Statistical Computing with R: Masters in Data Sciences 503 (S21) Fourth Batch, SMS, TU, 2025

Shital Bhandary

Associate Professor

Statistics/Bio-statistics, Demography and Public Health Informatics

Patan Academy of Health Sciences, Lalitpur, Nepal

Faculty, Masters in Medical Research, NHRC/Kathmandu University

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

Review Preview

- Measure of linear relationship
  - Covariance
  - Correlation

- Simple linear regression

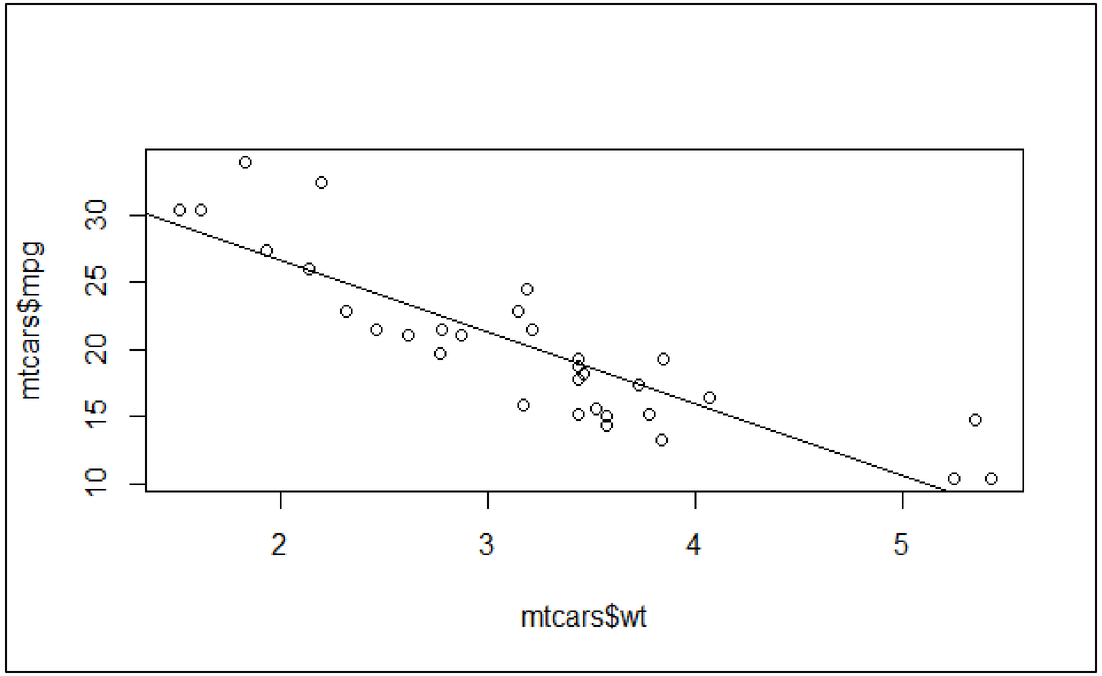
- Cause-effect?
- Prediction?

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)



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

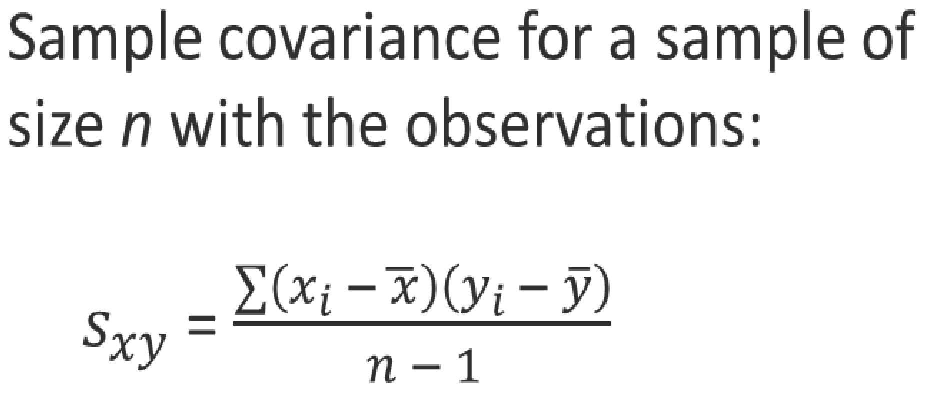
Covariance between WT and MPG variables:

cov(mtcars$wt, mtcars$mpg)

- **-5.116685**

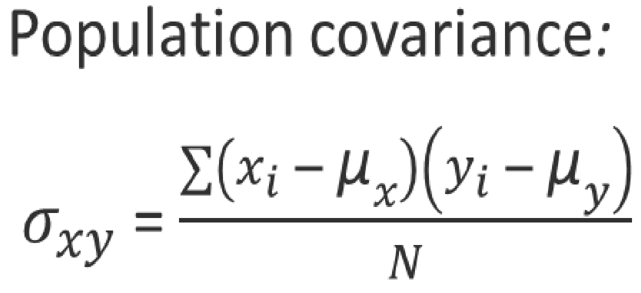
Do as follows now:

- Convert the weight (wt) variable measured in pound to kilogram and store it a new variable **wt2**
- 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:

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

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 (two parts, always!):

- 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 and t-test/1-way ANOVA

- **Thus, it comes with many assumptions!**

Simple Linear Regression:

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

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

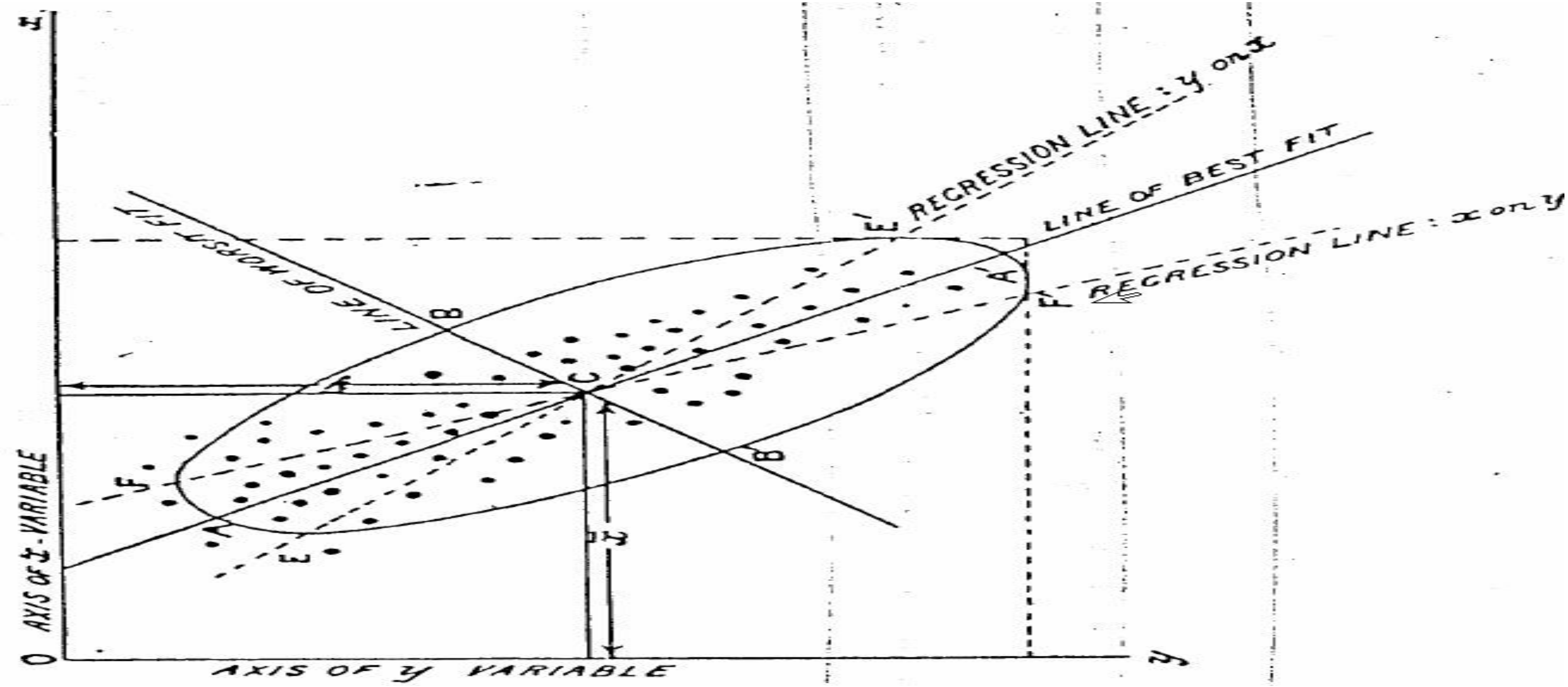
where α and β are parameters called y-intercept and slope respectively, and u is called error or disturbance term, which is erratic or random in nature.

- 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:

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

- 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 for minimizing error - OLS
(OLS = Ordinary Least Square Method)



Simple Linear Regression Assumptions:

- Dependent variable: Normal
- Dependent and Independent (continuous) 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 Best Linear Unbiased Estimation (**BLUE**)!

- To do predictions, Regression Model Residuals (Errors) i.e. "y – yhat" must also be:
  - **L**inear - Linearity of residuals
  - **I**ndependent - Independence of residuals (for time series)
  - **N**ormal - Normality of residuals
  - **E**qual variance - Homoscedasticity of residuals
- Also known as Residual **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?) (Variance explained by ID of DV!)
- Regression ANOVA p-value <0.05 (why?) (Test of Linearity!)
- Regression intercept (a) and coefficients (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

Question/queries so far?

Review Preview

- Simple Linear regression model fit, interpretation
  - Standard Error of Estimate (SEE) or Residual Standard Error (RSE as R calls it)
- Simple Linear Regression Residual Analysis
  - L = Linearity
  - I = Independence
  - N = Normality
  - E = Equal Variance
- Prediction with simple linear regression

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:α=0, H_1:α≠0
- Hypothesis testing of regression coefficient:

  H_0:β=0, H_1:β≠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)

Note: Use of RSE / SEE for Data Science?

- RSE or SEE is used HEAVILY in data science to compare the model accuracy of linear regression algorithm using different methods

- It is one of the “accuracy indices” for assessing linear model fit!

- For example if linear regression is fitted with Ordinary Least Square (OLS), like now, and then using Gradient Descent (GD) methods later then the model with less RSE or SEE will be chosen as the better model for Data Science projects

- **We will discuss this in detail in the next class!!**

Are these (BLUE) 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

<u>OR use them directly!</u>

Linearity of residuals: Do it!

- Graphical (suggestive):
  - Plot scatterplot of residuals (y-axis) and fitted values (x-axis)
  - 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)
```

LINE: Cross-sectional vs Time Series data

- For cross-sectional data, independence of residuals is not mandatory so valid LNE will do

- For time-series data i.e. dependent variable is time, independence of residuals is mandatory so valid LINE is a must

- The **E is most important assumptions of LINE test for both cross-sectional and time series data**, it is it not valid then the BLUE will also be not valid so be careful with this assumption!

If LINE is valid after BLUE then we can predict:

(More here: https://www.statology.org/r-lm-predict/

- 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: 3 more assumptions!

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

```
plot(lm1, which=4, col=c("blue"))
```

- **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
- Polynomial Regression
- Log-transformed Regression
- Assessing Model Accuracy
- Regularized Regression
- Model Concerns

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

- Non-linear Regression e.g. KNN, SVM, DT, RF etc.
- **Principal Component Regression (not covered here)**
- **Partial Least Squares (not covered here)**
- Assessing Model Accuracy
- Model Concerns

You must install “caret” package for next and subsequent classes:

- install.packages(“caret”)
- library(caret)

- Read more about the “caret” package here:

- https://www.stat.colostate.edu/~jah/talks_public_html/isec2020/caret_package.html
- https://topepo.github.io/caret/
- https://cran.r-project.org/web/packages/caret/vignettes/caret.html

Question/queries?

Thank you!

@shitalbhandary