Statistical Computing with R: Masters in Data Sciences 503 (S21) Third Batch, SMS, TU, 2024

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

- 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 $(y_i - \hat{y}_i)$:

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

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,

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

Done with T-test for b:

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

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

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

OR use them directly!

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, 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
- Regularized Regression etc.
- Assessing Model Accuracy
- Model Concerns

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

- Non-linear Regression
- Principal Component Regression
- Partial Least Squares
- 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