Statistical Computing with R: Masters in Data Sciences 503, S20 First Batch, SMS, TU, 2021

Shital Bhandary

Associate Professor

Statistics/Bio-statistics, Demography and Public Health Informatics

Patan Academy of Health Sciences, Lalitpur, Nepal

Faculty, Data Analysis and Decision Modeling, MBA, Pokhara University, Nepal

Faculty, FAIMER Fellowship in Health Professions Education, India/USA.

Review Preview

- One-way ANOVA

- Linear relationship

- Covariance
- Correlation
- Simple Linear regression model fit, interpretation and residual analysis
- Simple Linear Regression prediction and Machine Learning

Comparing means of an outcome variable across another variable with more than two categories:

- **One-way ANOVA**

- H_0: $\mu_1 = \mu_2 = \mu_3$
- H_1: At least one pair of means are not equal

- **If H_1 is accepted, pairwise comparison (post-hoc) test must be done to find the significant pairs!**

- Compare mpg (miles per gallon) by cars with different gear (numbers of gears) using "mtcars" data

- Dependent variable = mpg
- Independent variable = gears

Assumptions of 1-way ANOVA:

- Same as two-samples t-test:

- Dependent variable must be "normally distributed"

- Variance across categories must be same

- Normally distributed:
  - Test of normality by each category

- Homogenous variance:
  - var.test is not useful (>2 groups)
  - Levene's Variance test is preferred
  - It is available in the "car" package
  - library(car)
  - leveneTest(y~x, data=data)
  - **x must be categorical i.e. factor!**

1-way ANOVA assumptions checks:

Normality by categories:

- with(mtcars, shapiro.test(mpg[gear == 3]))

W = 0.95833, p-value = 0.6634

- with(mtcars, shapiro.test(mpg[gear == 4]))

W = 0.90908, p-value = 0.2076

- with(mtcars, shapiro.test(mpg[gear == 5]))

W = 0.90897, p-value = 0.4614

Equal variance among categories:

library(car)

leveneTest(mpg ~ gear, data=mtcars)

Result:

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	2	1.4886	**0.242429**

Levene's Test is a GOF test, so group variances are equal as p-value>0.05.

So, Classical 1-way ANOVA can be used now!

- summary(aov(mpg ~ gear, data = mtcars))
- Since F-test p-value <0.05, we accept H1. At least one of the mean pairs are not equal!
- This means, post-hoc test or pairwise comparison is required!
- **Fisher's LSD uses pairwise t-tests (not good)!**
- For classical 1-way ANOVA, Tukey HSD is the best post-hoc test!
- TukeyHSD (aov(mpg ~ gear, data = mtcars))

	Df	SumSq	MeanSq	Fvalue	Pr(>F)
gear	2	483.2	241.62	10.9	0.000295
Residuals	29	642.8	22.17		

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = mpg ~ gear, data = mtcars)

$gear

	diff	lwr	upr	p adj
4-3	**8.426667**	3.9234704	12.929863	0.0002088
5-3	5.273333	-0.7309284	11.277595	0.0937176
5-4	-3.153333	-9.3423846	3.035718	0.4295874

Check this result with the simple linear model (regression):

- summary(lm(mpg ~ gear, data = mtcars))

- P-value are reported without correcting them i.e. simple t-test were used, which can be checked with this command in R/R Studio:

- **pairwise.t.test(mtcars$mpg, mtcars$gear, p.adj = "none")**
- 3 4
- 4 7.3e-05 (3 vs 4) --
- 5 0.038 (3 vs 5) 0.218 (4 vs 5)

- What is the interpretation?
- **Why gear = 3 category is omitted in the result?**

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	16.107	1.216	13.250	7.87e-14	***
gear[T.4]	**8.427**	1.823	4.621	7.26e-05	***
gear[T.5]	**5.273**	2.431	2.169	0.0384	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- R automatically creates 3 dummy variables for 3 categories of gear variable i.e. 3, 4 and 5 and uses only last two of them in the model and takes the first one as reference!
- gear[T.3] = 1 if gear = 3, else 0
- gear[T.4] = 1 if gear = 4, else 0
- gear[T.5] = 1 if gear = 5, else 0

Measures of linear relationship:

- Two continuous variables
- Assumption:
- Two continuous variables have linear or “tentative” linear relationship
- Assessed using “scatterplot”

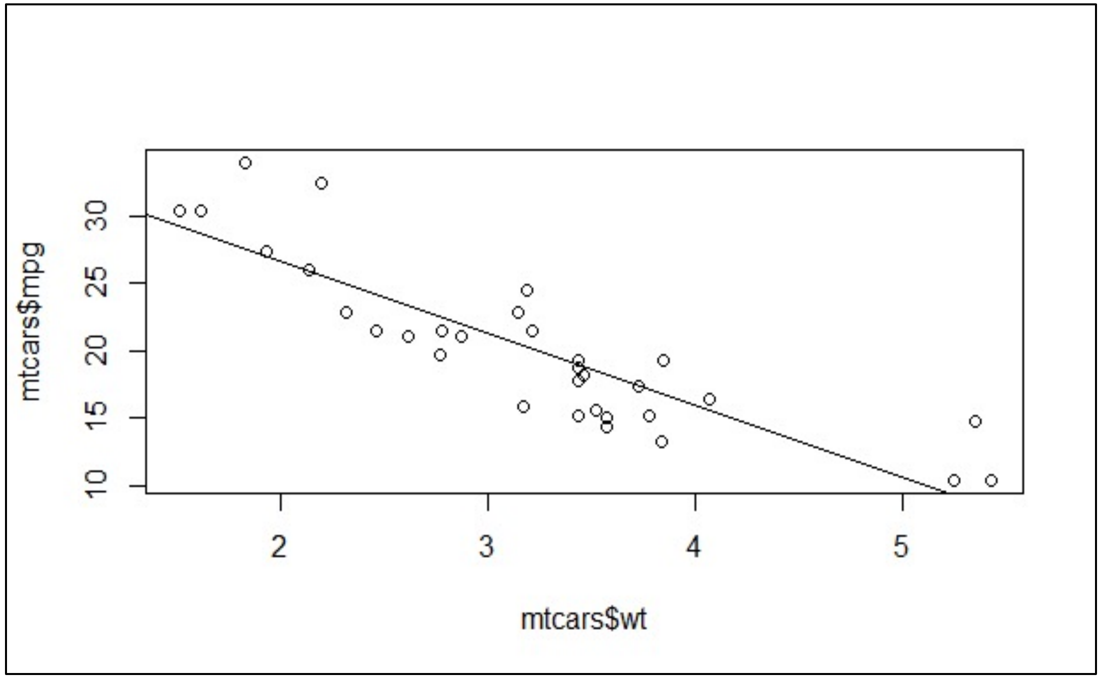
- Measures of linear relationship:
- Covariance
  - Limitations
- Pearson’s Correlation Coefficient
  - Limitations
- Simple Linear regression

Covariance

- It measures the **linear** relationship between two quantitative variables.
    1. Positive values indicate a positive linear relationship; negative, a negative linear relationship.
    2. Close to zero means there is not much of a linear relationship.
    3. The magnitude of covariance is difficult to interpret.
    4. Covariance has problems with units (like feet compared to inches).

Example: which one is more linear?

- plot(mtcars$wt, mtcars$mpg)
- plot(mtcars$disp, mtcars$mpg)
- plot(mtcars$hp, mtcars$mpg)
- plot(mtcars$drat, mtcars$mpg)
- plot(mtcars$qsec, mtcars$mpg)



There is a “tentative” linear relationship between mpg and weight variables! So, we can use measures of linear relationship for these variables!

Covariance between WT and MPG variables:

cov(mtcars$wt, mtcars$mpg)

- **-5.116685**

Do as follows now:

- Convert the weight (wt) variable measured in pound to kilogram and store it a new variable
- Compute the covariance of weight in KG and MPG now!
- -2.325766

Sample covariance for a sample of size *n* with the observations:

$$S_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Population covariance:

$$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Pearson’s Correlation Coefficient (r) to measure linear relationship:

- Measure the strength and direction of the linear relationship between two quantitative variables.
- A relative measure of strength of association (relationship) between 2 variables or a measure of strength per unit of standard deviation, s_x * s_y .
- **Solves “units” and “magnitude” problems of covariance.**

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$s_{xy} = \text{sample covariance} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_x = \text{sample standard deviation of } x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$s_y = \text{sample standard deviation of } y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

Correlation of WT, WT2 and MPG variables:

cor(mtcars$wt, mtcars$mpg)

- **-0.8676594**

cor(mtcars$wt2, mtcars$mpg)

- **-0.8676594**

Interpretation (Pearson):

- Low degree: <0.25
- Medium degree: 0.25-0.75
- High degree:>0.75

- How to check if this correlation is a valid linear correlation?
- We need to do the hypothesis testing:
- H_0: Linear correlation is zero i.e. $\rho = 0$.
- H_1: Linear correlation is NOT zero i.e. $\rho \neq 0$.

Test of “true” linear correlation of WEIGHT and MPG variables:

- cor.test(mtcars$wt, mtcars$mpg)
- cor.test(mtcars$wt2, mtcars$mpg)

Interpretation:

- Since p-value < 0.05, we accept H1 (Decision)
- This means the true linear correlation coefficient is NOT zero so computed sample estimate of this correlation coefficient as -0.87 is a valid estimate (Conclusion)

```
Pearson's product-moment correlation

data:  mtcars$wt and mtcars$mpg
t = -9.559, df = 30, p-value = 1.294e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9338264  -0.7440872
sample estimates:
       cor
-0.8676594
```

Limitation of Linear correlation coefficient:

- It provides the magnitude and direction of the relationship between two linearly related quantitative variables
- It does not provide the estimate of change in dependent variable with respect to the change in the independent variable

- Thus, it is required to use a simple linear regression i.e.

$$y = a + bx$$

- Simple linear regression is an extension of the simple linear correlation
- **But it come with many assumptions!**

Simple Linear Regression:

A simple linear regression model of Y on X in stochastic form (population) in statistics is written as:

$$Y = \alpha + \beta X + u$$

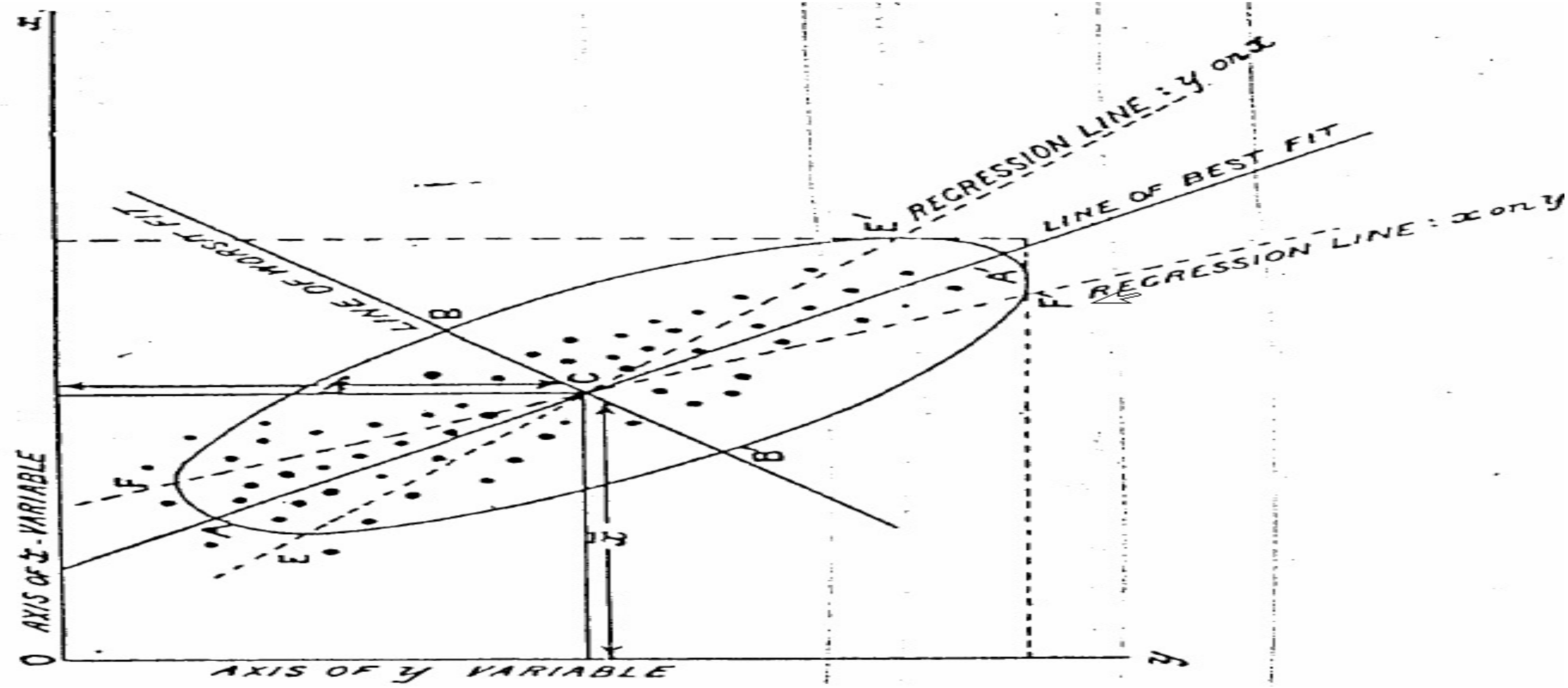
where α and β are parameters called y-intercept and slope respectively, and u is called error or disturbance term, which is erratic or random in nature.

- For given n pairs of data values (x_1, y_1), (x_2, y_2), (x_3, y_3),. . ., (x_n, y_n) of (X, Y), the estimated model is written as:

$$\hat{y} = a + bx$$

- y-hat is estimated value of Y based on a and b, which are least square estimates of α and β respectively.
- We need to calculate best solutions of n equations each containing two unknown parameters α and β using **OLS method**.

LINE OF BEST FIT with minimizing error - OLS (Ordinary Least Square Method)



Simple Linear Regression Assumptions:

- Dependent variable: Normal
- Dependent and Independent variables: Linear

Regression Model:

- Coefficient of determination > 0.50
- Regression ANOVA must be significant statistically
- Y-intercept (a) an slope (b) must be statistically significant

If these conditions are satisfied then it is called a BLUE estimate!

- Regression Model Residuals or Errors i.e. “y – yhat”:
  - Linearity of residuals
  - Independence of residuals (for time series)
  - Normality of residuals
  - Equal variance (Homoscedasticity) of residuals
- Also known as LINE test
  - Each of these assumptions must be checked with graphs and statistical methods

Simple Linear Regression between MPG and WT variables:

- Dependent variable MPG follows normal distribution (checked!)
- Dependent variable MPG and independent variable WT has "tentative" linear relationship
- We can move forward!

- We need to check after fitting the simple linear regression:
- R-square > 0.50 (why?)
- Regression ANOVA p-value <0.05 (why?)
- Regression coefficients i.e. a and b p-values < 0.05.

Let’s fit the model and get the summary:

```
lm1 <- lm(mtcars$mpg ~ mtcars$wt)
lm1
```

The outputs shows the “minimum” results for the model

R gives the “minimalist” output!

```
Call:
lm(formula = mtcars$mpg ~ mtcars$wt)

Coefficients:
        (Intercept)    mtcars$wt
        37.285         -5.344
```

Let's ask R to provide summary of lm1:

- **summary(lm1)**

- The coefficient of determination (R-square) = 0.7528, which means the independent variable (wt) is able to explain around 75.28% of variance (variability) in the dependent variable (mpg)

The regression ANOVA, hypothesis:

- H0: Intercept only model (y = a) is better
- H1: Intercept only model is significantly reduced than the full model (y=a +bx)

- Regression ANOVA (given by F-Test) p-value <0.05, we accept H1.

- It confirms that intercept only model is significantly reduced than the full model!

Residuals:

- Min 1Q Median 3Q Max
- -4.5432 -2.3647 -0.1252 1.4096 6.8727

Coefficients:

- Estimate Std. Error t value Pr(>|t|)
- (Intercept) 37.2851 1.8776 19.858 < 2e-16 ***
- mtcars$wt -5.3445 0.5591 -9.559 1.29e-10 ***
- ---
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom

Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446

F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10

What is the “residual standard error”?

The **residual standard error, s, (standard error of estimate, SEE),** for *n* sample data points is calculated from the residuals ($y_i - \hat{y}_i$):

$$\boxed{s = \sqrt{\frac{\sum residual^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}}$$

s is an unbiased estimate of the regression standard deviation ***σ***.

- Why is this important?
- It is used to test whether a and b are equal to zero or not.
- Hypothesis test of regression constant:

  $H_0: \alpha=0$, $H_1: \alpha \neq 0$
- Hypothesis testing of regression coefficient:

  $H_0: \beta=0$, $H_1: \beta \neq 0$

Testing a and b in simple linear regression: The "lm" function of R does it for us!

Done with T-test for a:

- Hypothesis: H_0:α=0, H_1:α≠0
- t_a = a/SE(a)
  - Where,

$$SE_a = SEE * \sqrt{\frac{1}{n} + \frac{\overline{(x)}^2}{\sum\left(x-\overline{x}\right)^2}}$$

Done with T-test for b:

- Hypothesis: H_0:β=0, H_1:β≠0
- t_b = b/SE(b)
- Where,

$$SE_b = \frac{SEE}{\sqrt{\sum (x-\overline{x})^2}}$$

Let’s interpret the model coefficients now:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.2851	1.8776	19.858	< 2e-16 ***
mtcars$wt	-5.3445	0.5591	-9.559	1.29e-10 ***

- ---
- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Since this is a BLUE estimate, we can say that: One unit increase in weight of the car decreases the miles per gallon by 5.3445 unit!

The average mileage is 37.2851 miles per gallon!

Residual standard error: 3.046 on 30 degrees of freedom (lower is better!)

Multiple R-squared: 0.7528 (Higher is better!), Adjusted R-squared: 0.7446

F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10 (must be significant to use the coefficients)

Are these results valid?

- **No, not yet!**

Objects saved in the lm1 model can be seen with:

```
names(lm1)
```

- You need to do the “residual” analysis or the LINE tests:

- L = Linearity of residuals
- I = Independence of residuals
- N = Normality of residuals
- E = Equal variance of residuals

You can save the residuals of the model:

```
lm1.resid <- lm1$residuals
```

You can save the fitted value of the model:

```
lm1.fitted <- lm$fitted.values
```

OR use them directly!

Linearity of residuals: Do it!

- Graphical (suggestive):
  - LOESS scatterplot of residuals (y-axis) and predicted values (x-axis)
  - If the LOESS line lies in the zero line of the y-axis then residuals are linear

```
plot(lm1, which=1, col=c("blue"))
```

- Calculation (confirmative):
  - Calculate mean of the residuals
  - If the mean of the residuals is zero then the residuals are linear

```
summary(lm1$residuals)
```

Independence of residuals: Do it!

- Graphical (suggestive):
  - Get Autocorrelation Function Plot (ACF) of the residuals
  - If the plot show is “decreasing” or “increasing” bars then autocorrelation is present
  - If the plot shows “ups” and “down” bars on x-axis then no autocorrelation

```
acf(lm1$residuals)
```

- Calculation (Confirmative):
  - Calculate Durbin-Watson test of residuals
  - If the p-value > 0.05, no autocorrelation
  - If the p-value <= 0.05, autocorrelation present

```
library(car)
durbinWatsonTest(lm1)
```

Normality of residuals: Do it!

- Graphical (Suggestive):
  - Histogram/Normal Q-Q plot
  - If histogram is bell-shaped or values line in the diagonal like of the Q-Q plot then residuals are normally distributed

```
plot(lm1, which=2, col=c("blue"))
```

- Calculation (Confirmative):
  - Get Shapiro-Wilk test or Kolmogorov-Smirnov test of residuals
  - If the p-value > 0.05, residuals follow the normal distribution
  - If the p-value <= 0.05, residuals do not follow the normal distribution

```
shapiro.test(lm1$residuals)
```

Equal variance (homoscedasticity) of residuals: most important residual assumption, DO IT!

- Graphical (Suggestive):
  - Scatterplot of standardize residuals (y-axis) and standardized predicted values (x-axis)
  - If the values are distributed randomly in the plot then homoscedasticity
  - If the values shows some pattern then heteroscedasticity (unequal variances)

```
plot(lm1, which=3, col=c("blue"))
```

- Calculation (Confirmative):
  - Get the Breusch-Pagan test of residuals
  - If the p-value > 0.05, residual variances are equal (homoscedasticity)
  - If the p-value <= 0.05, residual variances are not equal (heteroscedasticity)

```
library(lmtest)
bptest(lm1)
```

If LINE is valid after BLUE then we can predict:

(More here: http://www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-intervals/)

- We need to save independent variable value/values in a new data

```
p <-  as.data.frame(6)
colnames(p) <- "wt”
```

- We can then use this data to predict dependent variable based on the fitted model

```
predict(lm1, newdata = p)
```

- 5.218297 (Cars with 6000 lbs weight will give 5.22 miles per gallon!)

Outliers, Leverage points and Influential observations in Linear Model

- Why Outliers, Leverage points and Influential observations are important in the linear regression validation?

- **Self-learning (Use the link given below to start exploring):**

- https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html

Machine Learning (ML) and Linear Regression: Next class

- Split the data into Train and Test data
- Fit the linear model in the Train data
- Predict the Test data using the Fitted model
- *Linear regression*, a staple of classical statistical modeling, is one of the simplest algorithms for doing supervised learning: https://bradleyboehmke.github.io/HOML/linear-regression.html

Linear Regression Algorithms for ML: https://bradleyboehmke.github.io/HOML/linear-regression.html

- Simple Linear Regression
- Multiple Linear Regression
- Assessing Model Accuracy
- Model Concerns
- Polynomial Regression
- Principal Component Regression
- Partial Least Squares
- Regularized Regression etc.

Question/queries?

Thank you!

@shitalbhandary