DS Part IV – Machine Learning
(Model Building ,Evaluation & Communication)

Sasken Training, 9840014739, Adyar , Chennai – 600 020

Machine Learning Model Building

Supervised Machine Learning – Linear Regression

Agenda

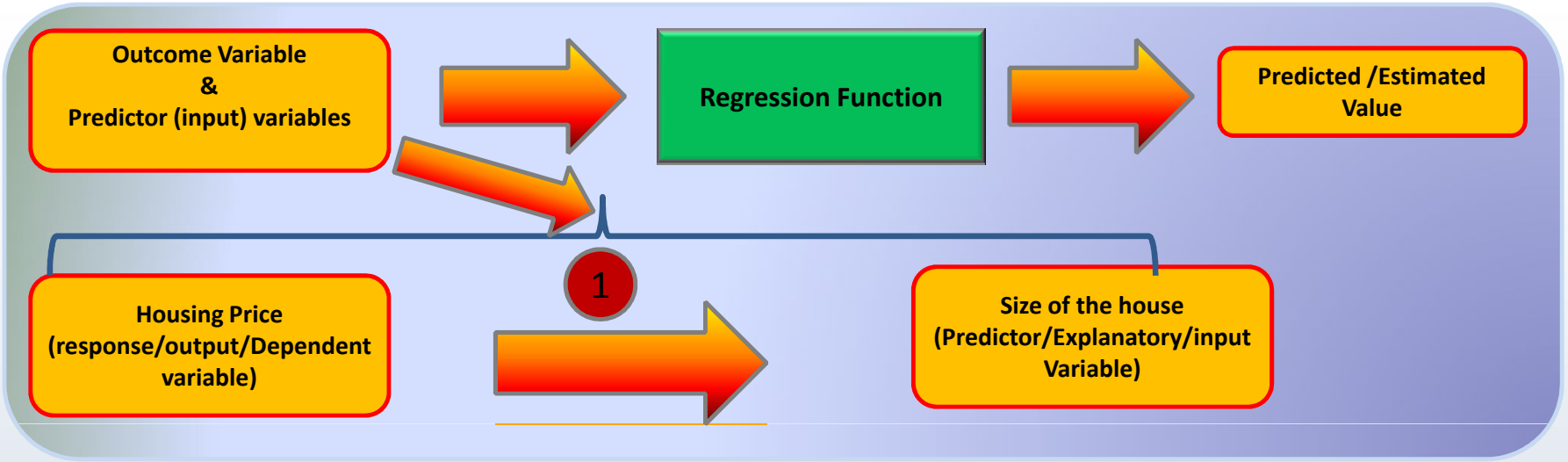
- Regression Introduction
- Simple Linear Regression
- Multiple Linear Regression
- Regression Assumptions
- Implementation in R

What is Regression Analysis ?

- Used to estimate a continuous value as a linear function of explanatory variable or variables
  - House price as a function of median home price in the squre feet, number of bed rooms, etc.,
- Its mainly used for prediction where its use has substantial overlap with the field of machine learning.
- It's also used to understand which among the explanatory variables are related to outcome variable.

To understand and quantify cause and effect relationship

What is Regression Analysis ?



1. Are there any relationship exists between Housing Prices and the size of the houses ? Is there any change in price correlate linearly with a change in size

2. If yes, we can predict a price of a house given their size

Regression Analysis

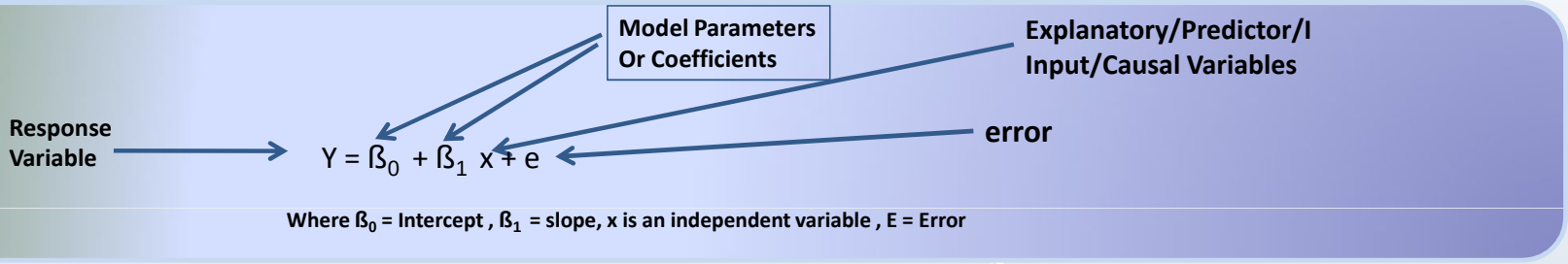
Lets say you work for a realtor and are looking to understand factors that may influence the price of a house

What factors would you think of having an influence on housing price ?

- Size of an House
- Number of bed rooms
- Car Park
- Area

Simple Linear Regression

- We can define the regression equation as the straight line equation between the two variables
- We have to account for the unobserved variables in the equation, which is typically captured as “error” or disturbance.
- The Simple Linear Regression Model is usually denoted by



Where $ß_0$ = Intercept , $ß_1$ = slope, x is an independent variable , E = Error

- Slope - how much the line rises for each increase in x
- Betas needs to be estimated so that we can understand the relationship between Y and the X
- Using OLS (Ordinary Least Squares) technique the model estimates the unknown parameters (Intercept & Slope) , with the goal of minimizing the sum of squared distances

Terminology

Dependent Variable - The variable y (refer previous slide) is dependent variable and also it is called as predicted variable or outcome variable

Independent Variable - The variable x (refer previous slide) is independent variable and also it is called as predictor variable or explanatory variable

Model parameters - It is also referred as beta cofficients

E – is an error term, the impact of the unobserved variables on the dependent variable

Simple Linear Regression

In the previous example,

Housing Price = $ß_0 + ß_1 * (\text{Size of a house}) + e$

these beta coefficients are estimated by OLS (Ordinary Least Squares) method,

Input => variables can be continuous or discrete

Output => A set of coefficients that indicate the relative impact of each driver. A linear expression for predicting outcome as a function of drivers.

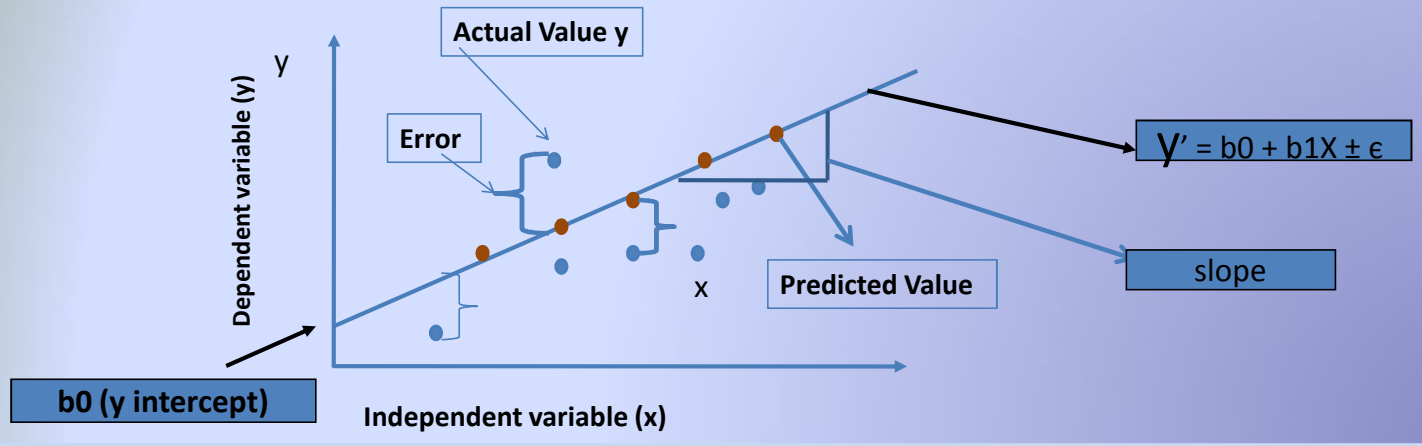
Ordinary Least Squares Estimation

It is a method used for estimating the optimal values for the beta coefficients.

OLS minimizes $\sum_{i=1}^{N} e_i^2$ (i = 1, 2, …., N)

The aim of this approach is to minimize the sum of squared residuals (the vertical distance between the observed and the estimated values)

Simple Linear Regression



What is intercept ? It's the value of Y when X = 0

- What if $ß_0 = 0$?

What is $ß_1$ which is Slope ?

What if $ß_1 = 0$?

Y is what you want to predict , say weight

$ß_1$ is the amount of influence on x, say 8

x is a predictor of y, say age

$ß_0$ is a constant , say 1

Multiple Linear Regression

When more than one independent variable is used to predict the outcome variable

price = side + uds + park

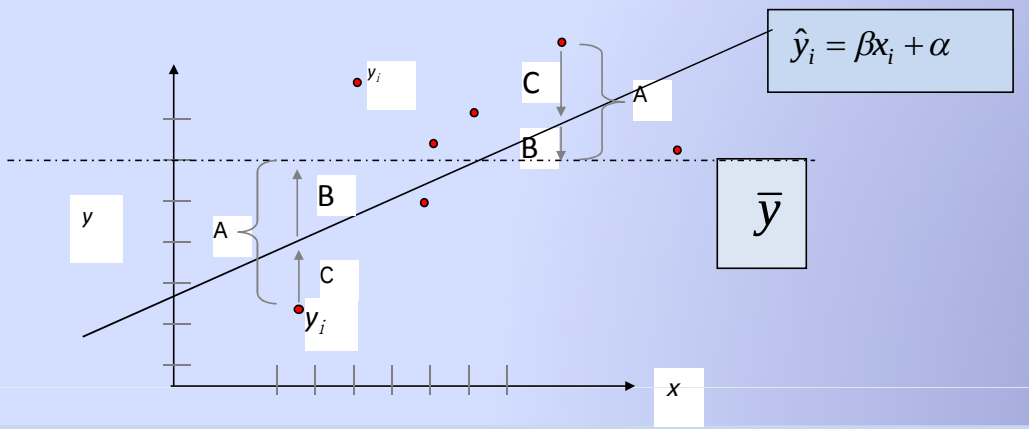
Linear Regression Model Validation

Goodness of fit measures & Regression Assumptions

Measures of Goodness of fit

Metrics		
R-Squared	Higher the better	
Adj R Squared	Higher the better	
Std. Error	Lesser the better (close to 0)	
t-statistic	p value less than 0.05	
AIC	Lower the better	
MSE	Lower the better	
MAPE (Mean Absolute percentage error)	Lower the better	$\text{mean}\left(\frac{\text{abs(pred – actuals)}}{\text{actuals}}\right)$

Regression picture



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

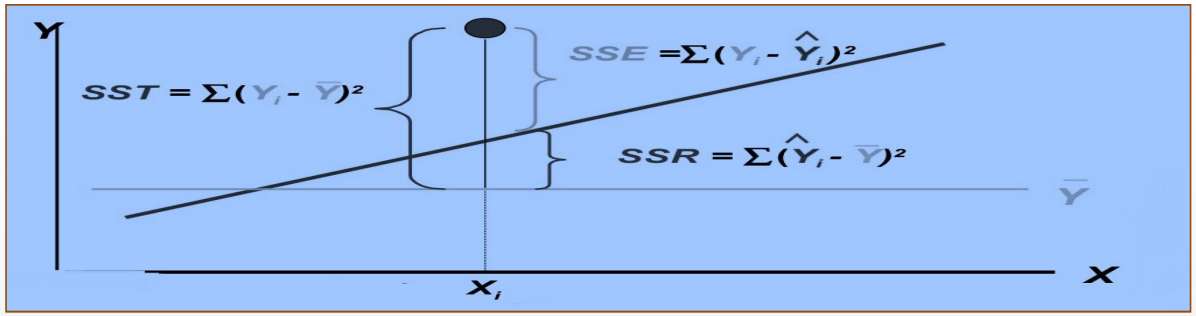
A^2 B^2 C^2

SS_{total}
Total squared distance of observations from naïve mean of y
Total variation

SS_{reg}
Distance from regression line to naïve mean of y
Variability due to x (regression)

$SS_{residual}$
Variance around the regression line
Additional variability not explained by x—what least squares method aims to minimize

Goodness of fit measures - Coefficient of Determination



R^2 Measures the proportion of variation in Y that is explained by the independent variable X in the regression model

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} = 1 - SSE/SST$$

SST = Total Sum of Squares (Measures the variation of the Y values around their mean)

SSR = Regression Sum of Squares
(Explained variation attributable to the relationship between x and y)

SSE = Sum of Squared Errors

The value of R can range between 0 and 1, and the higher its value the more accurate the regression model is. It is often referred to as a percentage.

Adjusted R Squared, Mean Squared Error, Root mean Squared error

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

R squared - Sample R Square
p - Number of predictors
N - Total sample size

Mean Square error = MSE = ∑ SSR / (n)

Root Mean Square error - This is a measure which is used often to judge the quality of prediction

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y)^2}{n}}.$$

Mean Absolute Deviation (MAD)

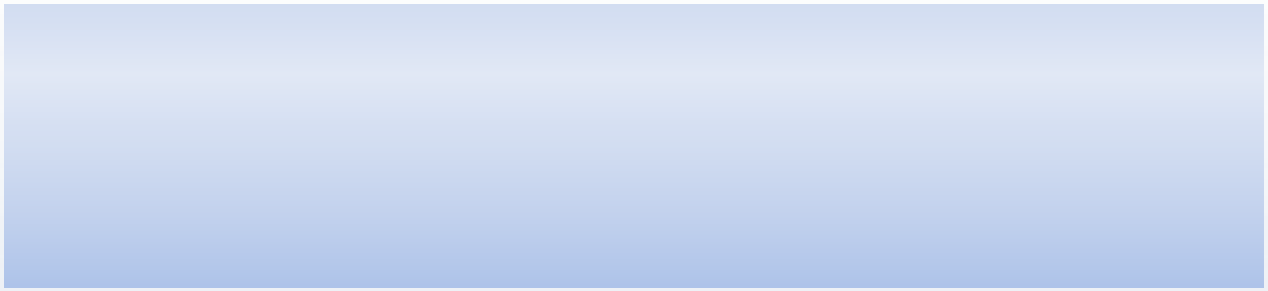
$$MAD = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n} = \frac{\sum | \text{forecast errors} |}{n}$$

Linear Regression Assumptions

Regression Assumptions

- **Linearity**
- **Homoscadasticity** – For each value of the predictors the variance of the error term should be constant
- **Multicollinearity –** No multicollinearity , predictors must not be highly correlated
- **Independence of Errors –** For any pair of observations, the error terms should be uncorrelated
- **Normality of error distribution –** Residuals should be normally distributed.

Linearity



Homoscadasticity

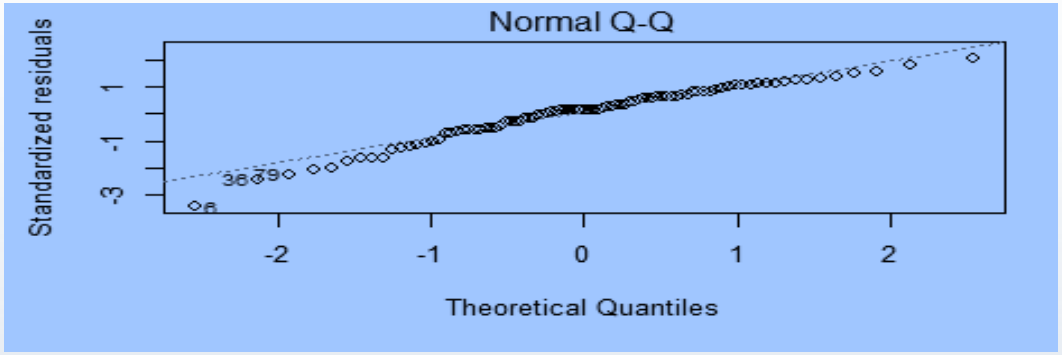
What is Heteroscedasticity ?

If the error terms do not have constant variance , they are said to be heteroscedastic. If the variance of the error term is constant , they are said to be homoscadasticity

How to detect & fix heteroscedasticity ?

- Visual Inspection
- Breusch-Pagan test
  - if the test is positive (low p value) , you should see if transformation of dependent variable

Normality of Residuals



he 'Normal Q-Q' plot gives us an visual idea whether our errors are normally distributed.

It plots the quantiles of the standardized residuals versus the theoretical quantiles. A quick visual inspection of the plot shows the data points fitting nicely around the 45-degree diagonal line, which means that our assumption of errors being normally distributed is validated.

Multicollinearity - Multiple Linear Regression

- What is multicollinearity ?
  - Multicollinearity occurs when two or more predictors in the model are correlated and provide redundant information about the response

- How to detect Multicollinearity and solve ?
  - Compute correlations between all pairs of predictors. If some r are close to -1 or 1
  - Remove one of the correlated predictors from the model
  - Calculate the **variance inflation factors** for each predictor. If the vif >= 10 then there is a problem with multicollinearity
  - The mean vif of the variance inflation factors is substantially greater than one

- Effects of Multicollinearity
  - Increased standard error of estimates of the B's (decreased reliability

    Often confusing and misleading results

Detecting and fixing Multicollinearity

```
> library(car)
Warning message:
package 'car' was built under R version 3.2.5
> vif(mcol)
      size        uds       park       misc
  1.019435   1.194178 353.505974 354.443389
>
```

since the Vif value is more than 5 that represents multicollinearity's presence

```
> mcol <- lm(tr$price~size+uds+park+misc,data=tr)
> summary(mcol)

Call:
lm(formula = tr$price ~ size + uds + park + misc, data = tr)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3343 -0.5264  0.3575  1.2106  2.8194

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.948013   0.448918   6.567 1.96e-09 ***
size         0.048641   0.001994  24.396  < 2e-16 ***
uds          0.176381   0.012446  14.172  < 2e-16 ***
park         0.105051   0.142449   0.737    0.462
misc        -0.108781   0.142077  -0.766    0.446
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.864 on 106 degrees of freedom
Multiple R-squared:  0.8957,    Adjusted R-squared:  0.8917
F-statistic: 227.5 on 4 and 106 DF,  p-value: < 2.2e-16
```

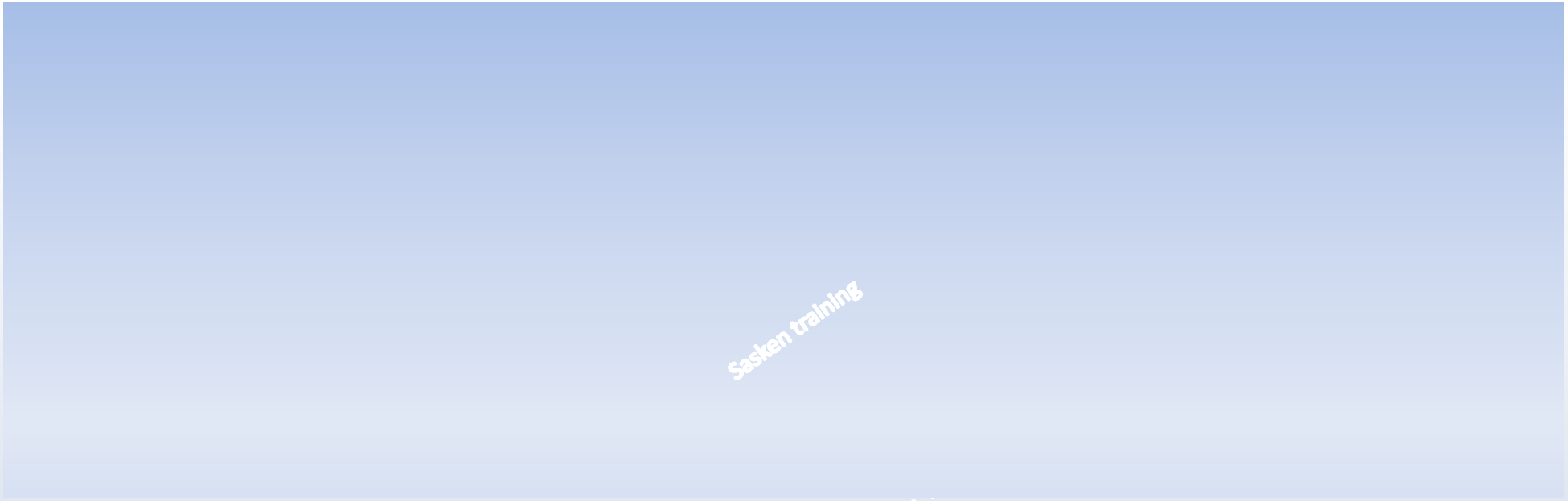
```
> mcol <- lm(tr$price~size+uds+park,data=tr)
> summary(mcol)

Call:
lm(formula = tr$price ~ size + uds + park, data = tr)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2912 -0.5676  0.3878  1.2246  2.8318

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.935477   0.447751   6.556 2.01e-09 ***
size         0.048803   0.001979  24.664  < 2e-16 ***
uds          0.175628   0.012383  14.183  < 2e-16 ***
park        -0.003833   0.008212  -0.467    0.642
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.86 on 107 degrees of freedom
Multiple R-squared:  0.8951,    Adjusted R-squared:  0.8921
F-statistic: 304.3 on 3 and 107 DF,  p-value: < 2.2e-16
```

How to fix the Multicollinearity

```
> # Model 2
> lfit2 <- lm(tr$price~size+uds,data=tr)
> summary(lfit2)

Call:
lm(formula = tr$price ~ size + uds, data = tr)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1651 -0.6111  0.3601  1.2428  2.7794

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.864900   0.419917   6.823 5.39e-10 ***
size        0.048796   0.001971  24.751  < 2e-16 ***
uds         0.173385   0.011370  15.249  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.854 on 108 degrees of freedom
Multiple R-squared:  0.8949,    Adjusted R-squared:  0.8929
F-statistic: 459.7 on 2 and 108 DF,  p-value: < 2.2e-16
```


After discarding the explanatory variables that has higher VIF value from your model, you can notice that R squared value got reduced and adj. r-squared value got increased

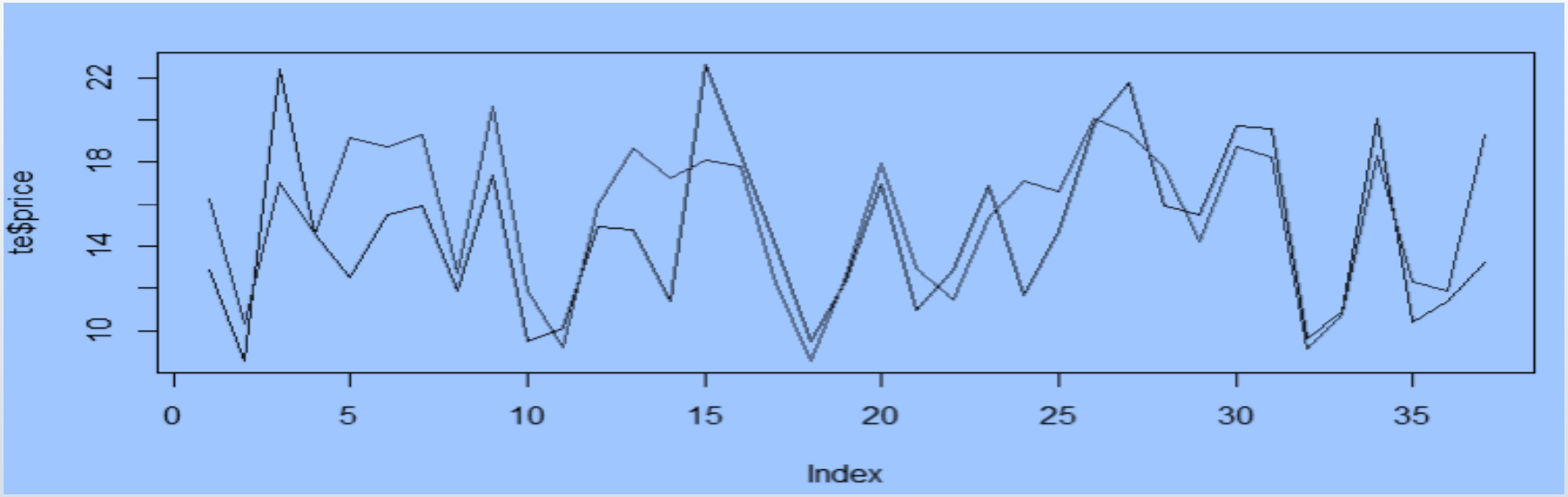
Regression Assumptions - Graphs



- ✓The two plots on the left lets us to examine the presence of heteroscadascity of errors and non linearity. Heteroscedasticity errors will appear to be u shaped or inverted u shape or will cluster close together on the left side of the plot and become wider as the fitted values increase (a funnel shape)
- ✓Normal Q-Q plot in the upper right corner , displays if the residuals are normally distributed
- ✓ It appears that the observations 30,60 and 131 , may be causing a violation of the assumption. The residuals Vs fitted plot will tell us which observations , influencing the model

Plot the Actual Vs Predicted

- plot(act_price,col="blue",type="l",ylim=c(0,50000)) lines(pre_price,col="red", type="l")



Appendix - Linear Regression Model Interpretations

Model Interpretations

Residuals	The residuals are the difference between the actual values of the variable you're predicting and predicted values from your regression--y - ŷ. For most regressions you want your residuals to look like a normal distribution when plotted. If our residuals are normally distributed, this indicates the mean of the difference between our predictions and the actual values is close to 0 (good) and that when we miss, we're missing both short and long of the actual value, and the likelihood of a miss being far from the actual value gets smaller as the distance from the actual value gets larger. Think of it like a dartboard. A good model is going to hit the bullseye some of the time (but not everytime). When it doesn't hit the bullseye, it's missing in all of the other buckets evenly (i.e. not just missing in the 16 bin) and it also misses closer to the bullseye as opposed to on the outer edges of the dartboard.
Significance Stars	The stars are shorthand for significance levels, with the number of asterisks displayed according to the p-value computed. *** for high significance and * for low significance. In this case, *** indicates that it's unlikely that no relationship exists b/w heights of parents and heights of their children

Model Interpretations

Estimated Coefficient	The estimated coefficient is the value of slope calculated by the regression. It might seem a little confusing that the Intercept also has a value, but just think of it as a slope that is always multiplied by 1. This number will obviously vary based on the magnitude of the variable you're inputting into the regression, but it's always good to spot check this number to make sure it seems reasonable.
Standard Error of the Coefficient Estimate	Measure of the variability in the estimate for the coefficient. Lower means better but this number is relative to the value of the coefficient. As a rule of thumb, you'd like this value to be at least an order of magnitude less than the coefficient estimate. In our example, the std error or the parent variable is 0.04 which is 16x less than the estimate of the coefficient (or 1.6 orders of magnitude greater).
t-value of the Coefficient Estimate	Score that measures whether or not the coefficient for this variable is meaningful for the model. You probably won't use this value itself, but know that it is used to calculate the p-value and the significance levels.
Variable p-value	Probability the variable is NOT relevant. You want this number to be as small as possible. If the number is really small, R will display it in scientific notation. In or example 2e-16 means that the odds that parent is meaningless is about 1⁄5000000000000000

Model Interpretation

Significance Legend	The more punctuation there is next to your variables, the better. Blank=bad, Dots=pretty good, Stars=good, More Stars=very good
Residual Std Error / Degrees of Freedom	The Residual Std Error is just the standard deviation of your residuals. You'd like this number to be proportional to the quantiles of the residuals in #1. For a normal distribution, the 1st and 3rd quantiles should be 1.5 +/- the std error. The Degrees of Freedom is the difference between the number of observations included in your training sample and the number of variables used in your model (intercept counts as a variable).
R-squared	Metric for evaluating the goodness of fit of your model. Higher is better with 1 being the best. Corresponds with the amount of variability in what you're predicting that is explained by the model. In this instance, ~21% of the cause for a child's height is due to the height their parent. WARNING: While a high R-squared indicates good correlation, correlation does not always imply causation.
Variable p-value	Probability the variable is NOT relevant. You want this number to be as small as possible. If the number is really small, R will display it in scientific notation. In or example 2e-16 means that the odds that parent is meaningless is about 1⁄5000000000000000

Model Interpretation

F-statistic & resulting p-value	Performs an F-test on the model. This takes the parameters of our model (in our case we only have 1) and compares it to a model that has fewer parameters. In theory the model with more parameters should fit better. If the model with more parameters (your model) doesn't perform better than the model with fewer parameters, the F-test will have a high p-value (probability NOT significant boost). If the model with more parameters is better than the model with fewer parameters, you will have a lower p-value. The DF, or degrees of freedom, pertains to how many variables are in the model. In our case there is one variable so there is one degree of freedom.