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

Supervised Machine Learning – Linear Regression

Agenda

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

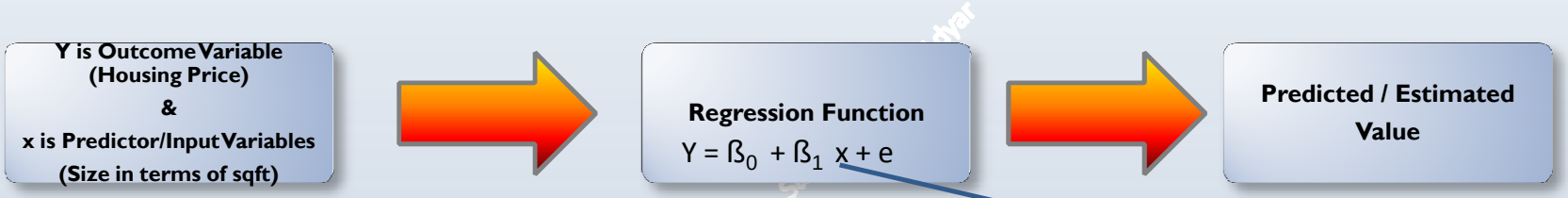
What is Regression Analysis ?

Linear regression is a predictive modeling technique, it is a supervised machine learning algorithm it is used to solve regression problems

- In Regression problem, we are trying to predict a continuous(numeric) outcome variable

Linear Regression is used to

- Understand or to Model the relationship between variables
- Predict Continuous value. To estimate/predict the value of one variable (continuous variable which is dependent variable, for ex, Housing Price) based on the other variables (for ex., sqft, No. of bed rooms)



To understand and quantify cause and effect relationship

For ex, let x be sqft of an house

What is Regression Analysis ?

Outcome Variable
&
Predictor (input) variables

Regression Function
$Y = ß_0 + ß_1 x + e$

Predicted /Estimated Value

Y = Housing Price (response/output/Dependent variable)

1

x = Size of the house (Predictor/Explanatory/input Variable)

1. Is there any relationship exist between Housing Price 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

Simple Linear Regression

Lets say you work for a realtor and you would like 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

Response /Dependent Variable

$$Y = \beta_0 + \beta_1 x + e$$

Model Parameters Or Coefficients

Explanatory/Predictor/I Input/Causal Variables

error

Where β_0 = population Y-Intercept , β_1 = Population slope, x is an independent variable , E = Random Error

- Slope - how much the line rises for each increase in x $\quad \beta_1 = \frac{\text{cov}[x,y]}{\text{var}[x]}$ $\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- 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

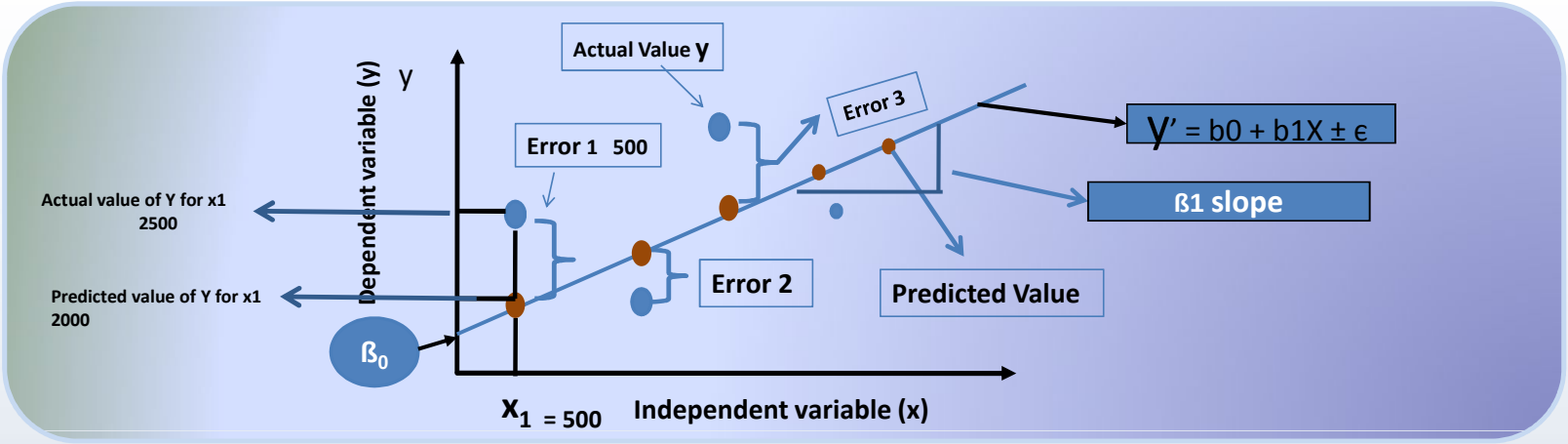
Dependent Variable - The variable y (refer previous slide) is dependent variable and also it is called as predicted variable or outcome variable

Independent Variable - The variable x (refer previous slide) is independent variable and also it is called as predictor variable or explanatory variable. It can be numeric or categorical

Model parameters - It is also referred as beta coefficients

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

Simple Linear Regression



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

- What if $ß_0 = 0$?

If $\beta_0 = 4$, then Average *Y* Is Expected to Be 4 When *X* Is 0

What is $ß_1$ which is Slope ?

What if $ß_1 = 0$?

— **Estimated *Y* Changes by β_1 for Each 1 Unit Increase in *X***

- **If $\beta_1 = 2$, then *Y* Is Expected to Increase by 2 for Each 1 Unit Increase in *X***

Ordinary Least Squares

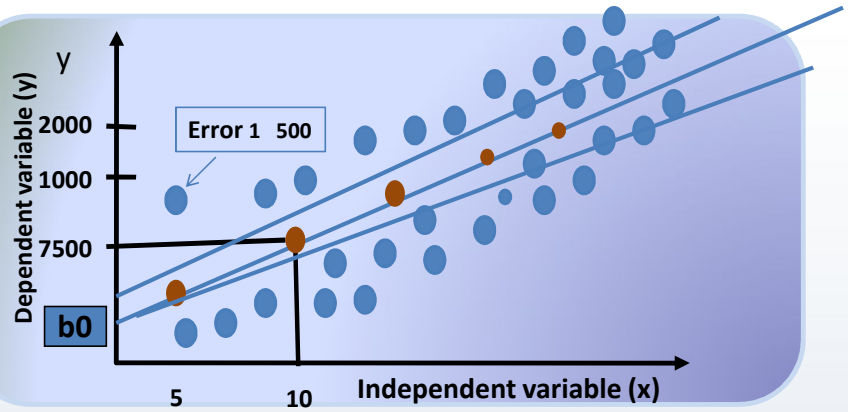
In the previous example,

Housing Price = $ß_0$ + $ß_1$ * (Size of a house) + e

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

How do you determine which line 'fits best'?

'Best fit ' means difference between actual Y values and predicted y values



The procedure to find the best fit is known as least squares method . It is a method used for estimating the optimal values for the beta coefficients.

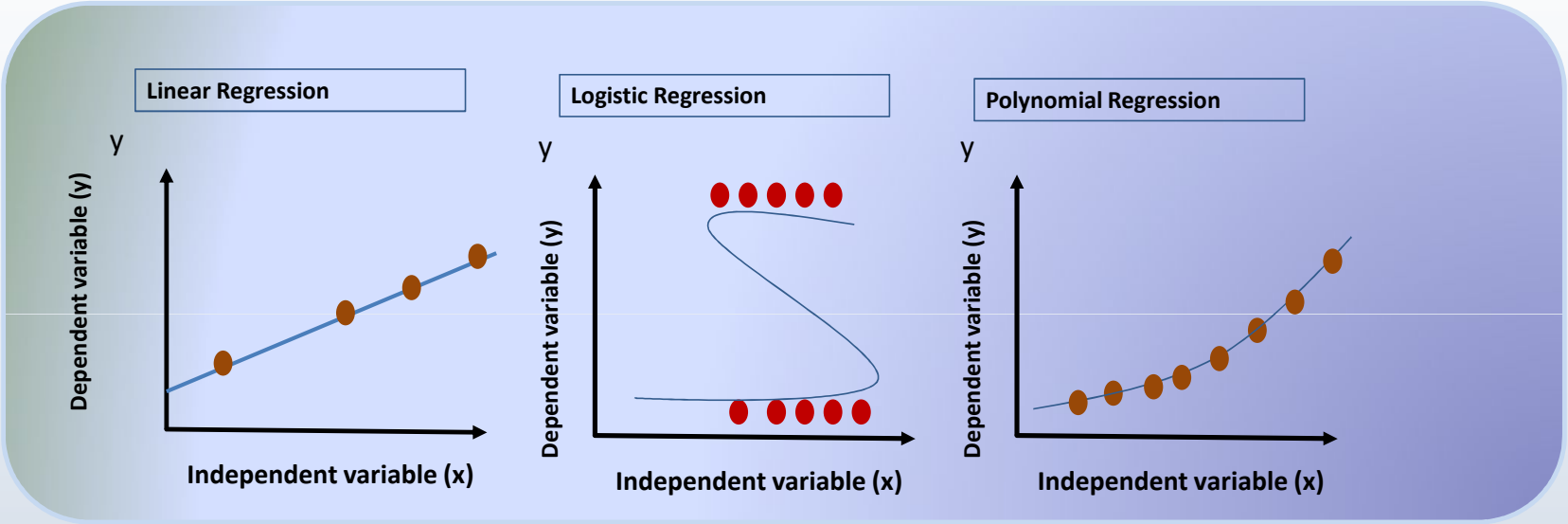
OLS minimizes $\sum_{i=1}^{N} e_i^2$ (i = 1, 2, …., N)

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

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.

Types of Regression



Simple Linear Regression in R & Python

Now, lets see now we can implement Linear Regression in R & Python

Ordinary Least Squares Interpretation

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- Model Validation
- Perform Prediction

This is the simple data set we have used to fit the linear regression model

housingdf <- read.csv("D:/dat1/housingData.csv",sep=",")

	size	uds	park	price	misc
1	230.1	37.8	69.2	22.1	69.3
2	44.5	39.3	45.1	10.4	58.4
3	17.2	45.9	69.3	9.3	69.2
4	151.5	41.3	58.5	18.5	58.5
5	180.8	10.8	58.4	12.9	45.1
6	8.7	48.9	75.0	7.2	75.0
7	57.5	32.8	23.5	11.8	23.5
8	120.2	19.6	11.6	13.2	11.6
9	8.6	2.1	1.0	4.8	1.0
10	199.8	2.6	21.2	10.6	21.2

Ordinary Least Squares in R

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- Model Validation
- Perform Prediction

Split the dataset into training and test data sets

Training dataset -> to fit/train the model
Test dataset -> to perform cross validation and make predictions

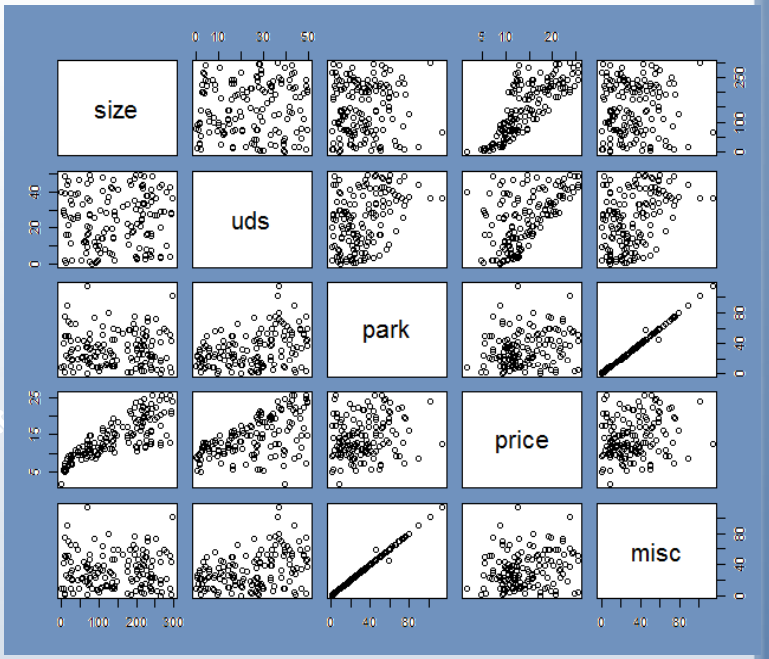
```
library(caret)
ind <- createDataPartition(housingdf$price, p=0.6,list = FALSE)
train <- housingdf[ind,]
dim(train)
head(train)
test <- housingdf[-ind,]
```

Here we divide the data into 60:40 ratio so that 60% will be present as training set and remaining 40% as the test set.

Ordinary Least Squares

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- Model Validation
- Perform Prediction

- The relationship for both size and price , we could see a positive trend
- We could see there was no linear relationship exist between park and price



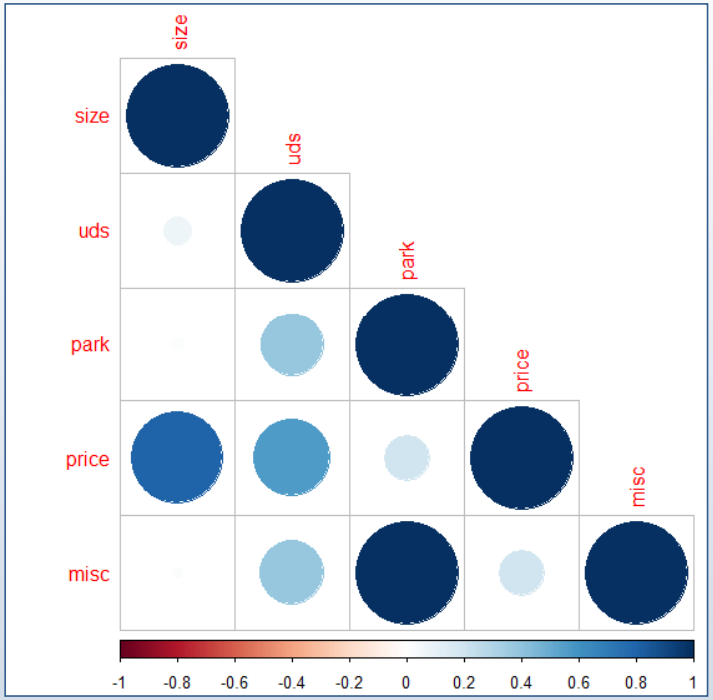
Ordinary Least Squares

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- Model Validation
- Perform Prediction

- By performing Correlation Analysis, we can understand not only the relationship between the independent and dependent variable but also we can understand if there are any relation exists among independent variables
- For visualizing the same , we can use a corrplot()
- library(corrplot)
- corrplot(housing.cor, type="lower")

- We could see a strong relationship exists between the independent variables park and misc itself
- When correlation exist among X's is high, OLS has very little information to estimate. This makes us relatively uncertain about our estimate

Positive correlations are displayed in blue and **negative correlations** in red color. Color intensity and the size of the circle are proportional to the **correlation coefficients**.



Ordinary Least Squares

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- Model Validation
- Perform Prediction

```
> 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 (Variable Inflation Factor) 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
```

ulticollinearity exists whenever two or more of the predictors in a regression model are moderately or highly rrelated. From the top image, we could that the Vif value is more than 5 that represents multicollinearity's presence

Ordinary Least Squares- How to fix the multicollinearity

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- Model Validation
- Perform Prediction

```
> 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 (Variable Inflation Factor) 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
```

ter discarding the misc variable, we could notice that the adjusted R squared value got slightly increased

Ordinary Least Squares

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- Model Validation
- Perform Prediction

```
#    fit the model
 housing.fit <- lm(price~size,    data=train)

# To describe model , use summary() as below
 summary(housing.fit)
```

Some of the important metrics are
1. Rsquared value
2. p-value


```
> summary(housing.fit)

Call:
lm(formula = price ~ size, data = housingdf)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5934 -1.7324  0.0496  1.8996  6.8348

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.054122   0.511608   13.79   <2e-16 ***
size        0.049327   0.003041   16.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.163 on 146 degrees of freedom
Multiple R-squared:  0.6431,    Adjusted R-squared:  0.6406
F-statistic:   263 on 1 and 146 DF,  p-value: < 2.2e-16
```

Ordinary Least Squares – Model Validation using metrics

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- **Model Validation**
- Perform Prediction

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 output Interpretation

```
> summary(lmfit)

Call:
lm(formula = price ~ size, data = hdf)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5934 -1.7324  0.0496  1.8996  6.8348

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.054122   0.511608   13.79   <2e-16 ***
size        0.049327   0.003041   16.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.163 on 146 degrees of freedom
Multiple R-squared:  0.6431,    Adjusted R-squared:  0.6406
F-statistic:   263 on 1 and 146 DF,  p-value: < 2.2e-16
```

This is the hypothesis that the slope for size is 0

Gives an Idea of how far Observed price are from the estimated or fitted Price

Its an estimated mean Y value when all X's are Zero

This is the estimated effect of Housing size on Housing Price

*** 99.9% confident – *** represents your model is 99% confident that this indep. variable or value is significant

** 99% confident

* 95% confident

. 90% confident

This tests the Null hypothesis that all the model coefficients are 0

Anova Table

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- Model Validation
- Perform Prediction

```
> anova(housing.fit )
Analysis of Variance Table

Response: price
           Df  Sum Sq Mean Sq F value    Pr(>F)
size        1 1545.99 1545.99  177.33 < 2.2e-16 ***
Residuals 88  767.21    8.72
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

SSR/k => 1545.99/1 => Mean Sq
SSQ/Residuals => 767.21/88 => 8.72 => Mean Sq Residuals
F value = Mean Sq/ Residuals = 1545.99/8.72

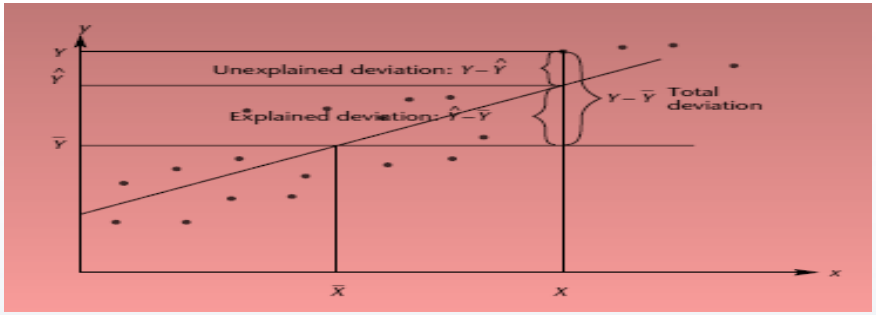
Ordinary Least Squares Interpretation

For every unit increase in housing size , we expect to see an increase in Housing Price by 40$

Positive sign on the coefficient on housing size implies a positive relationship between housing size and Housing price

What does unit increase implies ?

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 = SSR + SSE

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.

Mean Square error = MSE = ∑ SSR / (n)

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

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

Ordinary Least Squares – performing prediction

- Import the data
- Split the dataset
- Exploratory Analysis
- Fit the model
- Model Validation
- Perform Prediction

```
#    perform prediction
 pred <- predict(housing.fit, newdata=test)
 print (pred)
              ( OR )
fitted(housing.fit)
```

```
> summary(housing.fit)

Call:
lm(formula = price ~ size, data = housingdf)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5934 -1.7324  0.0496  1.8996  6.8348

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.054122   0.511608   13.79   <2e-16 ***
size        0.049327   0.003041   16.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.163 on 146 degrees of freedom
Multiple R-squared:  0.6431,    Adjusted R-squared:  0.6406
F-statistic:   263 on 1 and 146 DF,  p-value: < 2.2e-16
```

Ordinary Least Squares – Confidence Level Interpretation

95% of the time housing size increases , housing price will increase by these values

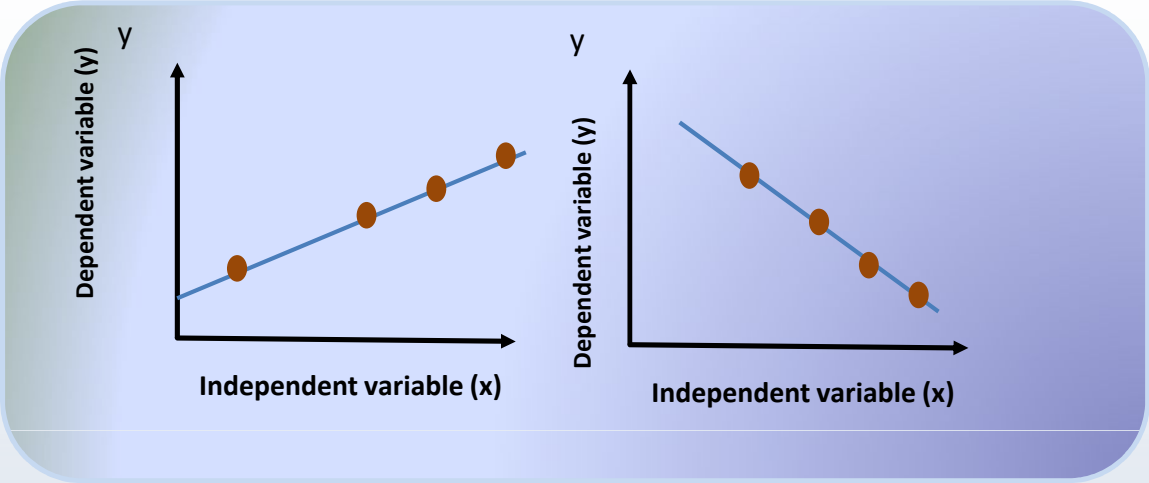
```
> confint(lmfit,level=0.95)
                2.5 %      97.5 %
(Intercept) 6.04300769 8.06523694
size        0.04331594 0.05533754
```

- At every time X (Housing size) increases by 1 unit, 95% of the time Y (Housing price) will increase between 0.04$ to 0.05$.
- 95% of the time y (Housing Price) will increase between these ranges. Lower the standard error the narrower the range of confidence interval

Linear regression Assumptions – Model Validation

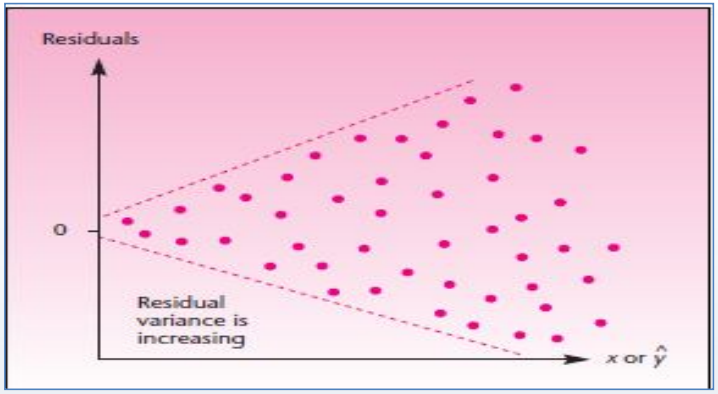
- **Linearity**
- **Homoscadasticity** – For each value of the predictors the variance of the error term should be constant
- **Multicollinearity –** No multicollinearity , predictors must not be highly correlated
- **Independence of Errors –** For any pair of observations, the error terms should be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

Linearity

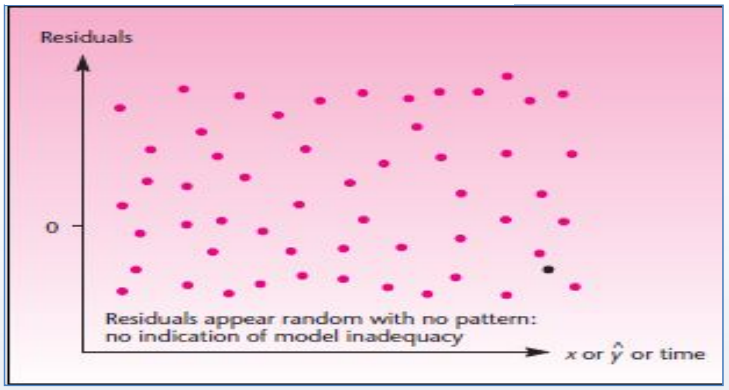


Homoscadasticity

A Residual Plot Indicating Heteroscedasticity



A Residual Plot Indicating No Heteroscedasticity



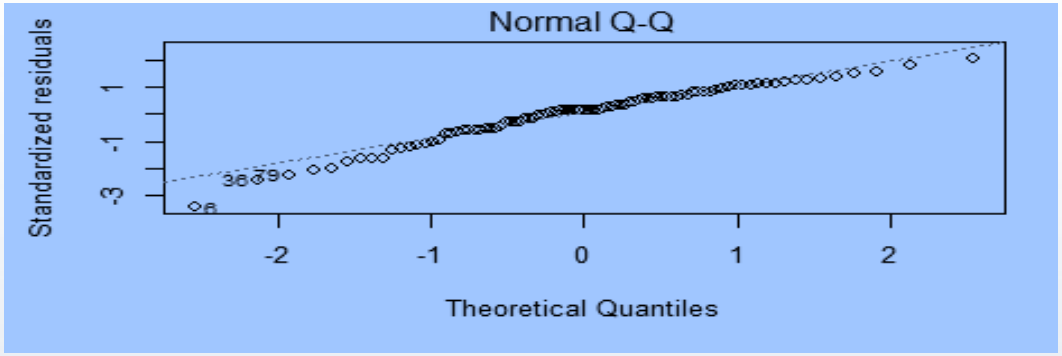
What is Heteroscedasticity ?

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



The '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

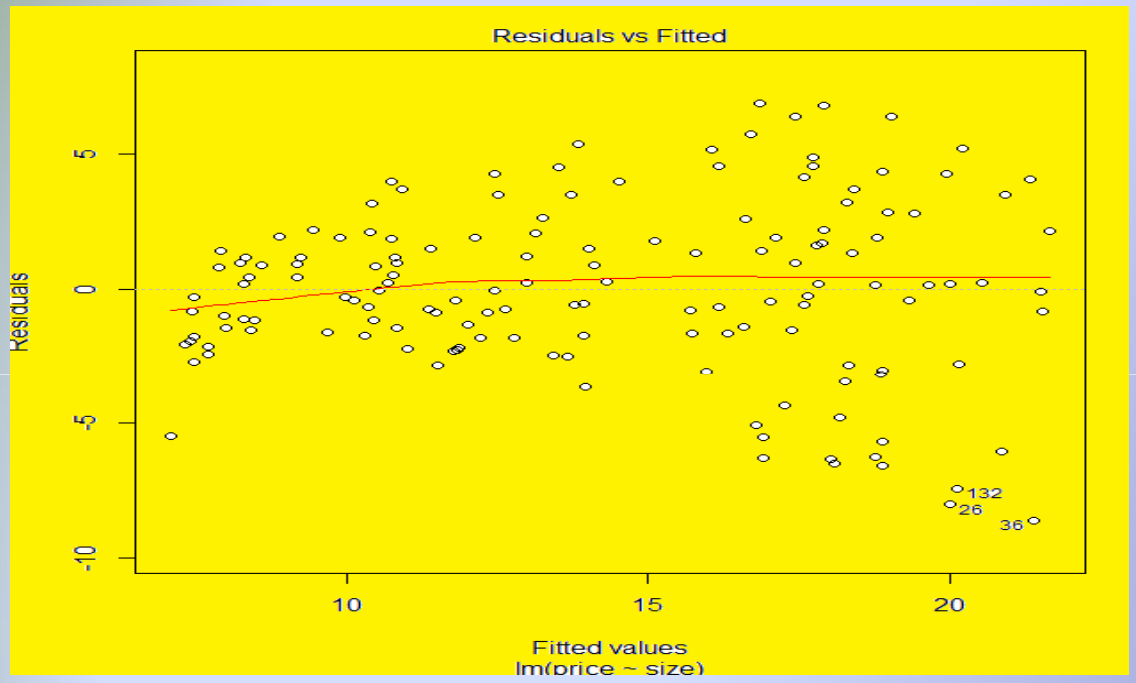
- What is multicollinearity ?
  - Multicollinearity occurs when two or more predictors in the model are correlated and provide redundant information about the response

- 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

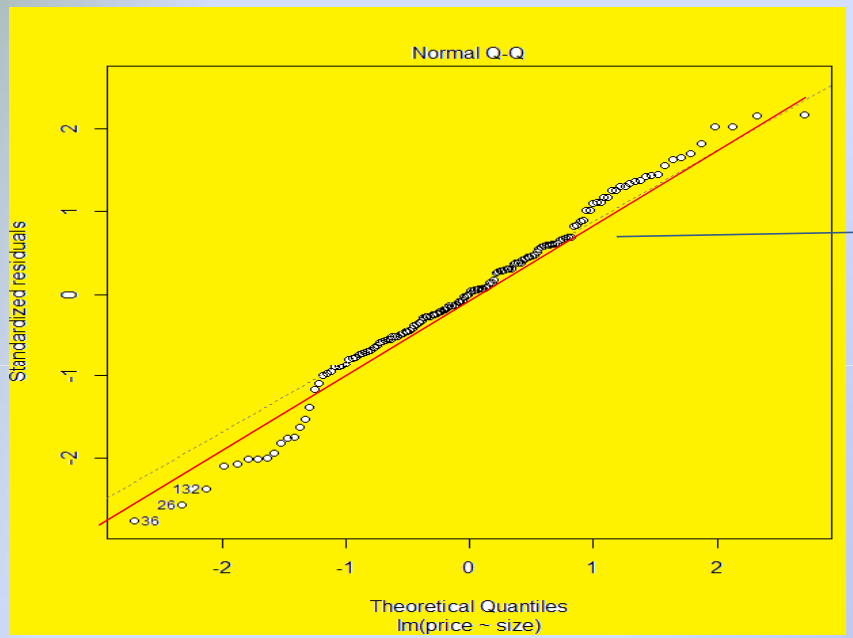
Linear Regression Assumptions – Interpreting the plots



The red line should be fairly flat

Linear Regression Assumptions – Interpreting the plots

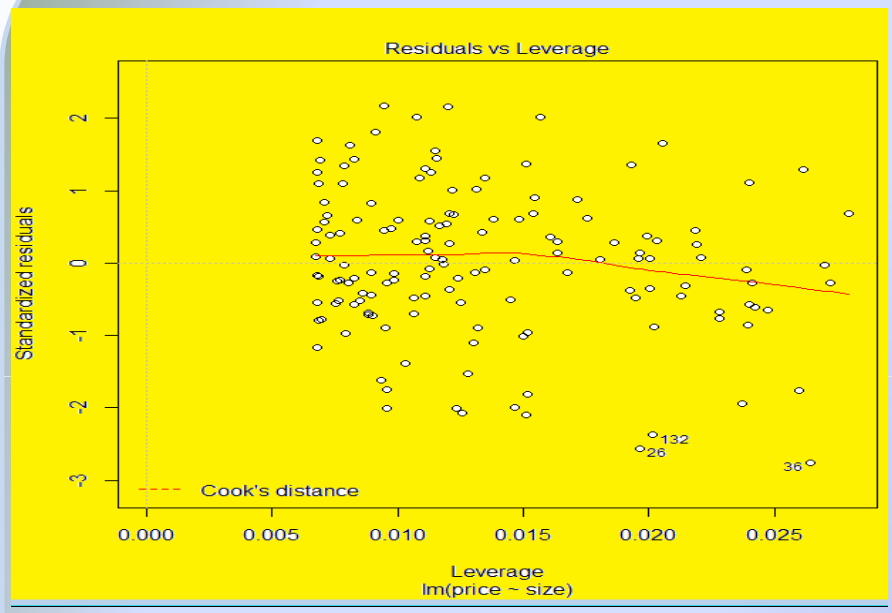
Quantile – Quantile plot

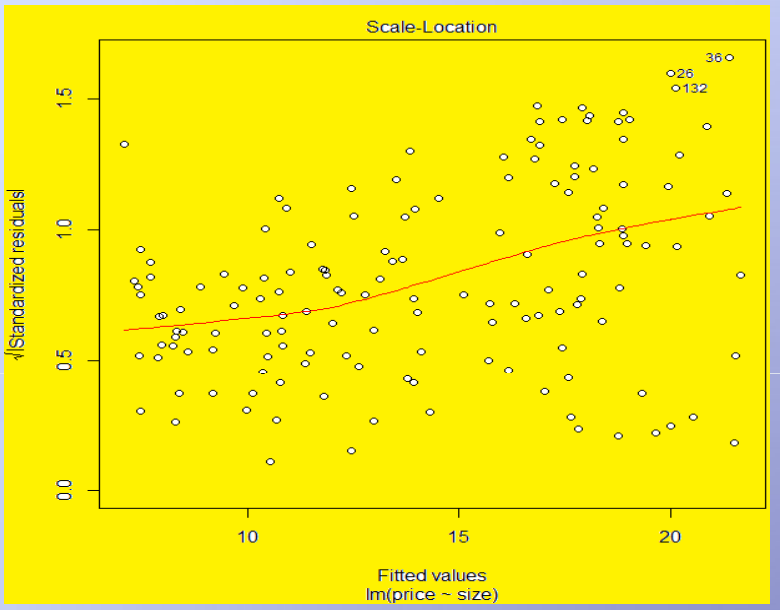


If our Y values or errors are normally distributed these points should fall roughly on a diagonal line

X axis is the expected residuals if the errors/residuals are normally distributed

Linear Regression Assumptions – Interpreting the plots





X axis is the expected residuals if the errors/residuals are normally distributed

Multiple Linear Regression

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

price = side + uds + park

Step wise Regression – Feature Selection

- Stepwise regression is a way to build a model by adding or removing predictor variables

We can use the below in R to perform stepwise regression

```
library(MASS)
stepAIC(step.model, direction = “backward”)
stepAIC(step.model, direction = “forward”)
stepAIC(step.model, direction = “both”)
```

How Step wise Regression works ?

- **Forward Selection**
  - Forward selection starts with a model with no variables.
- **Backward Elimination**
  - This procedure works in a manner opposite to forward selection. We start with a model containing all k variables.
- **Stepwise Regression**
  - This is probably the most commonly used, wholly computerized method of variable selection. The procedure is an interesting mixture of the backward elimination and the forward selection methods

How Step wise Regression works ?

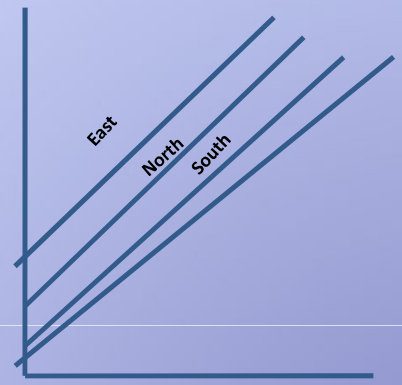
```
> summary(housing.fit)

Call:
lm(formula = price ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6174 -0.6108  0.3145  1.0593  2.4932

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.131157   0.426993   7.333 1.16e-10 ***
size         0.048861   0.002034  24.023  < 2e-16 ***
uds          0.179389   0.011699  15.334  < 2e-16 ***
park        -0.006813   0.007819  -0.871    0.386
misc               NA         NA      NA       NA
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.551 on 86 degrees of freedom
Multiple R-squared:  0.9184,    Adjusted R-squared:  0.9155
F-statistic: 322.4 on 3 and 86 DF,  p-value: < 2.2e-16
```

NA as a coefficient in a regression indicates that the variable in question is linearly related to the other variables

Dummy Variables

- Often we must work with categorical Independent variables
- In regression we call these dummy or indicator variables or one hot encoding.
- Categorical variable with k levels or categories , requires (k-1) dummy or indicator variables
  - Region variable has East, North, South and West , so 4-1 = 3 dummy variables

Region	X1	x2	X3
North	1	0	0
South	0	1	0
East	0	0	1
West	0	0	0



- By default, the category that R chooses to be the reference or baseline , is the first category that appears alphabetically or numerically (if categories are coded using 0,1,2,…)

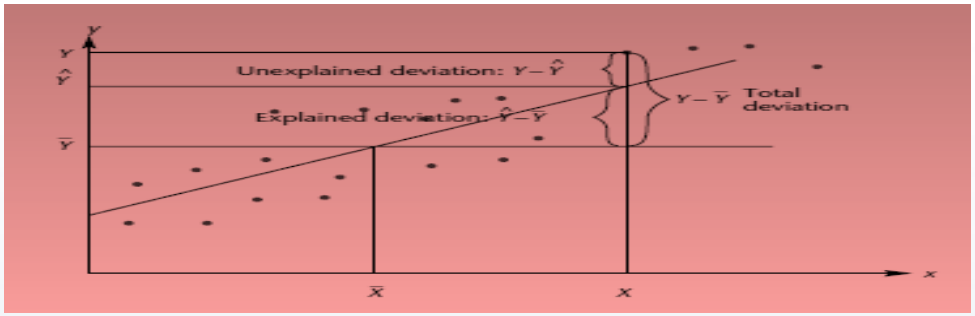
Dummy Variables Interpretation

$E(Y) = ß_0 + ß_1 x + ß_2 x2 + ß_3 x3 + ß_4 x + e$

Expected value of home price given the region is East x1 = 1
$E(Y \mid East) = ß_0 + ß_1 (1) + ß_2 (0) + ß_3 (0) + ß_4 (0) + e$

Expected value of home price given the region is NOT East,South,West x1 ,x3,x4= 0
$E(Y \mid Not\ East, South, and\ West) = ß_0 + ß_1 (0) + ß_2 (1) + ß_3 (0) + ß_4 (0) + e$

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 = SSR + SSE

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

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
>
```

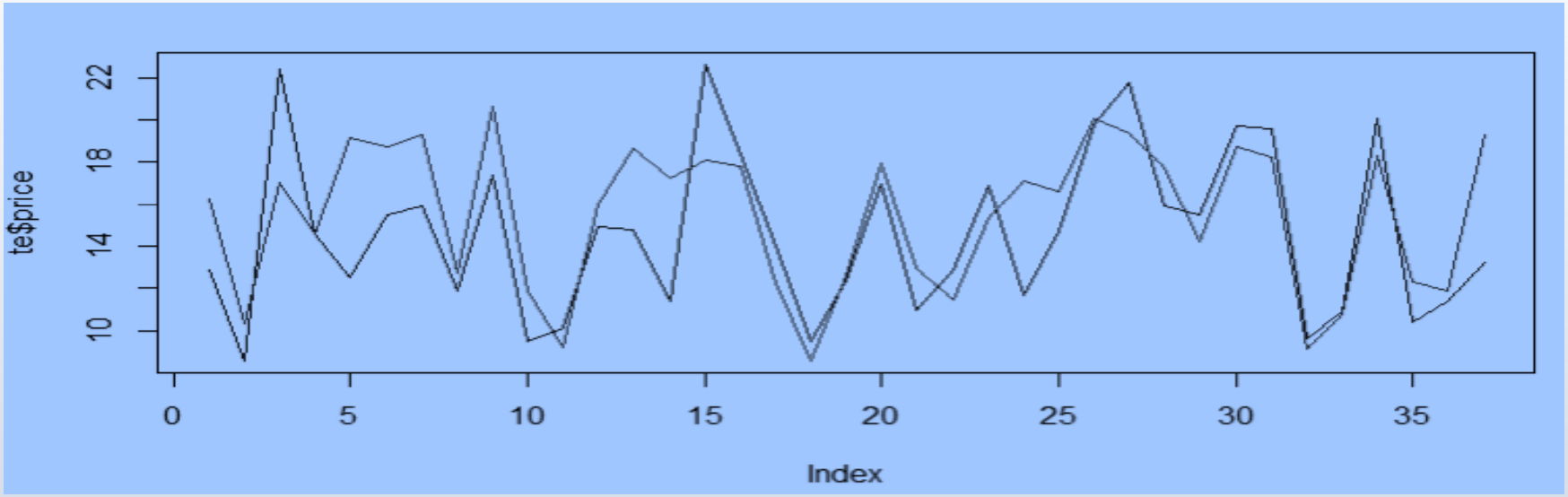
since the Vif value is more than 5 that represents multicollinearity's presence

```
> 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
```

Plot the Actual Vs Predicted

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



Linear Regression Quiz

What is the Difference between Anova and Linear Regression ?

Linear regression is used to analyze continuous relationships; however linear regression Is essentially the same as ANOVA

In Anova, we calculate means and deviations of our data from the means

In Linear regression, we calculate the best line through the data and calculate the deviations from the data from this line.

The F ratio can be calculated in both

Linear Regression Quiz

Question 1 1 pts

What is the difference between R2 and Adjusted R2 in a linear regression model?

- They are the same
- Adjusted R2 is better because it goes up only if significant variables are added to the model
- R2 is better because it goes up when significant variables are added to the model
- Adjusted R2 is better because it is always lower than R2

Question 2 1 pts

If you have a categorical variable with 4 levels, and you have created dummy variables for each of the categories, how many of these can you include in a linear regression model?

- The first three based on alphabetic or numeric order
- Any 3
- All 4

Linear Regression Quiz

Question 3 1 pts

If you include any three dummy variables based on a categorical variable that had 4 levels, how will you interpret the coefficients on these dummy variables?

- The beta coefficient on a dummy variable captures the relative difference of the change in DV for a unit change in the dummy variables, where relative impact is over the baseline dummy that was not included in the model
- The beta coefficient on any of the three dummy variables captures the change in the DV for a unit change in the dummy variables
- The beta coefficient on a dummy variable captures the relative difference of the change in DV for a unit change in the dummy variables, where relative impact is over the intercept in the model

Question 4 1 pts

Is it possible to include a logarithmic relationship in a linear model?

- Yes
- No

Linear Regression Quiz

Question 5 1 pts

Is it possible to model a non-linear relationship between an IV and a DV in a linear model?

- ○ Yes
- ○ No

Question 6 1 pts

Beta coefficients must always have a p value of 0.05 or lower to be considered significant

- ○ True
- ○ False

Linear Regression Quiz

Question 7 1 pts

What is the correct interpretation of R2?

- It is a measure of accuracy captured by the model
- It is a measure of variance captured by the model
- It is a measure of preciseness captured by the model

Question 8 1 pts

Which of the following is a violation of OLS assumptions?

- There should not be high correlation between the DV and the IVs
- There should not be high correlation between the IVs
- There should not be any correlation between IVs
- There should not be any correlation between the DV and the IVs

Linear Regression Quiz

Question 9 1 pts

A parsimonious model is one that

- Is easy to understand
- Includes all variables required to explain all the variation in the DV
- Is a mis-specified model
- Includes a few variables that explain most of the variation in the DV

Question 10 1 pts

Which of the following is a violation of OLS assumptions?

- The average error term is zero
- The variance of the error term is not constant
- The model is linear in parameters
- The IVs are not highly correlated

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.

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.

Leverage, Influence, and Cooks Distance

Cook's D is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question