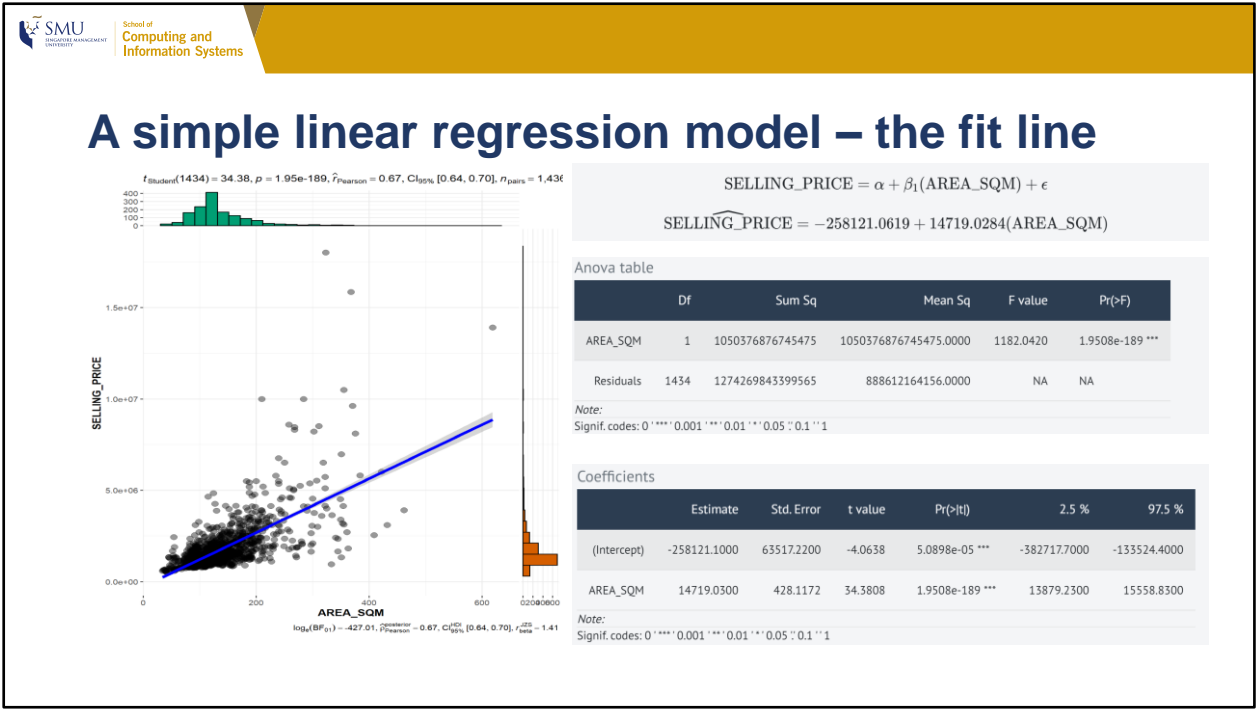
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	456.91423	456.914	443.4547
Error	98	100.97444	1.030	Prob > F
C. Total	99	557.88868		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	1.9759645	0.161493	12.24	<.0001*
x	2.3523873	0.111708	21.06	<.0001*

Cautions about interpreting a simple linear regression

It is important for us to visualise the relationship between a dependent variable and a predictor variable in a simple regression model. This is because a simple regression model might be statistically significant, but it does not mean a linear relationship exist between the dependent variable and the explanatory variable. The scatterplot in the figure above shows that there is a relationship between the x- and y-variables but the relationship is not linear.



The scatterplot in the figure above shows that there is a direct linear relationship between the re-sale prices of condominium units (dependent variable) and the floor area of the re-sale units, the predictor (also known as independent variable). Notice that a regression line has been fitted in the scatter plot and the following reports have been generated:


- Linear Fit
- Summary of Fit
- Lack of Fit
- Analysis of Variance
- Parameter Estimates

We will discuss how to interpret the report in later. First, you will describe the model. The reports reveal that the relationship between re-sale price and floor area of the condominium re-sale units can be defined as follow:

$SELLING_{PRICE} = -258121.0619 + 14719.0284(AREA_SQM)$

The positive slope value indicates that the re-sale price of a condominium unit is

directly proportional to the floor area of the unit. More specifically, the model shows that any single unit increase of the floor area (i.e. in terms of psm), the price of the re-sale unit will increase by S\$14,719.0284.



School of
Computing and
Information Systems

Overall Model Fit

- Sums of squares, R and R^2

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

Total sum of squares (TSS)


Regression (explained) sum of squares (ESS)

Residual (unexplained) sum of squares (RSS)

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

The proportion of the total explained variation in y is called the coefficient of determination (R^2)

How well the model fits the data?




We can assess the overall model fit using the R^2 and significance of the F-value.

The R^2 (or coefficient of determination) indicates the degree to which the model explains the observed variation in the dependent variable, relative to the mean. A good regression line should explain a substantial amount of variation (have a high ESS) relative to the total variation (TSS).

The R^2 always lies between 0 and 1, where a higher R^2 indicates a better model fit. When interpreting the R^2 , higher values indicate that more of the variation in y is explained by variation in x , and therefore that the ESS is high relative to the RSS.

A discussion on adjusted R^2 will be given in the section on Multiple Linear Regression.



School of
Computing and
Information Systems

Describing the overall fit

Model Summary			
R	0.672	RMSE	942005.596
R-Squared	0.452	MSE	888612164155.903
Adj. R-Squared	0.451	Coef. Var	53.829
Pred R-Squared	0.445	AIC	43587.753
MAE	530082.245	SBC	43603.562

RMSE: Root Mean Square Error

MSE: Mean Square Error

MAE: Mean Absolute Error

AIC: Akaike Information Criteria

SBC: Schwarz Bayesian Criteria

The R^2 and Adjusted R^2 of the model are 0.452 and 0.451 respectively. The value of R^2 tells us that the floor area of the re-sale condominium can account for 45.2% of the variation in the re-sale price of the condominium. In other words, if we are trying to explain why the re-sale values of some condominium are higher than others, we can look at the variation in floor area of the units. There might be many factors that can explain this variation, but our model, which include only floor area of the condominium unit, can explain approximately 45.2% of it. This means that 54.8% of the variation in re-sale price cannot be explained by floor area alone. Therefore, there must be other variables that have an influence also.

Significance testing in regression

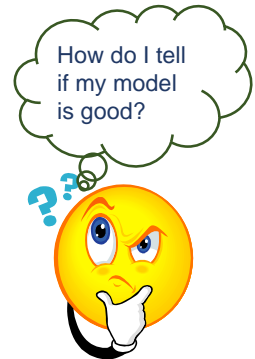
- Test hypothesis: That the variation explained by the models is not due to chance (F-test)

$$F = \frac{MS_M}{MS_R}$$

where


$$MS_M = \frac{\sum(\hat{y}_i - \bar{y})^2}{1}$$

$$MS_R = \frac{\sum(y_i - \hat{y}_i)^2}{n - 2}$$



The test statistic's F-value is the result of an one-way ANOVA that test the null hypothesis that all regression coefficients together are equal to zero.

If a model is good, then we expect the improvement in prediction due to the model to be large (so MS_M will be large) and the difference between the model and the observed data to be small (so MS_R will be small). In short, a good model should have a large F-ratio (greater than 1 at least) because the top of the equation will be bigger than the bottom. The exact magnitude of this F-ratio can be assessed using critical values for the corresponding degrees of freedom (refer to statistical distribution table).



School of
Computing and
Information Systems

Analysis of variance

ANOVA					
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	1.050377e+15	1	1.050377e+15	1182.042	0.0000
Residual	1.27427e+15	1434	888612164155.903		
Total	2.324647e+15	1435			

The Analysis of Variance report provides the calculations for comparing the fitted model to a simple mean model. The hypotheses for the F-test are:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \text{not all equal to 0}$$

Where k is the number of independent variables.

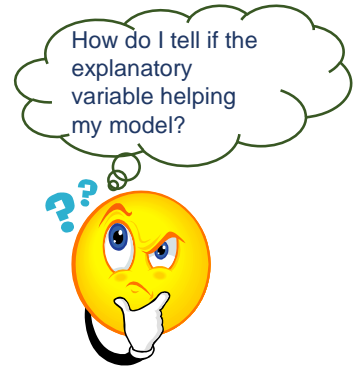
It reveals that the F-ratio is 1182.042 which is significant at $p < 0.0001$. This result tells us that there are less than 0.01% chance that an F-ratio this large will happen if the null hypothesis is true. Therefore, we can conclude that our regression model result is significantly better explanatory model of re-sale prices than if we used the mean value of re-sale prices. In short, the regression model overall estimates re-sale prices significantly well.

Assessing individual parameters

- Null hypothesis: $b = 0$
- Test statistics, t


$$\frac{b_{\text{observed}} - b_{\text{expected}}}{SE_b}$$

- The degrees of freedom are $N - p - 1$, where N is the total sample size and p is the number of explanatory variables.



The t-statistics tests the null hypothesis that the value of b is 0; therefore, if it is significant, we gain confidence in the hypothesis that the b -value is significantly different from 0 and that the explanatory variable contributes significantly to our ability to estimate values of the outcome.

A good parameter should have a large t -value (greater than 1 at least) because the top of the equation will be bigger than the bottom. The exact magnitude of this t -statistics can be assessed using critical values for the corresponding degrees of freedom (refer to statistical distribution table).



School of
Computing and
Information Systems

Describing the parameter estimates

Parameter Estimates							
model	Beta	Std. Error	Std. Beta	t	Sig.	lower	upper
(Intercept)	-258121.062	63517.224		-4.064	0.000	-382717.698	-133524.426
AREA_SQM	14719.028	428.117	0.672	34.381	0.000	13879.225	15558.832

This report contains the model parameters (the beta values) and the significance of these values. Based on the general formula of a simple regression model, β_0 is the Y intercept and this value is the value in the Estimate column for the intercept. With reference to the Parameter Estimates report, we can say that β_0 is -258121.062 and this can be interpreted as, when there is no information on the floor area of the re-sale unit (when $X = 0$), the model estimates that the re-sale price will be S\$258,121.062.

We can also read off the value of β_1 from the row labeled AREA_SQM and this value represents the gradient of the regression line. It is 14719.028. The positive sign indicates that the relationship is a direct one.

Although the value of β_1 is the slope of the regression line, it is more useful to think of this value as representing the change in the outcome associated with a unit change in the predictor. Therefore, if our predictor variable is increased by one unit (if the age of the car is increased by 1), then our model predicts that the price of the re-sale condominium unit will be increased by 14,719.028 unit. Our unit of measurement are Singapore dollar and square-meter. So, we can say that for with an increase in floor area by 1 square-meter, the re-sale price of the condominium will increase by S\$14,719.028.

The hypotheses for t-test are:

$$H_0: \beta_k = 0$$

$$H_1: \beta_k \neq 0$$

k = 1, 2, 3, K

The parameter Estimates reports also reveals that the t-ratios for Intercepts and AREA_SQM are -258,121.062 and 14,719.028 respectively. Both t-ratios are significant at $p < 0.0001$. Therefore, the β_s are different from 0 and we can conclude that the floor area of a unit makes a significant contribution ($p, 0.0001$) to predicting re-sale prices.

Assumptions of linear regression models

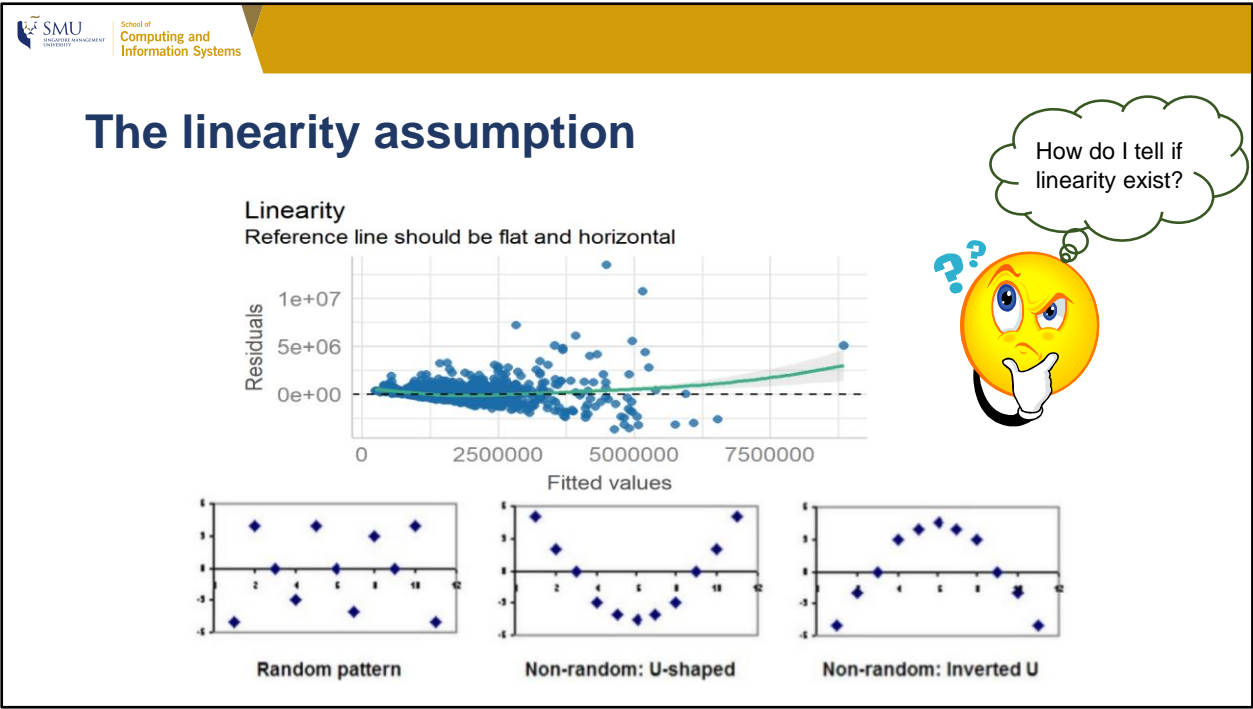


- The relationship between the dependent variable and independent variables is (approximately) linear.
- The expected mean error of the regression model is zero.
- The residuals do not vary with x .
 - constant variance, no heteroskedasticity
- The residuals are uncorrelated with each other.
 - serial correlation, as with time series
- (*Optional*) The errors (residuals) are normally distributed and have a 0 population mean.

Similar to the statistical techniques that you have learned previously, it is important for us to take note at some of the basic assumptions of linear regression. There are at least five major assumption of linear regressions model we should pay attention to. In this lesson, we will not go into detail to discuss the mathematical foundation of these assumptions but in the later discussion, we will share with you methods to verify if the multiple regression derived conforms to these basic assumptions.

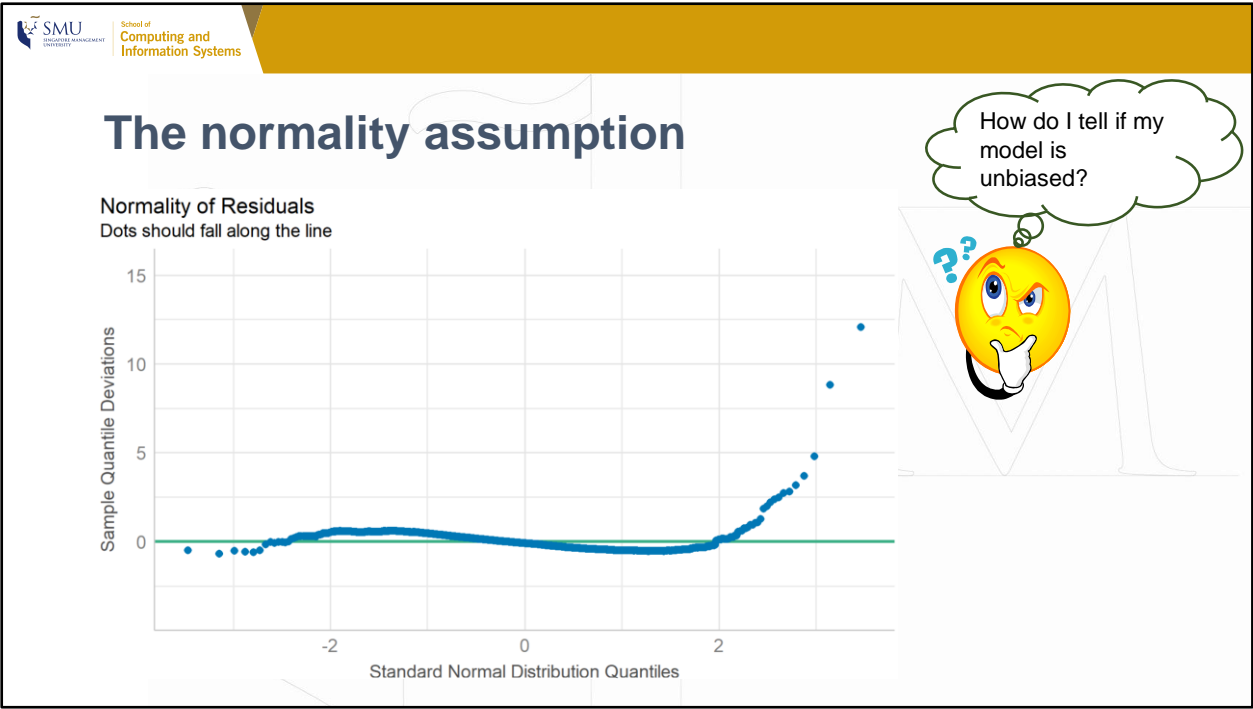
This list of assumptions probably seems pretty daunting assumptions are important. When the assumptions of regression are met, the model that we get for a sample can be accurately applied to the population of interest (the coefficients and parameters of the regression equation are said to be unbiased). Some people assume that this means that when the assumptions are met the regression model from a sample is always identical to the model that would have been obtained had we been able to test the entire population. Unfortunately, this belief isn't true.

What an unbiased model does tell us is that **on average the regression model from the sample is the same as the population model**. However, you should be clear that even when the assumptions are met, it is possible that a model obtained from a sample may not be the same as the population model – but the likelihood of them being the same is increased.



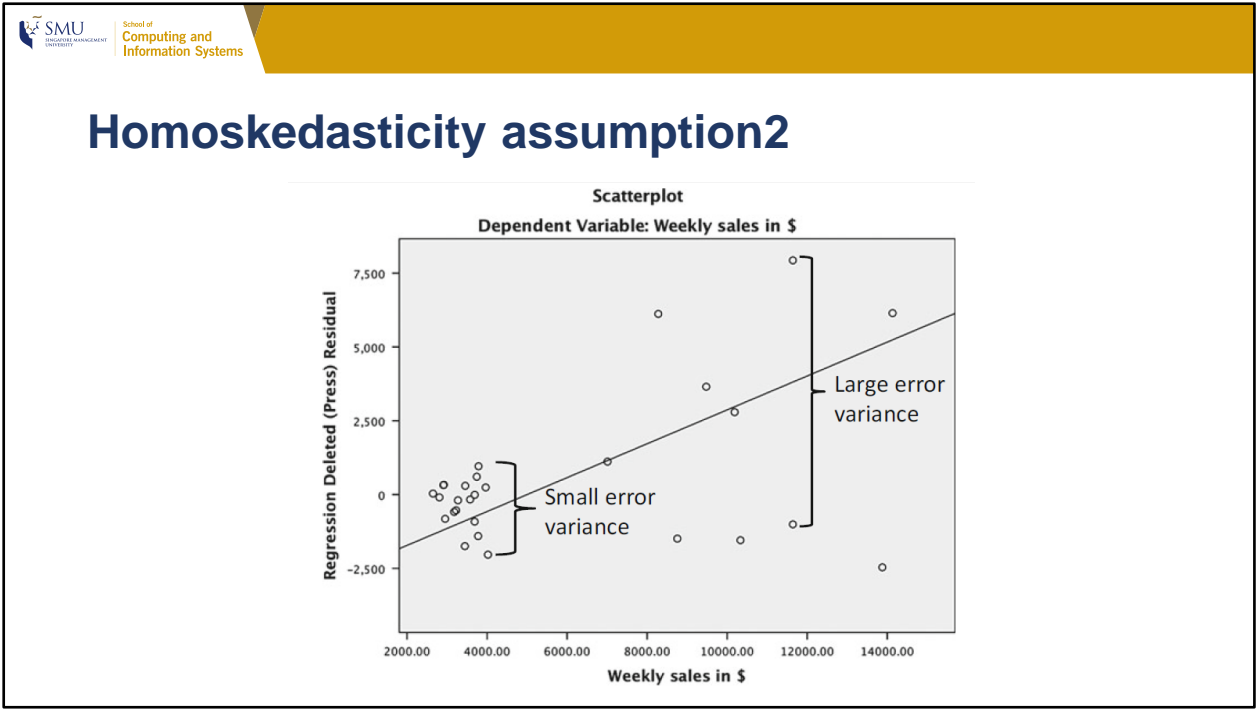
One of the basic assumption of a linear regression model is that the relationship between the dependent variable and independent variables is (approximately) linear. If you fit a linear model to data which are nonlinearly related, your predictions are likely to be seriously in error, especially when you extrapolate beyond the range of the sample data.

To verify the linearity assumption, we can examine the plot of residuals versus predicted values, which are a part of standard regression output. The points should be symmetrically distributed around a diagonal line figure above (e.g. random pattern). Look carefully for evidence of a "bowed" pattern (e.g. non-random: U-shaped and inverted U), indicating that the model makes systematic errors whenever it is making unusually large or small predictions.



One of the basic assumption of a linear regression model is that **normality assumption**. This mean that the error (residuals) are normally distributed and have a 0 population mean. Some people confuse this assumption with the idea that independent variables have to be normally distributed. In fact, independent variables do not need to be normally distributed.

The Distribution analysis report above shows that the standardised residual of the model is indeed normally distribution with mean very close to 0. We can safely conclude that the model conforms to the normality assumption.



The homoskedasticity assumption is that the errors’ variance is constant, a situation we call **homoskedasticity**. Imagine that we want to explain the weekly sales of various supermarkets in \$. Clearly, large stores have a much larger spread in sales than small supermarkets. For example, if you have average weekly sales of \$50,000, you might see a sudden jump to \$60,000 or a fall to \$40,000. However, a very large supermarket could see sales move from an average of \$5,000,000–\$7,000,000. This issue causes weekly sales’ error variance to be much larger for large supermarkets than for small supermarkets. We call this non-constant variance **heteroskedasticity**. We visualize the increasing error variance of supermarket sales in figure above, in which we can see that the errors increase as weekly sales increase.

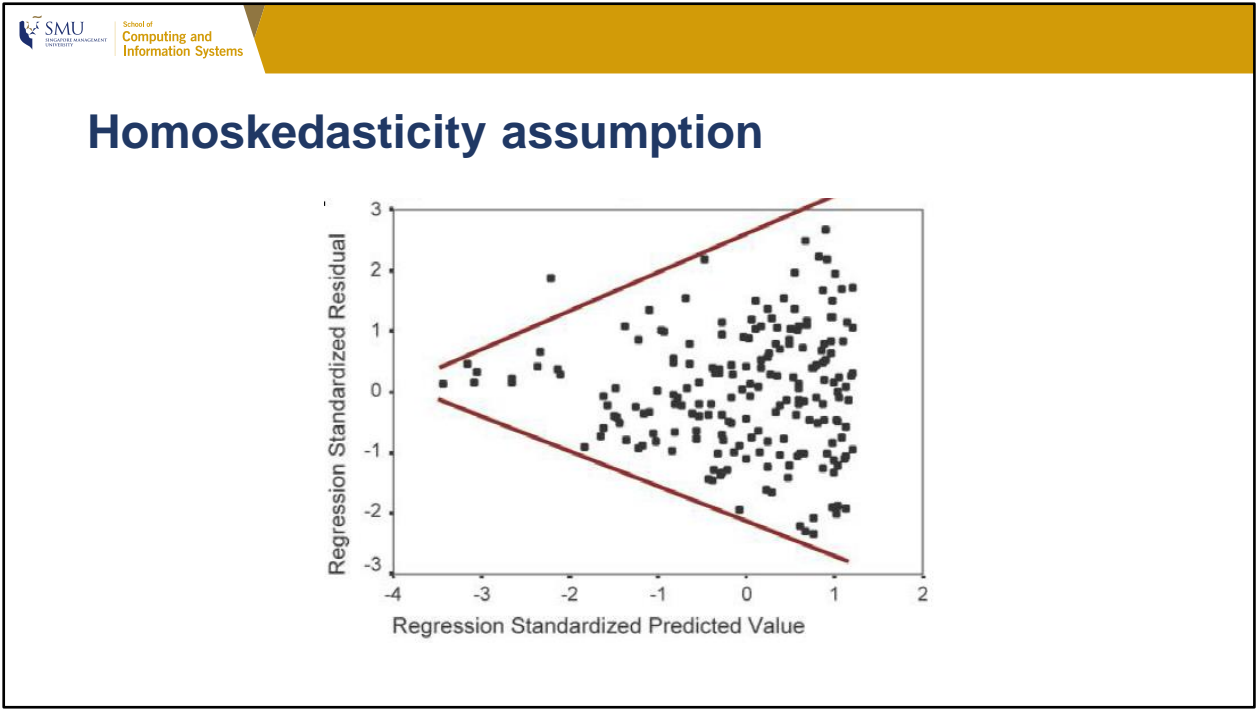
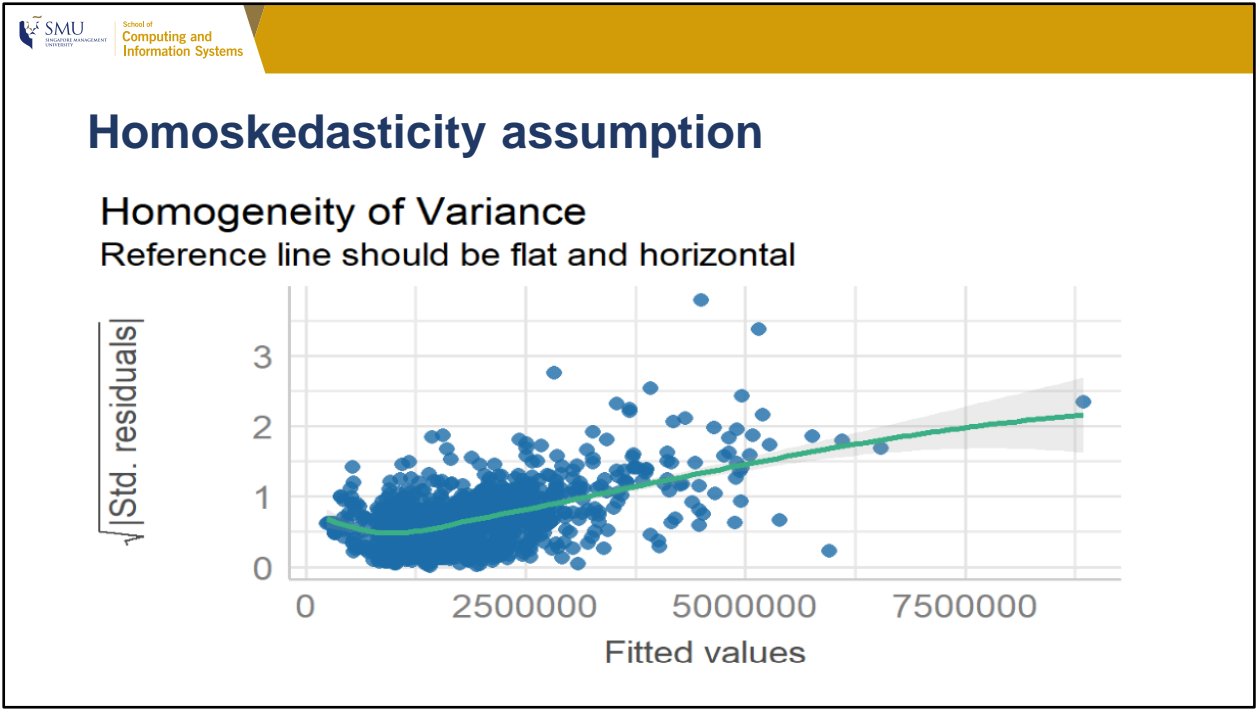



Figure above shows a plot of some data set that violates the assumption of homoscedasticity. Note that the points form the shape of a funnel so they become more spread out across the graph. This funnel shape is typical of heteroscedasticity and indicates increasing variance across the residuals.



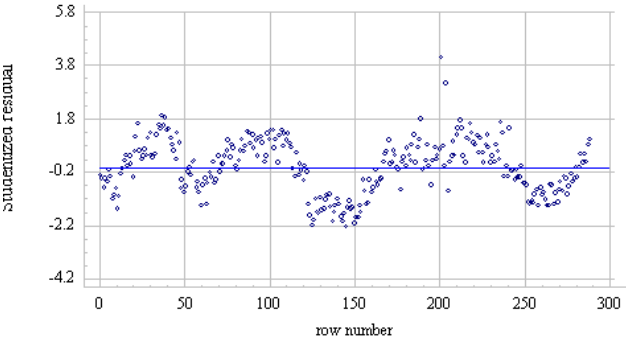


School of
Computing and
Information Systems


Checking for serial correlation

- At Regression report,
 - Row diagnostic s-> Plot residual by row

Residual Plot

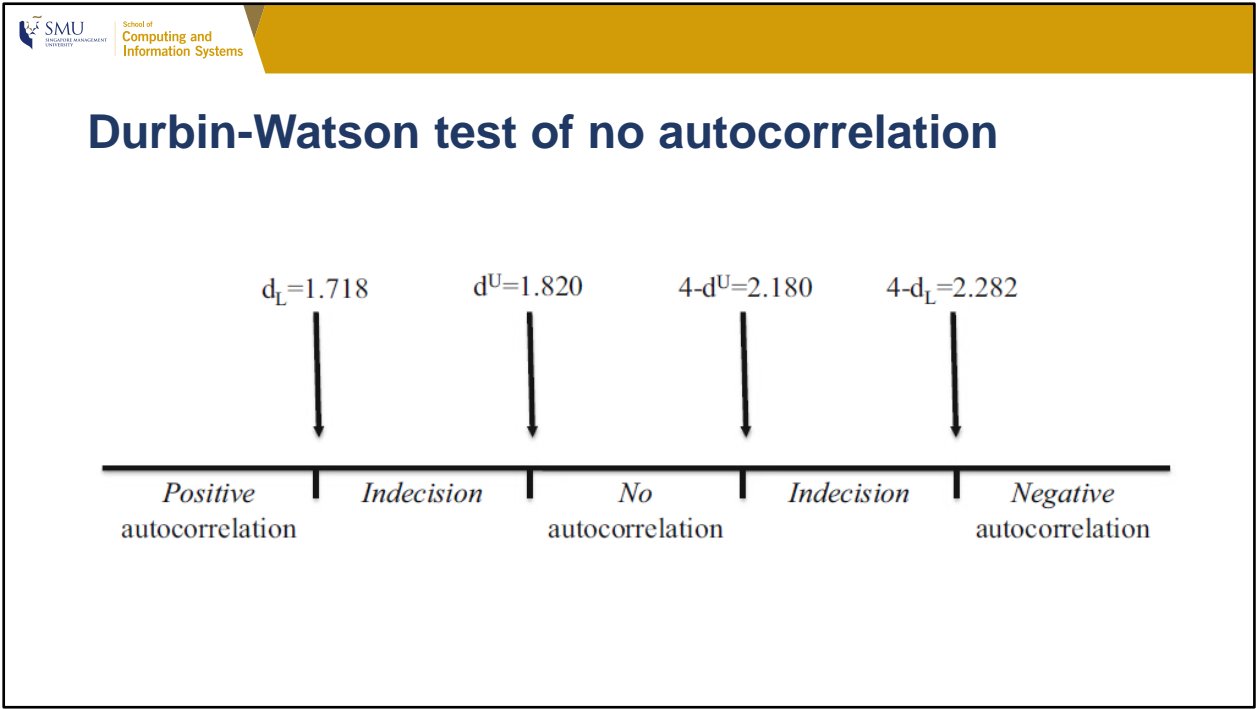


How do I tell if the dataset is temporally uncorrelated?




In regression analysis, we tend to assume the data given are time independent (or temporal independent). Plotting residuals versus row number (i.e., versus time) is always a good idea when you want to check if the residuals are correlated (also known as serial correlation).

The scatter plot above shows residuals clearly have a very strong pattern of positive autocorrelation--notice the long runs of errors with the same sign.



The **No Autocorrelation assumption** assumes that the regression model errors are independent; that is, the error terms are uncorrelated for any two observations. Imagine that you want to explain the sales of a particular supermarket using that supermarket’s previous week sales. It is very likely that if sales increased last week, they will also increase this week. This may be due to, for example, a growing economy, or other reasons that underlie supermarket sales growth. This issue is called autocorrelation and means that regression errors are correlated positively, or negatively, over time.

Fortunately, we can identify this issue using **the Durbin–Watson test**. The Durbin–Watson test assesses whether there is autocorrelation by testing a null hypothesis of no autocorrelation, which is tested against a lower and upper bound for negative autocorrelation and against a lower and upper bound for positive autocorrelation. Thus, there are four critical values. If we reject the null hypothesis of no autocorrelation, we find support for an alternative hypothesis that there is some degree of autocorrelation.



Condominium resale case: A revisit

- One dependent variable
- Multiple independent variables
- What are the combined effects of all the eight predictor variables on the used-car prices?

In real world explanatory modelling, it is very unusual that only one explanatory (or independent) variable is used. For example, to discover the re-sale price of a condominium unit, the model will be over simplified if we only consider the floor area of the unit as the sole explanatory variable. As all of us know, the price of a re-sale condominium might be also affected by its lease, and neighboring locational factors.

In this section, I will share with you the concepts and methods of Multiple Linear Regression by using the Hedonic Pricing Model. Instead of using the floor area of the condominium unit as the sole explanatory model, this time, nineteen explanatory variables will be used. These variables are *AREA_SQM*, *AGE*, *PROX_CBD*, *PROX_CHILDCARE*, *PROX_ELDERLYCARE*, *PROX_URB_GROWTH_AREA*, *PROX_HAWKER_MARKET*, *PROX_KINDERGARTEN*, *PROX_MRT*, *PROX_PARK*, *PROX_PRIMARY_SCH*, *PROX_TOP_PRIMARY_SCH*, *PROX_SHOPPING_MALL*, *PROX_SUPERMARKET*, *PROX_BUS_STOP*, *NO_Of_UNITS*, *FAMILY_FRIENDLY*, *FREEHOLD* and *LEASEHOLD_99YR*.



Multiple Linear Regression

- Regression establishes relationship among a dependent variable and a set of independent variable(s)
- A typical linear regression model looks like:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni}) + \varepsilon_i$$

- With Y_i the dependent variable, X_{ji} (j from 1 to n) the set of independent variables, $\beta_0 \dots \beta_n$ are the regression coefficients and ε_i the residual.

In order to address the task given in previous slide, we need to use multiple linear regression methods.

Similar to simple regression model, the least square method will be employed to determine the best fit linear regression line, in this case a multi-dimensional planer. The method used is call **Ordinary Least Square** method (OLS) in R.

When one uses MLR for explanatory purposes, that person is exploring relationships between multiple variables in a sample to shed light on a phenomenon, with a goal of generalising this new understanding to a population.

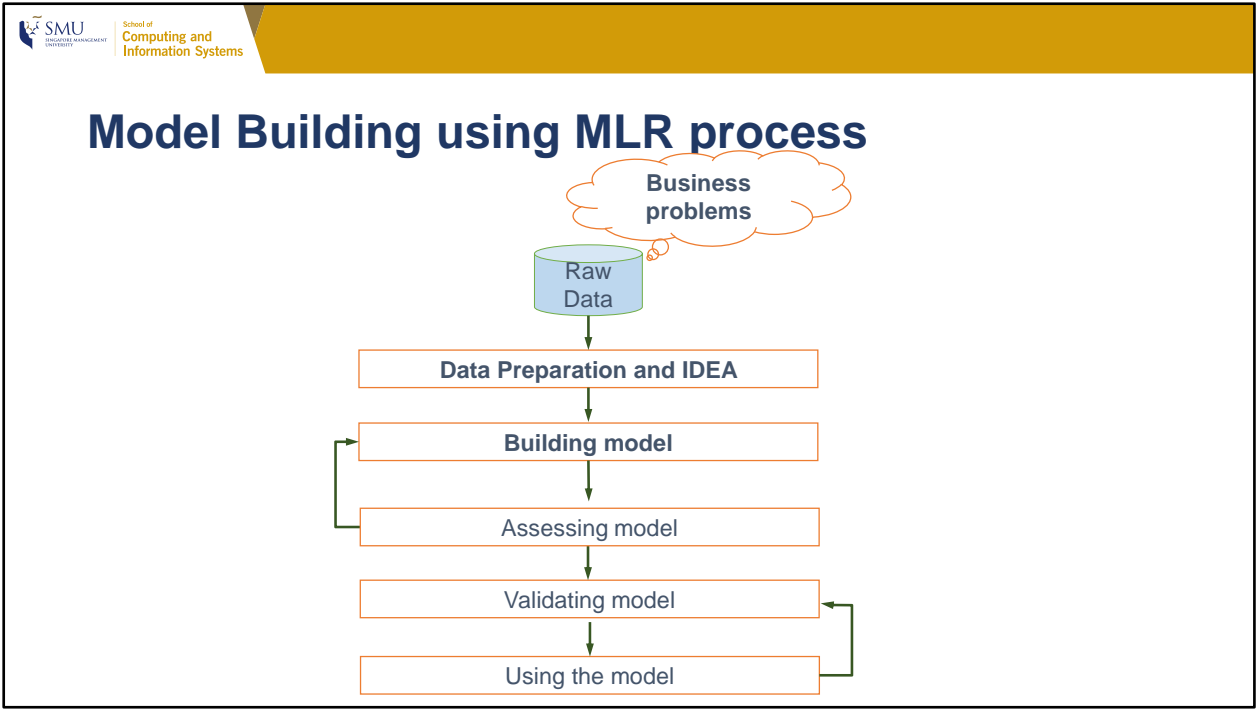


Figure above shows the model building process using multiple linear regression analysis. It is important to note that the process is an iterative process.

Data Requirement Considerations

- Scale type of the dependent variable
- Variables need to vary
- Sample size
- Collinearity

Several data requirements have to be considered before we undertake a regression analysis. These include: sample size, variables need to vary, scale type of the dependent variable, and collinearity.

The first data requirement is that the dependent variable needs to be interval or ratio scaled. If the data are not interval or ratio scaled, alternative types of regression need to be used. You should use binary logistic regression if the dependent variable is binary and only takes on two values (e.g., zero and one). If the dependent variable consists of a nominal variable with more than two levels, you should use multinomial logistic regression. This should, for example, be used if you want to explain why people prefer product A over B or C.

A regression model cannot be estimated if the variables have no variation.

Specifically, if there is no variation in the dependent variable (i.e., it is constant), we also do not need regression, as we already know what the dependent variable's value is. Likewise, if an independent variable has no variation, it cannot explain any variation in the dependent variable.

Size of the analytical dataset



Multiple Linear Regression is a large-data procedure. Unreliable results might be obtained if the dataset does not include at least 100 observations, preferably 200. The greater the number of explanatory variables included in the multiple regression equation, the greater the number of observations that will be necessary to obtain reliable results. Most experts recommend at least 15 to 30 observations per explanatory variable. See Cohen (1992).


The curse of multicollinearity

- Multicollinearity exists whenever two or more of the explanatory variables in a regression model are moderately or highly correlated.
- There are two types of multicollinearity:
 - **Structural multicollinearity** is a mathematical artifact caused by creating new explanatory variables from other explanatory variables — such as, creating the predictor x_2 from the predictor x .
 - **Data-based multicollinearity**, on the other hand, is a result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

Low levels of collinearity pose little threat to the model generated, but as collinearity increases there are three problems that arise:

- Untrustworthy bs: As collinearity increases so do the standard errors of the b coefficients.
- It limits the size of R .
- Multicollinearity between explanatory variables makes it difficult to assess the individual importance of a predictor. If the explanatory variables are highly correlated and each accounts for similar variance in the outcomes, then how can we know which of the two variables is important?

One way to detect multicollinearity is to scan a correlation matrix of all the explanatory variables and see if any correlate very highly (> 0.8). This is a good 'ball park' method but misses more subtle forms of multicollinearity. For example, it is possible that the pairwise correlations are small and yet a linear dependence exists among three or even more explanatory variables.




School of
Computing and
Information Systems

Variance Inflation Factors (VIF)

- Right-click in the report and select Columns -> VIF.

How do I tell if multicollinearity exist?



Parameter Estimates					
Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	-2373560	307223.1	-7.73	<.0001*	
Age_08_04	-23.11072	12.75058	-1.81	0.0703	32.075725
Mfg_Year	1178.6814	153.3588	7.69	<.0001*	31.87202
KM	-0.01475	0.001533	-9.62	<.0001*	1.9488232
HP	41.431786	3.523356	11.76	<.0001*	1.587666
CC	-4.039544	0.383233	-10.54	<.0001*	3.031268
Quarterly_Tax	6.2052892	1.751382	3.54	0.0004*	3.2367679
Weight	29.245774	1.523803	19.19	<.0001*	3.6346531
Boardcomputer	-300.2181	137.2895	-2.19	0.0290*	2.2082475

The variance inflation factors (VIF) are useful in determining which variables may be correlated or collinear. It can be shown that the variance of the estimate of the corresponding regression coefficient is larger by that factor than it would be if there were no collinearity. In other words, the VIF statistics show how collinearity has increased the instability of the coefficient estimates.


Mathematically, the VIF statistic is calculated using the following formula:

$$VIF = \frac{1}{1 - RSquare_j}$$

where $RSquare_j$ is the multiple $RSquare$ for the regression of X_j on the other covariates (a regression that does not involve the response variable Y).

There are no formal criteria for determining the magnitude of variance inflation factors that cause poorly estimated coefficients. Some authorities state that values exceeding **10** may be cause for concern, but this value is arbitrary. Actually, for models with low coefficients of determination (R^2) for the regression, estimates of coefficients that exhibit relatively small variance inflation factors may still be unstable, and vice versa.

Figure above shows that Age and the Manufacturing Year of the resale car have VIFs greater than 10.



School of
Computing and
Information Systems


Correlation of Estimates

Correlation of Estimates									
Corr	Intercept	Age_08_04	Mfg_Year	KM	HP	CC	Quarterly_Tax	Weight	Boardcomputer
Intercept	1.0000	-0.9477	-1.0000	-0.0135	0.0481	-0.0806	0.0181	0.0837	0.0925
Age_08_04	-0.9477	1.0000	0.9469	-0.1081	-0.0742	0.0632	-0.0272	0.0376	0.1169
Mfg_Year	-1.0000	0.9469	1.0000	0.0131	-0.0478	0.0815	-0.0161	-0.0880	-0.0936
KM	-0.0135	-0.1081	0.0131	1.0000	0.2620	-0.3144	-0.1146	0.1278	-0.0422
HP	0.0481	-0.0742	-0.0478	0.2620	1.0000	-0.2912	0.4951	-0.2022	-0.0943
CC	-0.0806	0.0632	0.0815	-0.3144	-0.2912	1.0000	-0.3729	-0.4340	0.0532
Quarterly_Tax	0.0181	-0.0272	-0.0161	-0.1146	0.4951	-0.3729	1.0000	-0.4261	-0.1075
Weight	0.0837	0.0376	-0.0880	0.1278	-0.2022	-0.4340	-0.4261	1.0000	0.1231
Boardcomputer	0.0925	0.1169	-0.0936	-0.0422	-0.0943	0.0532	-0.1075	0.1231	1.0000

Now, let us try to understand why Age and Manufacturing Year of the resale car have high VIF values.

Notice that the correlation coefficient between Age_08_04 and Mfg_Year is 0.9469 and is marked in dark blue. This indicates that Age_08_04 is strongly and positively correlated to Mfg_Year (Manufacturing Year).

Since these two explanatory variables are highly correlated (i.e. 0.9469), it is very safe for you to drop one of them in the subsequent analysis.



School of
Computing and
Information Systems

Assessing and Interpreting the Whole Model


- Describe the regression model
- Describe the overall fit of the multiple regression model

Model Summary


R	0.807	RMSE	750537.537
R-Squared	0.652	MSE	571262902261.223
Adj. R-Squared	0.647	Coef. Var	43.160
Pred R-Squared	0.637	AIC	42971.173
MAE	412117.987	SBC	43081.835

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

How do I tell if my model is good?



Let’s look first at the R^2 and Adjusted R^2 of the model under the Summary of Fit report. The R^2 and Adjusted R^2 of the model are 0.652 and 0.647 respectively. The value of adjusted R^2 tells us that the nineteen explanatory variables can account for 64.7% of the variation in the price of the re-sale condominium units. In other words, if we are trying to explain why the prices of some condominium units are higher than others, we can look at the variation in *AREA_SQM*, *AGE*, *PROX_CBD*, *PROX_CHILDCARE*, *PROX_ELDERLYCARE*, *PROX_URA_GROWTH_AREA*, *PROX_HAWKER_MARKET*, *PROX_KINDERGARTEN*, *PROX_MRT*, *PROX_PARK*, *PROX_PRIMARY_SCH*, *PROX_TOP_PRIMARY_SCH*, *PROX_SHOPPING_MALL*, *PROX_SUPERMARKET*, *PROX_BUS_STOP*, *NO_Of_UNITS*, *FAMILY_FRIENDLY*, *FREEHOLD* and *LEASEHOLD_99YR* of the re-sale condominium units.



School of
Computing and
Information Systems

Adjusted R-squared

$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

- R^2 = sample R-square
- p = Number of predictors
- N = Total sample size.

Vars	R-Sq	R-Sq(adj)
1	72.1	71.0
2	85.9	84.8
3	87.4	85.9
4	89.1	82.3
5	89.9	80.7


Multiple regression can be a beguiling, temptation-filled analysis. It’s so easy to add more variables as you think of them, or just because the data are handy. Some of the explanatory variables will be significant. Perhaps there is a relationship, or is it just by chance? You can add higher-order polynomials to bend and twist that fitted line as you like, but are you fitting real patterns or just connecting the dots? All the while, the R-squared (R²) value increases, teasing you, and egging you on to add more variables!

These problem can be rectified by using adjusted R-squared. The adjusted R-squared compares the explanatory power of regression models that contain different numbers of explanatory variables.

Suppose you compare a five-predictor model with a higher R-squared to a one-predictor model. Does the five predictor model have a higher R-squared because it’s better? Or is the R-squared higher because it has more explanatory variables? Simply compare the adjusted R-squared values to find out!

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of explanatory variables 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. The adjusted R-squared can be negative, but it’s usually not. It is always lower than the R-squared.

In the simplified Best Subsets Regression output above, you can see where the adjusted R-squared peaks, and then declines. Meanwhile, the R-squared continues to increase.




School of
Computing and
Information Systems

Assessing and Interpreting the Whole Model


- Describe the overall fit of the multiple regression model

ANOVA					
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	1.515738e+15	19	7.977571e+13	139.648	0.0000
Residual	8.089083e+14	1416	571262902261.223		
Total	2.324647e+15	1435			

How do I tell if my model is good?




The Analysis of Variance report provides the calculations for comparing the fitted model to a simple mean model. It reveals that the F-ratio is 139.648 and the p-value is less than $\alpha = 0.05$. This result tells us that there are less than 5% chance that an F-ratio of this large will happen if the null hypothesis is true. Therefore, we can reject the null hypothesis and conclude that our regression model result is significantly better estimation of re-sale prices of condominium than if we used the mean price of there-sale condominium unit. In short, the multiple linear regression model overall explains prices of the re-sale condominiums significantly well.



School of

Computing and

Information Systems



Assessing and Interpreting model parameters

Parameter Estimates							
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	543071.420	136210.918		3.987	0.000	275874.535	810268.305
AREA_SQM	12688.669	370.119	0.579	34.283	0.000	11962.627	13414.710
AGE	-24566.001	2766.041	-0.166	-8.881	0.000	-29991.980	-19140.022
PROX_CBD	-78121.985	6791.377	-0.267	-11.503	0.000	-91444.227	-64799.744
PROX_CHILDCARE	-333219.036	111020.303	-0.087	-3.001	0.003	-551000.984	-115437.089
PROX_ELDERLYCARE	170949.961	42110.748	0.083	4.060	0.000	88343.803	253556.120
PROX_URA_GROWTH_AREA	38507.622	12523.661	0.059	3.075	0.002	13940.700	63074.545
PROX_HAWKER_MARKET	23801.197	29299.923	0.019	0.812	0.417	-33674.725	81277.120
PROX_KINDERGARTEN	144097.972	82738.669	0.030	1.742	0.082	-18205.570	306401.514
PROX_MRT	-322775.874	58528.079	-0.123	-5.515	0.000	-437586.937	-207964.811
PROX_PARK	564487.876	66563.011	0.148	8.481	0.000	433915.162	695060.590
PROX_PRIMARY_SCH	186170.524	65515.193	0.072	2.842	0.005	57653.253	314687.795
PROX_TOP_PRIMARY_SCH	-477.073	20597.972	-0.001	-0.023	0.982	-40882.894	39928.747
PROX_SHOPPING_MALL	-207721.520	42855.500	-0.109	-4.847	0.000	-291788.613	-123654.427
PROX_SUPERMARKET	-48074.679	77145.257	-0.012	-0.623	0.533	-199405.956	103256.599
PROX_BUS_STOP	675755.044	138551.991	0.133	4.877	0.000	403965.817	947544.272
NO_OF_UNITS	-216.180	90.302	-0.046	-2.394	0.017	-393.320	-39.040
FAMILY_FRIENDLY1	142128.272	47055.082	0.056	3.020	0.003	49823.107	234433.438
FREEHOLD1	300646.543	77296.529	0.117	3.890	0.000	149018.525	452274.561
LEASEHOLD_99YR1	-77137.375	77570.869	-0.030	-0.994	0.320	-229303.551	75028.801

The Parameter Estimates report contains the model parameters (the beta values) and the significance of these values. With reference to the Parameter Estimates report, the Estimate column reveals that resale prices of the condominiums are directly related AREA_SQM, PROX_ELDERCARE, PROX_URA_GROWTH_AREA, etc. and inversely related to AGE, PROX_CBD, PROX_CHILDCARE, etc.

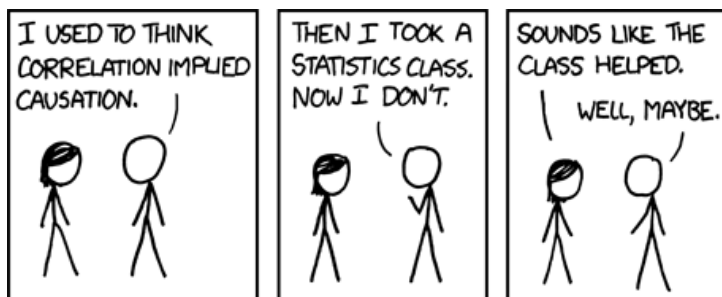
By referring to the Sig>|t| (also known as p-value) column, we can see that except PROX_HAWKER_MARKET, PROX_KINDERGARTEN, PROX_TOP_PRIMARY_SCH, PROX_SUPERMARKET and LEASEHOLD_99YR1, the p-values of the rest of the explanatory variables are smaller than the critical value (i.e. $\alpha = 0.05$). Hence, we can conclude that except PROX_HAWKER_MARKET, PROX_KINDERGARTEN, PROX_TOP_PRIMARY_SCH, PROX_SUPERMARKET and LEASEHOLD_99YR1, the rest of the explanatory variables contribute significantly in explaining the resale price of the condominium units.

The explanatory model above reveals that among the nineteen explanatory variables, AREA_SQM contributes the most in explaining the resale price of the used car.

It is also important to note that their relationship of not the same. For example, if the AREA_SQM increases by 1 unit, the resale price will increase by 12688.669. On the other hand, if the AGE of the re-sale condominium increases by 1 unit, it's resale price will decrease by24566.001.

Cautions about Interpreting Correlation

- Relationship between an explanatory variable (i.e. x) and a dependent variable does not mean a **cause-and-effect** relationship is present between x and y .

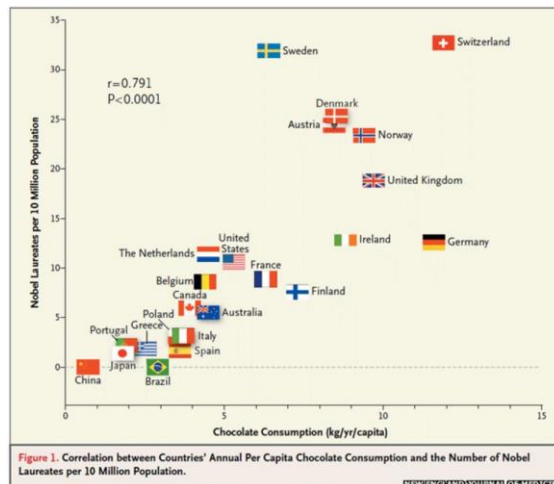


Multiple regression can determine whether a given set of variables is useful for explaining a response variable. Among other things, this means that multiple regression can be used to determine:

- whether the relationship between the response variable and explanatory variables (taken as a group) is statistically significant
- how much variance in the response is accounted for by the explanatory variables
- which explanatory variables are relatively important explanatory variables of the response.

Although the preceding section often refers to causal models, it is important to remember that the procedures discussed in this section do not provide strong evidence concerning cause-and-effect relationships between explanatory variables and the response. The most that you can say is that your findings are consistent with the causal model. It is incorrect to say that the results prove that the model is correct.

Stop the Fallacy of Data Storytelling



Source: [Franz H. Messerli \(2012\) Chocolate Consumption, Cognitive Function, and Nobel Laureates](#), *The New England Journal of Medicine*.

SMU

SCHOOL OF

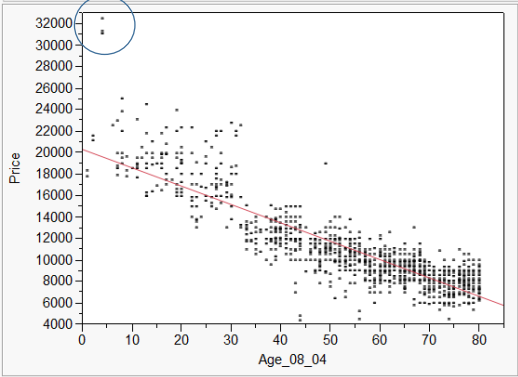
COMPUTING AND

INFORMATION SYSTEMS

Outliers and influential of data points

- Do you think the following data set contains any outliers and/or any influential data points?

Bivariate Fit of Price By Age_08_04



Linear Fit

Price = 20294.059 - 170.93359*Age_08_04

Summary of Fit

RSquare	0.768411
RSquare Adj	0.768249
Root Mean Square Error	1746.038
Mean of Response	10730.82
Observations (or Sum Wgts)	1436

Lack Of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	20294.059	146.0973	138.91	<.0001*
Age_08_04	-170.9336	2.478079	-68.98	<.0001*

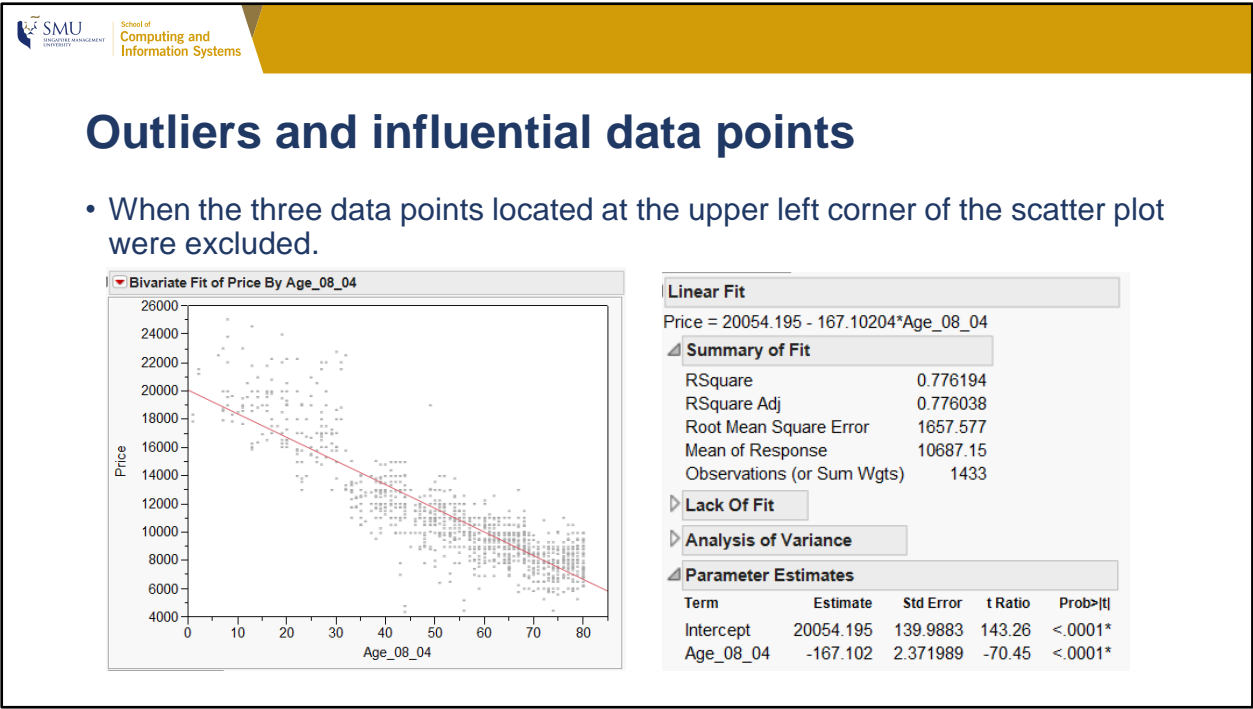
An **outlier** is a data point whereby response y does not follow the general trend of the rest of the data.

A data point is **influential** if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.

Note that, for our purposes, we consider a data point to be an outlier *only if* it is extreme with respect to the other y values, not the x values.

In this case, the circled blue data point are most certainly outliers and influential! They don't follow the general trend of the rest of the data

5-39



Strategies for dealing with problematic data points

- Decide whether or not deleting data points is warranted:
 - Do not delete data points just because they do not fit your preconceived regression model.
 - You must have a good, objective reason for deleting data points.
 - If you delete any data after you've collected it, justify and describe it in your reports.
 - If you are not sure what to do about a data point, analyze the data twice - once with and once without the data point and report the results of both analyses.

First, foremost, and finally - it's okay to use your common sense and knowledge about the situation.

With JMP, you can use the Row -> Exclude/Unexclude feature to exclude the problematic data points from the analysis instead of deleting them from the dataset. This approach is relatively more flexible and effective than deleting the problematic data points from the dataset because we can easily bring back the problematic data points by unexcluding them.

Categorical explanatory variables and Multiple Regression

- Technically, no categorical explanatory variables should be used in fitting a multiple regression.
- In order to use categorical explanatory variables, **dummy variables** have to be created.
- Dummy coding is a way of representing groups of observations using only zeros and ones.

Dummy coding process

- Count the number of groups you want to recode and subtract 1.
- Create as many new variables as the value you calculated in Step 1. These are your dummy variables.
- For the first dummy variable, assign the value 1 to the first group that you want to compare with and assign all other groups 0.
- Repeat this until you run out of dummy variables.


Stepwise Regression Techniques

- Goal: Find parsimonious model (the simplest model that performs sufficiently well)
 - More robust
 - Higher predictive accuracy
- Exhaustive Search
- Stepwise (Partial Search) Algorithms
 - Forward
 - Backward
 - Stepwise

Running a regression model with many variables including irrelevant ones will lead to a needlessly complex model. Stepwise regression is a way of selecting important variables to get a simple and easily interpretable model.

Stepwise regression is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves adding or removing potential explanatory variables in succession and testing for statistical significance after each iteration.

The availability of statistical software packages makes stepwise regression possible, even in models with hundreds of variables.



School of
Computing and
Information Systems

Forward Stepwise Regression

Forward stepwise selection example with 5 variables:

Start with a model with no variables

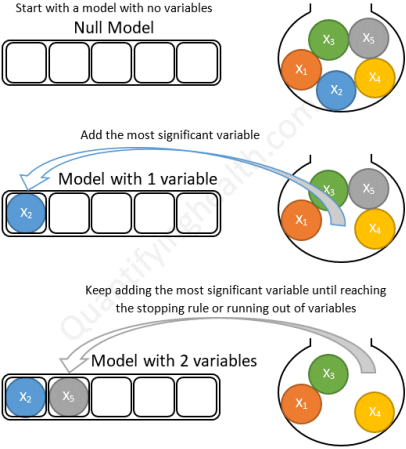
Null Model

Add the most significant variable

Model with 1 variable

Keep adding the most significant variable until reaching the stopping rule or running out of variables

Model with 2 variables



Forward Stepwise Regression procedures

- Begins with a model that contains no variables (called the Null Model)
- Then starts adding the most significant variables one after the other
- Until a pre-specified stopping rule is reached or until all the variables under consideration are included in the model

SMU

SCHOOL OF
ENGINEERING MANAGEMENT
UNIVERSITY

School of
Computing and
Information Systems

Stopping Rule and P-value Threshold

Stopping Rule:

P-value Threshold

P-value Threshold

Minimum AICc

Minimum BIC

Max Validation RSquare

Max K-Fold RSquare

Direction:

Rules:

P-value Threshold

Prob to Enter0.05

Prob to Leave0.05

In order to fully understand how forward selection works, we need to know:

1. How to determine the most significant variable at each step
2. How to choose a stopping rule

1. Determine the most significant variable to add at each step

The most significant variable can be chosen so that, when added to the model:

- It has the smallest p-value, or
- It provides the highest increase in R^2 , or
- It provides the highest drop in model RSS (Residuals Sum of Squares) compared to other predictors under consideration.

- ## 2. Choose a stopping rule

The stopping rule is satisfied when all remaining variables to consider have a p-value larger than some specified threshold, if added to the model. When we reach this state, forward selection will terminate and return a model that only contains variables with p-values $<$ threshold.

How to determine the threshold?

The threshold can be:

- A fixed value (for instance: 0.05 or 0.01 or 0.001)
- Determined by AIC (Akaike Information Criterion)
- Determined by BIC (Bayesian information criterion)

If we choose a fixed value, the threshold will be the same for all variables. However, if we let AIC or BIC automatically determine the threshold, it will be different for each variable. Fortunately, computers nowadays calculate these thresholds automatically so we do not have to bother with the details. However, I think it is interesting to have at least some understanding of what is going on under the hood.

How does AIC and BIC determine the threshold?



AIC chooses the threshold according to how many degrees of freedom the variable under consideration has. Take for example the case of a binary variable (by definition it has 1 degree of freedom): According to AIC, if this variable is to be included in the model, it needs to have a p-value < 0.157 . The more degrees of freedom a variable has, the lower the threshold will be.

BIC chooses the threshold according to the effective sample size n . For instance, for $n = 20$, a variable will need a p-value < 0.083 in order to enter the model. The larger n is, the lower the threshold will be.

BIC is a more restrictive criterion than AIC and so yields smaller models. Therefore it is only recommended when working with large sample sizes — where the sample size (or number of events in case of logistic regression) exceeds 100 per independent variable [Heinze et al.].

Note that both AIC (and BIC) can be applied to the pooled degrees of freedom of all unselected predictors. But applying it to individual variables (like we described above) is far more prevalent in practice.

SMU

SEMIKOTA UNIVERSITY

School of

Computing and

Information Systems

Backward Stepwise Regression

Backward stepwise selection example with 5 variables:

Start with a model that contains all the variables

Full Model

Remove the least significant variable


Model with 4 variables

Keep removing the least significant variable until reaching the stopping rule or running out of variables

Model with 3 variables

Backward stepwise selection (or backward elimination) is a variable selection method which:

- Begins with a model that contains all variables under consideration (called the Full Model)
- Then starts removing the least significant variables one after the other
- Until a pre-specified stopping rule is reached or until no variable is left in the model



School of
Computing and
Information Systems

Mixed Stepwise Regression

ToyotaCorolla_MLR-v1 - Fit Stepwise - JMP Pro

Stepwise Fit for Price

Stepwise Regression Control

Stopping Rule: P-value Threshold

Prob to Enter 0.25

Prob to Leave 0.95

Direction: Mixed

Go Stop Step

rows not used due to excluded rows or missing values.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	RSquare Validation	RMSE Validation	RSquareTest	RMSE Test
3.1509e+9	273	3397.2965	0.0000	0.0000	1646.0256	1	5236.264	5243.446	-0.004	3690.432	-0.001	3660.788

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	10894.8832	1	0	0.000	1
<input type="checkbox"/>	<input type="checkbox"/>	Age_08_04	0	1	2.599e+9	1280.897	7e-105
<input type="checkbox"/>	<input type="checkbox"/>	KM	0	1	1.062e+9	138.191	4.5e-26
<input type="checkbox"/>	<input type="checkbox"/>	HP	0	1	2.677e+8	25.255	9.12e-7
<input type="checkbox"/>	<input type="checkbox"/>	cc	0	1	1.207e+8	10.831	0.00113
<input type="checkbox"/>	<input type="checkbox"/>	Quarterly_Tax	0	1	73858741	6.529	0.01116
<input type="checkbox"/>	<input type="checkbox"/>	Sport_Model	0	1	41209129	3.605	0.05868

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	RSquare Validation
------	-----------	--------	------------	--------	---------	----	---	------	-----	--------------------

In JMP Pro, there is a third algorithm called Mixed Stepwise Regression. It is available only when the P-value Stopping Rule is selected. It alternates the forward and backward steps. It includes the most significant term that satisfies Prob to Enter an removes the least significant term satisfying Prob to Leave. It continues removing terms until the remaining terms are significant and then it changes to the forward direction.

Which stepwise selection methods should be used?



Where forward stepwise is better.

Unlike backward elimination, forward stepwise selection can be used when the number of variables under consideration is very large, even larger than the sample size! This is because forward selection starts with a null model (with no predictors) and proceeds to add variables one at a time, and so unlike backward selection, it DOES NOT have to consider the full model (which includes all the predictors). In fact, it will only consider models with number of variables less than:

- The sample size (for linear regression)
- The number of events (for logistic regression)

Where backward stepwise is better

Starting with the full model has the advantage of considering the effects of all variables simultaneously. This is especially important in case of collinearity (when variables in a model are correlated with each other) because backward stepwise may be forced to keep them all in the model unlike forward selection where none of them might be entered.

BOTTOM LINE:

Unless the number of candidate variables > sample size (or number of events), use a backward stepwise approach.

Should you use stepwise selection?

Stepwise variable selection has been a very popular technique for many years, but if this procedure had just been proposed as a statistical method, it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing.

Regression modeling strategies – Frank Harrell

Be warned! The selection of variables using a stepwise regression will be highly unstable, especially when we have a small sample size compared to the number of variables we want to study. This is because many variable combinations can fit the data in a similar way! You can test the instability of the stepwise selection by rerunning the stepwise regression on different subsets of your data. When there is instability, you will notice that you'll get a different selection of variables each time. This instability is reduced when we have a sample size (or number of events) > 50 per explanatory variable. In case you didn't notice, 50 is a really HUGE number. Imagine that for a stepwise regression with only 10 candidate variables you will need 500 events to reduce the instability of the stepwise selection algorithm!

Tips on how to build a rigorous regression model

- Uses large amounts of trustworthy data and a small number of explanatory variables that have well established causal relationships.
- Uses sound reasoning for including variables in the model.
- Brings together different lines of research as needed.
- Effectively presents the results using graphs, confidence intervals, and prediction intervals in a clear manner that ensures proper interpretation by others.

This slide shows four useful tips on how to build a regourous Multiple Linear Regression model.

Reference

- CHENG Lijuan & YAN Qiang (2012) "Affect Factor Analysis of the Chinese Government Administrative Costs Empirical Analysis Based on the Data from 1978 to 2010" *Cross-Cultural Communication*, Vol. 8, No. 5, 2012, pp. 103-110.
- Leif Atle Beisland (2011) "The Effects of Earnings Variables on Stock Returns among Public Companies in Norway: A Multiple Regression Analysis". *International Journal of Management*, Vol. 28 No. 3 Part 1 Sept 2011.
- Chalikia, M.H. & Hinsz, V.B. (2013) "Sex-Based Salary Disparity and the Uses of Multiple Regression for Definition and Remediation". *Current Psychology*, Vol. 32, pp. 374–387.
- Jasrai, L. (2014) "Predicting Customer Satisfaction Towards Mobile Value-Added Services: An Application of Multiple Regression". *The IUP Journal of Marketing Management*, Vol. XIII, No. 1, pp. 29-44.
- Keskin, B (2008) "Hedonic Analysis of Price in the Istanbul Housing Market" *International Journal of Strategic Property Management*, Vol. 12, pp. 125–138.
- Muganiwa, K. & Lambrechts, H. (2017) "Determining soybean futures prices on Safex using multiple linear regression". *Management Dynamics*, Vol 26, No.3, pp. 2-15.

We are coming to the end of this lesson. For further reading, students are encouraged to consult the references provided.

Online Resource

- Regression Analysis Tutorial and Examples (<http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-tutorial-and-examples>)
- Violations of the Assumptions for Linear Regression (<http://blog.minitab.com/blog/statistics-and-quality-data-analysis/violations-of-the-assumptions-for-linear-regression-the-trial-of-lionel-loosefit-day-1>)
- “When Can You Safely Ignore Multicollinearity?” (<http://www.statisticalhorizons.com/multicollinearity>)
- “Bad weather or bad management? What’s really driving New York City’s notorious pothole problem.” (<https://www.informs.org/ORMS-Today/Public-Articles/June-Volume-41-Number-3/POTHOLE-ANALYTICS>)
- “Large and Small Regression Coefficients” (<http://davegiles.blogspot.sg/2013/08/large-and-small-regression-coefficients.html>).