Assignment2 2017159092

Mkhulekeli Nkosi

2025-02-28

## Read data

d <- openxlsx::read.xlsx('BayesAssignment2of2025.xlsx', 'Data')  
names(d)

## [1] "y" "x1" "x2" "x3"

1. They should have not because it will now be hard to work with the data cause we won’t know which technique for random sampling was used and how they dealt with missing values.(ATLAS.ti. (n.d.). Random sampling: Definition, methods, & examples. Retrieved February 28, 2025, from <https://atlasti.com/research-hub/random-sampling>)

2.Date and time are not annoying factors removing them will make the data useless since it is time series data.And will help us to know other reasons for that particular sale in that day other than the given variables.(Date and Time may be critical factors in retail sales. Let’s Data Science. (2023). Date and Time Features. Retrieved from <https://letsdatascience.com/date-and-time-features/> ).

3.Email is not the only way to share a data set one can use online services, git large file storage and cloud data base management.Making the data small information is lost and this might lead to invalid inferences due to loss of statistical power and makes it hard for data to met some assumptions (Understanding normality tests: A guide to Shapiro-Wilk and other methods. The Southwest Respiratory and Critical Care Chronicles, 9(40), 45–50. Retrieved from <https://pulmonarychronicles.com/index.php/pulmonarychronicles/article/view/1251/2689>.).

4.we should have also send them a list of things they should not do to data to avoid loss of information and also asking for raw data so we can clean data ourselves.

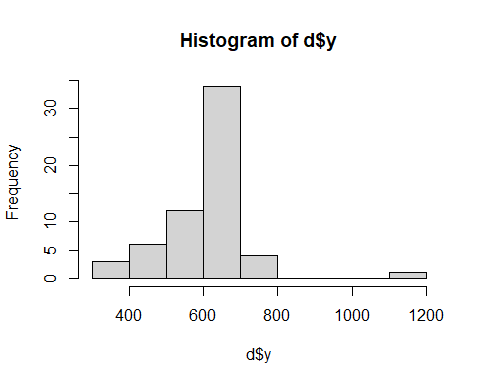
## clean data

1. A simple summary of the data is a good tool to use to check for issues in the data and it gives the descriptive statistics and helps check for missing values, inconsistencies, duplicates and outliers. Even visualization of the data can help check for issues and the shape of the data using histograms and density plots (R Core Team. (2023).)

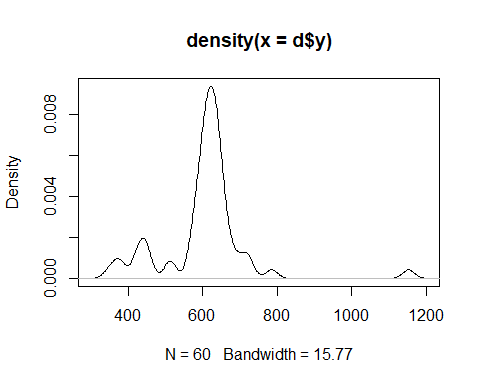
summary(d)

## y x1 x2 x3   
## Min. : 351.0 Min. : 3.900 Min. :10.00 Min. :0.00   
## 1st Qu.: 588.0 1st Qu.: 6.500 1st Qu.:13.00 1st Qu.:0.00   
## Median : 617.5 Median : 7.750 Median :15.00 Median :0.00   
## Mean : 604.0 Mean : 7.790 Mean :14.95 Mean :0.15   
## 3rd Qu.: 641.2 3rd Qu.: 9.225 3rd Qu.:17.00 3rd Qu.:0.00   
## Max. :1152.0 Max. :11.800 Max. :20.00 Max. :1.00

hist(d$y)



plot(density(d$y))



1. One can use more accurate methods to find outliers like the boxplot diagram, scatter plots or statistical tests which are Genralised Extreme Studentized Deviate Test (NIST/SEMATECH e-Handbook of Statistical Methods).

d$y[d$y > (mean(d$y) + 3\*sd(d$y))]

## [1] 1152

library(EnvStats)

##   
## Attaching package: 'EnvStats'

## The following objects are masked from 'package:stats':  
##   
## predict, predict.lm

## The following object is masked from 'package:base':  
##   
## print.default

rosnerTest(d$y, k = 5)

## $distribution  
## [1] "Normal"  
##   
## $statistic  
## R.1 R.2 R.3 R.4 R.5   
## 4.805688 2.731008 2.679942 2.818974 2.532107   
##   
## $sample.size  
## [1] 60  
##   
## $parameters  
## k   
## 5   
##   
## $alpha  
## [1] 0.05  
##   
## $crit.value  
## lambda.1 lambda.2 lambda.3 lambda.4 lambda.5   
## 3.199662 3.193214 3.186628 3.179900 3.173022   
##   
## $n.outliers  
## [1] 1  
##   
## $alternative  
## [1] "Up to 5 observations are not\n from the same Distribution."  
##   
## $method  
## [1] "Rosner's Test for Outliers"  
##   
## $data  
## [1] 577 595 718 639 635 427 643 669 632 709 589 374 617 612 648  
## [16] 1152 595 513 637 620 594 439 617 640 609 452 645 785 677 617  
## [31] 619 642 513 613 614 690 620 443 618 641 653 588 573 664 647  
## [46] 569 605 619 420 452 640 351 635 605 380 727 618 622 594 588  
##   
## $data.name  
## [1] "d$y"  
##   
## $bad.obs  
## [1] 0  
##   
## $all.stats  
## i Mean.i SD.i Value Obs.Num R.i+1 lambda.i+1 Outlier  
## 1 0 603.9833 114.03501 1152 16 4.805688 3.199662 TRUE  
## 2 1 594.6949 89.23260 351 52 2.731008 3.193214 FALSE  
## 3 2 598.8966 83.91844 374 12 2.679942 3.186628 FALSE  
## 4 3 602.8421 79.05080 380 55 2.818974 3.179900 FALSE  
## 5 4 606.8214 73.78103 420 49 2.532107 3.173022 FALSE  
##   
## attr(,"class")  
## [1] "gofOutlier"

7.Removing the outlier would be not a good idea since we working with profit data which depends on sales thus one must investigate first the reason for the outliers (Adams et al., 2019).

d <- d |> subset(d$y <= (mean(d$y) + 3\*sd(d$y)))

# Check Assumptions.

1. Normality is not required for the dependent variable in linear regression, but it is for the residuals. The estimation method used in linear regression, ordinary least squares (OLS) method, doesn’t not require the normality assumption,( Leech, N. L., Barrett, K. C., & Morgan, G. A. (2014)).

shapiro.test(d$y)

##   
## Shapiro-Wilk normality test  
##   
## data: d$y  
## W = 0.8677, p-value = 1.224e-05

1. linear regression can still be performed, but usually ANOVA will require data points to be transformed using log transformation or Box-Cox if the data does not follow normality after transformation we can still perform ANOVA.(Osborne, J. W. & Waters, E., (2002)).
2. The assumptions of ANCOVA like those of regression assume normality of residuals thus the alternative would be to fit a non - parametric regression.(Tibshirani, R.(2013))

[11 - 12]. ANCOVA assumptions are being checked and the F - test checks for homogeneity of variances in the two groups given the results y.( Leech, N. L., Barrett, K. C., & Morgan, G. A. (2014). SPSS for Intermediate Statistics: Use and Interpretation (5th ed.). Routledge)

var.test(y ~ x3, data = d)

##   
## F test to compare two variances  
##   
## data: y by x3  
## F = 0.57846, num df = 49, denom df = 8, p-value = 0.2307  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.1518506 1.4251412  
## sample estimates:  
## ratio of variances   
## 0.5784644

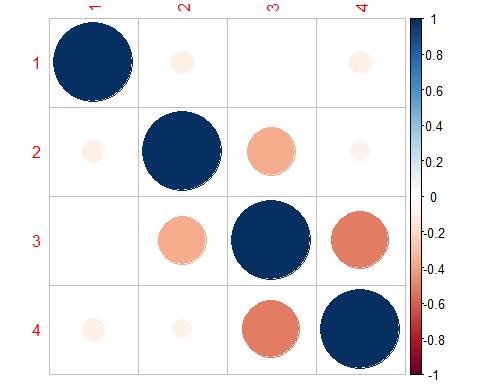
1. the p - value is not less than 0.05 which is the default alpha for the var.test in R thus we fail to reject the null hypothesis(<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/var.test>). this means that an equal variance t-test
2. In this case an equal variance t test should be used since we failed to reject HO of the f - test but it will be advisable to use another test since assumptions of normality are not met given the above Shapiro Wilk test (Kim, T. K., & Park, J. H. (2019)).
3. Yes we can but because we failed to reject null hypothesis is given the shapiro wilk test we cannot use it and also the population standard deviations are unknown(Curtis, A. E., Smith, T. A., Ziganshin, B. A., & Elefteriades, J. A. (2016).).
4. No we cannot take x3 as the main effect since the t - test only tests for equality of means not for the main effect.(Kim, T. K., & Park, J. H. (2019))

17.Its good to check for Auto correlation before performing ANCOVA because we working with time series data and this will check if the assumption of independence of the of the residuals holds.( ChatGPT. (2025, March 1).).

18.This is the correct but it should be done on the dependent variable and best way to calculate autocorrelation is to use Durbin - Waston test.

[19 - 20 ] No we need to calculate autocorrelation of the residuals to know if they are correlated the significance correlation of the dependent variable does not assume significant correlation of the residuals. ( ChatGPT. (2025, March 1)).

n\_obs <- nrow(d)   
n\_segments <- 4   
segment\_size <- ceiling(n\_obs/n\_segments)   
segmented\_matrix <- sapply(1:n\_segments, \(i) {   
 segment <- ((i-1)\*segment\_size + 1):min((i\*segment\_size), n\_obs)   
 c(d$x1[segment], rep(NA\_real\_, segment\_size - length(segment)))   
})   
segmented\_matrix |> cor(use = 'pairwise.complete.obs') |>   
 corrplot::corrplot()



21.Yes a mixed fixed effects model will be fitted if the residuals are know to be correlated but the correlation plot shows that there is weak correlation thus mixed fixed models are not fitted since autocorrelation is not present in x1.

1. That is true which means ANCOVA can still be fitted since no autocorrelation is observed meaning assumption of independence of residual still holds for the data.(Field, 2013).

23.ANCOVA can still be applied even if the sample is imbalanced. (Laerd Statistics, 2023).

aov(y ~ x1 + x2 + x3, data = d) |> summary()

## Df Sum Sq Mean Sq F value Pr(>F)   
## x1 1 46217 46217 15.594 0.000225 \*\*\*  
## x2 1 51 51 0.017 0.895807   
## x3 1 252552 252552 85.216 9.04e-13 \*\*\*  
## Residuals 55 163002 2964   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. This is true because it does not affect the dependent variable with p - value greater than 0.05.

aov(y ~ x1 + x3, data = d) |> summary()

## Df Sum Sq Mean Sq F value Pr(>F)   
## x1 1 46217 46217 15.74 0.000208 \*\*\*  
## x3 1 251232 251232 85.59 7.09e-13 \*\*\*  
## Residuals 56 164374 2935   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coef(aov(y ~ x1 + x3, data = d))

## (Intercept) x1 x3   
## 707.21961 -10.81188 -183.61300

1. One can choose to keep the variable.
2. stepwise is often associated with over fitting thus we need to worry about over fitting.

* as the model that will be used for prediction

1. No we need to choose values careful specially for $ x\_{3}$ it should only be 1 or 0 values for it can be any value as-long as it is realistic no one can have 100 years of experience.

30.We need to introduce confident intervals for our predicted values then we can be 95% confident that the predicted value lies in that interval.

# References.

1.(ATLAS.ti. (n.d.). Random sampling: Definition, methods, & examples. Retrieved February 28, 2025, from <https://atlasti.com/research-hub/random-sampling>)

1. Date and Time Features. Retrieved from <https://letsdatascience.com/date-and-time-features/>
2. The Southwest Respiratory and Critical Care Chronicles, 9(40), 45–50. Retrieved from <https://pulmonarychronicles.com/index.php/pulmonarychronicles/article/view/1251/2689>.
3. Investopedia. (2023). Normal Distribution. Retrieved from <https://www.investopedia.com/terms/n/normaldistribution.asp>

5.Linear Regression: Should Dependent and Independent Variables Be Distributed Normally? Retrieved from <https://www.statsimprove.com/en/linear-regression-should-dependent-and-independent-variables-be-distributed-normally/>

6.Analysis of Covariance (ANCOVA). Retrieved from <https://www.statisticssolutions.com/analysis-of-covariance-ancova/> ANCOVA might be a reasonable choice if x3 is a binary categorical variable.

8.Leech, N. L., Barrett, K. C., & Morgan, G. A. (2014). SPSS for Intermediate Statistics: Use and Interpretation (5th ed.). Routledge.

1. Built In. (n.d.). Z-Test: What It Is & How to Perform It. Retrieved from <https://builtin.com/data-science/z-test-statistics>. They used t-test instead of z-test like initially stated in the statement.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
3. Tibshirani, R., & Wasserman, L. (2013). Nonparametric regression. Statistical Machine Learning, Spring.

12.M. A. (2004). Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models.

1. R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org>. NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/>, April 2012.
2. Adams, J., Hayunga, D., Mansi, S., Reeb, D., & Verardi, V. (2019). Identifying and treating outliers in finance. Financial Management, 48(2), 345-384.
3. Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. Practical assessment, research, and evaluation, 8(1).
4. Tibshirani, R., & Wasserman, L. (2013). Nonparametric regression. Statistical Machine Learning, Spring.
5. Kim, T. K., & Park, J. H. (2019). More about the basic assumptions of t-test: normality and sample size. Korean journal of anesthesiology, 72(4), 331-335. 18.Curtis, A. E., Smith, T. A., Ziganshin, B. A., & Elefteriades, J. A. (2016). The mystery of the Z-score. Aorta, 4(04), 124-130.
6. ( ChatGPT. (2025, March 1). Why check for autocorrelation before performing ANCOVA. OpenAI. <https://chat.openai.com>).