Biostatistics 1 Project

Kimberly Schveder

Professor: Dr. Datta

Spring 2019

Table of contents:

The Cleveland Clinic Datasets, Data Dictionary images, and data set introductions in this report come from the Cleveland Clinic Lerner Research Institute's public data bank:

http://www.lerner.ccf.org/qhs/datasets/datasets.php

1. Analysis on Proportions: Liver Fibrosis

With this Liver Steatosis data, provided by the Cleveland Clinic Lerner Research Institute, the original study had the goal:

> to determine the prevalence of steatosis and fibrosis in morbidly obese subjects, as determined by liver biopsies—which are considered the "gold standard" diagnostic tool for this condition…. Our second goal was to determine the accuracy of ultrasonic diagnosis of liver steatosis in morbidly obese patients, considering clinical characteristics might also influence the diagnostic value of ultrasound for steatosis and help the clinician interpret ultrasound results…..The study enrolled patients who had laparoscopic gastric bypass, sleeve, or band surgery. We included patients who had no clinical evidence of other liver diseases and who had intraoperative needle liver biopsy with or without preoperative right upper quadrant ultrasound between 2005 and 2009. We recorded clinical characteristics including diabetes, plasma triglycerides, cholesterol, very-low-density lipoprotein, low-density lipoprotein, aspartate aminotransferase, alanine aminotransferase, NAS score, BMI, metabolic syndrome and duration of obesity.

So, I did the types of analyses that the authors did in this article, all using the methods discussed and described in class. My results matched up quite well with the results from the official study done at the Cleveland Clinic. I will be looking again at this data set in the ordinal logistic regression modeling technique below, to look at an alternative way to answer questions about what this dataset is telling us about these patients.

```
> n = length(positiveLSBiopsychar)    # valid responses count
> n
[1] 443
> k = sum(positiveLSBiopsychar == "1")
> k
[1] 311
> pbar = k/n; pbar
[1] 0.7020316
> SE = sqrt(pbar*(1-pbar)/n); SE     # standard error
[1] 0.02173009
> E = qnorm(.975)*SE; E              # margin of error
[1] 0.0425902
> pbar + c(-E, E)
[1] 0.6594414 0.7446218
> prop.test(k, n, p = 0.70,
+           alternative = c("two.sided"),
+           conf.level = 0.95, correct = TRUE)

        1-sample proportions test with continuity correction

data:  k out of n, null probability 0.7
X-squared = 0.0017199, df = 1, p-value = 0.9669
alternative hypothesis: true p is not equal to 0.7
95 percent confidence interval:
 0.6566879 0.7438101
sample estimates:
        p
0.7020316
```

Among 435 patients with conclusive biopsy results, estimated prevalence of liver steatosis was 70.2% (95% confidence interval 66%, 74.5%), as seen above. Just about three quarters of bariatric surgery patients have liver steatosis. As confirmed by the proportions test, which has the null hypothesis that the true p is equal to 0.70 vs. the alternative hypothesis that the true p (= prevalence of liver steatosis) is not equal to 0.70, we do not reject the null hypothesis in this case because the p-value = 0.9669. It is probably safe to assume that the p = 0.70, or just about. Which, because the sample size is large enough to the CLT to apply, and we know that the true population proportion of patients who had laparoscopic gastric bypass, sleeve, or band surgery with liver steatosis is 0.70.

Next, we see that the estimated prevalence of fibrosis was 26.2% (22.1%, 30.2%), as seen in the one-sample proportions test here. About a quarter of the patients have fibrosis. As confirmed by the proportions test, which has the null hypothesis that the true p is equal to 0.25 vs. the alternative hypothesis that the true p (= prevalence of liver steatosis) is not equal to 0.25, we do not reject the null hypothesis in this case because the p-value = 0.6022. It is probably safe to assume that the p = 0.25, or just about. Which, because

```
> n1 = length(fibrosischar)    # valid responses count
> n1
[1] 443
> k1 = n1-sum(fibrosischar== "0") #number of people with some kind of fibrosis in t$
> k1
[1] 116
> pbar1 = k1/n1; pbar1
[1] 0.261851
> SE1 = sqrt(pbar1*(1-pbar1)/n1); SE1     #standard error
[1] 0.02088802
> E1 = qnorm(.975)*SE1; E1                #margin of error
[1] 0.04093977
> pbar1 + c(-E1, E1)                      #95% CI
[1] 0.2209112 0.3027908
> prop.test(k1, n1, p = 0.25,
+           alternative = c("two.sided"),
+           conf.level = 0.95, correct = TRUE)

        1-sample proportions test with continuity correction

data:  k1 out of n1, null probability 0.25
X-squared = 0.27163, df = 1, p-value = 0.6022
alternative hypothesis: true p is not equal to 0.25
95 percent confidence interval:
 0.2220209 0.3058881
sample estimates:
       p
0.261851
```

the sample size is large enough, and we know that the true population proportion of patients who had laparoscopic gastric bypass, sleeve, or band surgery with fibrosis is 0.25.

```
> table(LiverSteatosisdata$LS..Biopsy, LiverSteatosisdata$LS..US)

       0 0.5   1
  0    81   1  38
  0.5   1   1   6
  1    40   2 251
```

```
> table(LiverSteatosisdata$LS..Biopsy, LiverSteatosisdata$LS..US)

       0 0.5   1
  0    81   1  38
  0.5   1   1   6
  1    40   2 251
```

My results continued to match up with the article's results.

The 0.5 columns in the table hereof the golden predictor of LS Biopsy vs. LS ultra sound stand for the inconclusive test results. Notice how few inconclusive test results there are. So, in the analysis, in order to do the calculations, I ignored these observations because there are so few, compared to the rest.

```
> LiverSteatosisdata_1 = subset(LiverSteatosisdata, LS..Biopsy!=0.5 & LS..US!=0.5)
> tbl2 = table(LiverSteatosisdata_1$LS..Biopsy, LiverSteatosisdata_1$LS..US)
> tbl2

     0   1
  0 81  38
  1 40 251
> chisq.test(tbl2)

        Pearson's Chi-squared test with Yates'
        continuity correction

data:  tbl2
X-squared = 117.21, df = 1, p-value < 2.2e-16
```

Subsetting the data to use to make a 2x2 contingency table of the 0s and 1s of the two variables ls. Biopsy and ls.us, a chi-square test of independence could be performed, and sensitivity, specificity, and measures of association can be easily calculated.

The chi-square test of independence test the null hypothesis that the two variables being compared are independent from each other vs. the alternative that they are not dependent on other. Note that because all of the observed counts in the cells o the 2x2 contingency table are larger than 5, we can use the powerful chi-square test of independence instead of the nonparametric, less powerful alternative test of fisher's exact test of independence. Here, the p-value is nearly zero, so we reject the null hypothesis at any reasonable significant level in favor of the alternative, and hence conclude that there is some dependence or association of LS Biopsy and LS by Ultra sound. This might tell us that Biospy is a good predictor, to spme extent, of liver steatosis. Now, let's look at sensitivity, specificity, and some measures of association.

Sensitivity of ultrasound for liver steatosis was (true positive/(true positive + false negative)) = (251/(251+38)) = 86.25% ; its specificity was = (true negative/(true negative + false positive)) = (81/(81+40)) = 67%. This provides evidence that, in general, ultrasound was moderately predictive of liver fibrosis, when compared to the golden standard of prediction.

```
> a = tbl2[1,1] #81 of the patients who do not have ls by biospy or ls by us
> a
[1] 81
> b = tbl2[1,2] #38 of the patients who do not have ls by biospy but have ls by us
> b
[1] 38
> c = tbl2[2,1] #40 of the patients who do have ls by biospy but do not have ls by us
> c
[1] 40
> d = tbl2[2,2] #251 of the patients who do have ls by biospy and have ls by us
> d
[1] 251
>
> oddsratio=(a*c)/(b*c)
> oddsratio
[1] 2.131579
```

Here are the calculations in R of the odds ratio, a measure of association as discussed in class. We can use this as a reliable measure because the sample size is large. If the sample size was small, then using the log odds ratio would and analyzing it would be the better option. The odds ratio is 2.13. According to a helpful resource (http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/presentations/Lobanova-2008-OddsRatio.pdf), because the odds ratio is larger than 1, the event of someone being diagnosed with liver steatosis is more likely with ultra sound than with biopsy, as expected.

The dataset represents data from the study by Wu et al. "Prevalence of Liver Steatosis and Fibrosis and the Diagnostic Accuracy of Ultrasound in Bariatric Surgery Patients". ObesSurg 2012; 22: 240-247.

R code:

```
###proportions analysis###

LiverSteatosisdata = read.csv(file="LiverSteatosis.csv", header=TRUE, sep=",")

ls(LiverSteatosisdata)

#converting numeric to character

diabetes <- LiverSteatosisdata$DM

diabeteschar <- as.character(diabetes)

metabolicsyndrome <- LiverSteatosisdata$MET..Syndrome

metabolicsyndromechar <- as.character(metabolicsyndrome)

hypertension <- LiverSteatosisdata$HTN

hypertensionchar <- as.character(hypertension)

hyperlipidemia <- LiverSteatosisdata$HPL

hyperlipidemiachar <- as.character(hyperlipidemia)

fibrosis <- LiverSteatosisdata$Fibrosis

fibrosischar <- as.character(fibrosis)

positiveLSUS <- LiverSteatosisdata$LS..US

positiveLSUSchar <- as.character(positiveLSUS)

positiveLSBiopsy <- LiverSteatosisdata$LS..Biopsy

positiveLSBiopsychar <- as.character(positiveLSBiopsy)

#help from #http://www.r-tutor.com/elementary-statistics/interval-estimation/interval-
estimate-population-proportion

n = length(positiveLSBiopsychar)    # valid responses count

n

k = sum(positiveLSBiopsychar == "1")

k

pbar = k/n; pbar

SE = sqrt(pbar*(1-pbar)/n); SE     # standard error

E = qnorm(.975)*SE; E              # margin of error

pbar + c(-E, E)

prop.test(k, n, p = 0.70,

          alternative = c("two.sided"),

          conf.level = 0.95, correct = TRUE)

n1 = length(fibrosischar)    # valid responses count

n1
```

```
k1 = n1-sum(fibrosischar== "0") #number of people with some kind of fibrosis in their
liver

k1

pbar1 = k1/n1; pbar1

SE1 = sqrt(pbar1*(1-pbar1)/n1); SE1     #standard error

E1 = qnorm(.975)*SE1; E1                #margin of error

pbar1 + c(-E1, E1)                      #95% CI

prop.test(k1, n1, p = 0.25,

          alternative = c("two.sided"),

          conf.level = 0.95, correct = TRUE)

#Ultrasound Liver Fibrosis

positiveLSUSchar = na.omit(positiveLSUSchar)

n2 = length(positiveLSUSchar)    # valid responses count

n2

k2 = sum(positiveLSUSchar== "1")

k2

pbar2 = k2/n2; pbar2

SE2 = sqrt(pbar2*(1-pbar2)/n2); SE2     #standard error

E2 = qnorm(.975)*SE2; E2                #margin of error

pbar2 + c(-E2, E2)                      #95% CI

prop.test(k2, n2)

#table includes inconclusive values, which are left out of the analysis because there
are so few.

tbl1 = table(LiverSteatosisdata$LS..Biopsy, LiverSteatosisdata$LS..US)

tbl1

chisq.test(tbl1)

#####

LiverSteatosisdata_1 = subset(LiverSteatosisdata, LS..Biopsy!=0.5 & LS..US!=0.5)

tbl2 = table(LiverSteatosisdata_1$LS..Biopsy, LiverSteatosisdata_1$LS..US)

tbl2

chisq.test(tbl2)

#http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-
independence

###chi square test of independence

a = tbl2[1,1] #81 of the patients who do not have ls by biospy or ls by us
```

```
b = tbl2[1,2] #38 of the patients who do not have ls by biospy but have ls by us
c = tbl2[2,1] #40 of the patients who do have ls by biospy but do not have ls by us
d = tbl2[2,2] #251 of the patients who do have ls by biospy and have ls by us
oddsratio=(a*c)/(b*c)
oddsratio
logoddsratio=log(oddsratio)
selogoddsratio = sqrt((1/a)+(1/b)+(1/c)+(1/d))
qnorm(0.975) #use the normal distribution because the sample size is very large
#95% CI for logoddsratio
L = logoddsratio - qnorm(0.975)*selogoddsratio
U = logoddsratio + qnorm(0.975)*selogoddsratio
cbind(L,U)
table(LiverSteatosisdata$LS..Biopsy, LiverSteatosisdata$Fibrosis)
```

2. Analysis on Hypothesis testing: Normothermia

This next data set that I am looking at and doing an appropriate analysis on is called Normothermia, from the study by Ruetzler et al, that's called "Forced-air and a novel patient-warming system (vitalHEAT vH2) comparably maintain normothermia during open abdominal surgery". (https://pdfs.semanticscholar.org/6ebe/0ae6e09f9477b7848c89848daafcb764d38b.pdf) This is a prospective study.

The data set background and description given by the Cleveland Clinic is stated below:

"Perioperative hypothermia causes adverse outcomes, including impaired drug metabolism, cardiac morbidity, shivering, impaired immune function, coagulopathy, and increased use of hospital resources. Therefore, maintaining perioperative normothermia significantly reduces morbidity and has become routine. Convective (forced-air) warming is by far the most common intraoperative warming strategy. Forced-air warming is relatively inefficient on a per-surface-area basis, but nonetheless transfers considerable heat to the anterior surface of patients because the warm air contacts a large surface area. One difficulty with forced-air warming, though, is that in patients having large procedures, especially in positions other than supine, it may be impossible to warm sufficient surface area to maintain normothermia, defined as a core temperature of 36.0°C. Recently, the vitalHEAT vH2 system was developed that potentially transfers adequate heat through a single extremity using a combination of conductive heat (circulating warm water within soft fluid pads) with mild vacuum, which enhances contact between the heating element and the skin surface.

Preliminary (uncontrolled and unpublished) studies suggest that the device is effective, even in open abdominal surgery. These investigators therefore tested the hypothesis that intraoperative distal esophageal (core) temperatures are not > 0.5°C lower during elective open abdominal surgery under general anesthesia in patients warmed with the warm-water sleeve on 1 arm than with an upper-body forced-air cover.

LIST OF VARIABLES:

Name	Codes/Values	Abbreviation
Hospital location	1 = CCF; 2 = Vienna	site
Subject ID		subject
Warming method patient randomized to receive	0 = Forced Air; 1 = VitalHEAT	randomization
ASA Physical Status	1 = I; 2 = II; 3 = III	asa
Gender	1 = Male; 2 = Female	gender
Age	years	age
Weight	kg	weight
Height	cm	height
Body Mass Index	kg/m^2	bmi
Race	1 = American Indian / Alaska Native 2 = Asian 3 = Native Hawaiian / Other Pacific Islander 4 = Black / African American 5 = White 6 = More than one race 7 = unknown	race
Ethnicity	1 = Hispanic / Latino 2 = Not Hispanic / Latino 3 = unknown	ethnicity
Arm tuck; left	0 = No; 1 = Yes	armTuckLeft
Arm tuck; right	0 = No; 1 = Yes	armTuckRight
Withdrawn	0 = No; 1 = Yes	withdrawn
Reached ≤ 35° C	0 = No; 1 = Yes	reached_le_35
Preoperative temperature	° C	preop_temp
Temperature at intubation	° C	intubation_temp

The study enrolled patients scheduled for elective major open abdominal surgery (liver, pancreas, and colon–rectal surgery) under general anesthesia scheduled to last at least 2 hours at the Cleveland Clinic Main Campus (Cleveland, Ohio) and at the Vienna General Hospital of the Medical University of Vienna (Vienna, Austria). Patients were randomly assigned to vitalHEAT (circulating-water sleeve) (n = 37) or forced-air warming (n=34). Temperature measurements from 15 minutes after intubation until the end of the case were used for analysis."

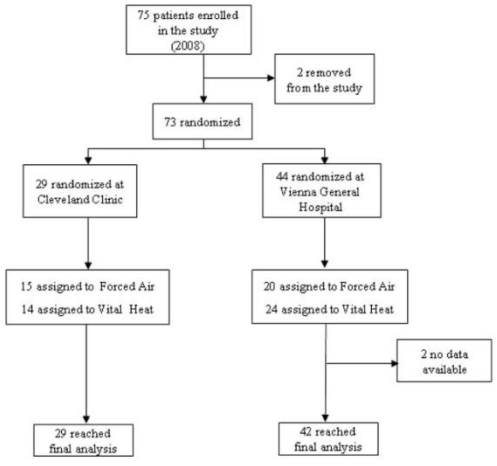


Figure 3. Trial profile.

The paper explains the protocol of the clinical trial and explains some of the difficulties of the experiment.

Here, there was a test of noninferiority of the drug treatment, as was talked about in the prospective randomized clinical trial lecture on bioequivalence testing. So, I replicated what they did by testing if the difference between the variances of the temperature at intubation (degrees Celsius) (completed during surgery) based on the two randomized treatments (VitalHeat and forced Air). We are looking to see if there is inferiority between the two treatments by comparing the means of the two groups.

The two groups are independent, as shown in the diagram above and as explained in the summary from the data set background above. The patients are from two different countries and areas of the world, and the clinical trial was randomized; that is, the patients were randomized to different treatments. The patients themselves are independent of each other. We see that there were enough people to make it to the final analysis (29 and 42) to have a meaningful and representative analysis of the data.

$$\text{H0: } \mu_C - \mu_F \leq -0.5°\text{C}$$

and

$$\text{HA: } \mu_C - \mu_F > -0.5°\text{C},$$

Here, the μ_C is the true population mean for the conventional circulating water sleeve. On the contrary, μ_F is the true population mean for the forced air temperature groups. The null hypothesis is that the "the mean temperature with the circulating-water sleeve is ≥0.5°C lower (worse) than is the mean forced-air temperature. The alternative hypothesis, which we assessed in our test for noninferiority, was that mean temperature in patients assigned to the circulating-water sleeve is at most 0.5°C lower than forced air, and perhaps higher." The noninferiority δ constant was chosen according from a pilot study, as it was -0.5 degrees Celsius in that study, according to the report.

First, in order to do the t-test, we need to see if there is equal variance or not. According to the results from the F test to compare two variances, we cannot conclude that the two variances are different, because the p-value is not significant at any reasonable significance level. We do know that the two populations are independent (participants from two hospitals in very different parts of the world). So, I will assume equal variances in the r-code of the t-test.

```
> vartest=var.test(intubation_temp~randomization,data=Normothermiadata)
> vartest

        F test to compare two variances

data:  intubation_temp by randomization
F = 1.4124, num df = 16, denom df = 20, p-value = 0.4602
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.554654 3.786483
sample estimates:
ratio of variances
          1.412449
```

```
> ttest=t.test(intubation_temp~randomization, data = Normothermiadata, var.equa$
+        mu = -0.5)
> ttest

        Two Sample t-test

data:  intubation_temp by randomization
t = 5.8911, df = 36, p-value = 4.858e-07
alternative hypothesis: true difference in means is greater than -0.5
95 percent confidence interval:
 -0.03238353         Inf
sample estimates:
mean in group 0 mean in group 1
       36.24118        36.08571
```

According to the results from the two-sample t-test, with the alternative hypothesis symbol set to "greater", we reject the null in favor of the alternative that the true mean difference is greater than -0.5 (in a pilot study, accordingly, this alternative hypothesis means that the two means are noninferior); These results line up quite well to the results in the study, that the "the results thus indicate that the circulating-water sleeve is noninferior to forced air during surgery, defined by core temperature being no > 0.5°C lower." Non-inferiority was detected here, and the p-value was less than any reasonable significance level, just like the study's results for the t-test of noninferiority. Because the core temperature is not > 0.5 degrees Celsius lower, the circulating water sleeve is non-inferior to forced air during surgery.

R code:

```
#####################Normothermia dataset#######################

Normothermiadata = read.csv(file="normothermia.csv", header=TRUE, sep=",")

ls(Normothermiadata)

install.packages("devtools") #https://cran.r-project.org/web/packages/webr/README.html

library(devtools)

vartest=var.test(intubation_temp~randomization,data=Normothermiadata)
#https://cran.r-project.org/web/packages/webr/README.html

vartest

install.packages("stats")

ttest=t.test(intubation_temp~randomization, data = Normothermiadata, var.equal=TRUE,
paired = FALSE,  alternative = c("less"),

       mu = -0.5)

ttest

plot(ttest)
```

The dataset represents data from the study by Ruetzler et al. "Forced-air and a novel patient-warming system (vitalHEAT vH2) comparably maintain normothermia during open abdominal surgery". Anesth Analg 2011; 112: 608-14.

3. Analysis on Kruskal-Wallis and post-hoc testing: Blood storage

This data is on blood storage has to do with prostate cancer in men. The data itself comes from the study done at the Cleveland Clinic done by Cata et al. It was called "Blood Storage Duration and Biochemical Recurrence of Cancer After Radical Prostatectomy". Mayo Clin Proc 2011; 86(2): 120-127.

Nobody can explain what the data set consists of and what the study was about than the Cleveland Clinic summary itself, which I have placed here for your reference:

> "Prostate cancer is the most common malignant neoplasm in men, and radical prostatectomy is among the primary therapies for localized prostate cancer. The biochemical recurrence rate 5 years after prostatectomy ranges from 70% to 90%. Improvements in the surgical technique have decreased the amount of intraoperative blood loss occurring during radical prostatectomy; however, substantial numbers of patients still require perioperative blood transfusions.
>
> Blood transfusions are associated with adverse reactions, including postoperative infections and transfusion-related immune perturbations. Allogeneic leukocytes present in the transfused blood are thought to suppress host cellular immune responses. Furthermore, the immunodepressant effect is secondary to an imbalance of accumulated cytokines and proinflammatory mediators in the transfused blood against decreased production of lymphocyte stimulating cell-mediated cytokines, such as interleukin 2 and increased release of immunosuppressive prostaglandins in the patient undergoing transfusion.

In cancer patients, perioperative blood transfusion has long been suspected of reducing long-term survival, but available evidence is inconsistent. It is also unclear which components of transfused blood underlie the cancer-promoting effects reported by some studies. An important factor associated with the deleterious effects of blood transfusion is the storage age of the transfused blood units. It is suspected that cancer recurrence may be worsened after the transfusion of older blood.

This study evaluated the association between red blood cells (RBC) storage duration and biochemical prostate cancer recurrence after radical prostatectomy. Specifically, tested was the hypothesis that perioperative transfusion of allogeneic RBCs stored for a prolonged period is associated with earlier biochemical recurrence of prostate cancer after prostatectomy.

LIST OF VARIABLES:

Name	Codes/Values	Abbreviation
RBC storage duration group	1 = ≤ 13 days (Younger) 2 = 13 - 18 days (Middle) 3 = ≥18 days (Older)	RBC Age Group
Median RBC age of all transfused units within each group	days	Median RBC Age
Patient age	years	Age
African American race	0 = non-African American 1 = African American	AA
Family history of disease	0 = No; 1 = Yes	FamHx
Prostate volume	g	PVol
Tumor volume	1 = Low 2 = Medium 3 = Extensive	TVol
Clinical T category	1 = atage T1-T2a 2 = stage T2b-T3	T Stage
Biopsy Gleason score	1 = score 0 - 6 2 = score 7 3 = score 8 - 10	bGS
Bladder neck positive	0 = No; 1 = Yes	BN+
Organ confined	0 = No; 1 = Yes	OrganConfined
Preoperative prostate specific antigen (PSA)	ng/mL	PreopPSA
Preoperative therapy	0 = No; 1 = Yes	PreopTherapy
Number of allogeneic units		Units
Surgical Gleason score	1 - Not assigned 2 = No residual disease or score 0 - 6 3 = score 7 4 = score 8 - 10	sGS
Any adjuvant therapy	0 = No; 1 = Yes	AnyAdjTherapy
Adjuvant radiation therapy	0 = No; 1 = Yes	AdjRadTherapy
Biochemical recurrence of prostate cancer	0 = No recurrence 1 = Recurrence	Recurrence
Censoring indicator	0 = Not censored 1 = Censored	Censor
Time to biochemical recurrence of prostate cancer	months	TimeToRecurrence

Patients were assigned to 1 of 3 RBC age exposure groups on the basis of the terciles (ie, the 33rd and 66th percentiles) of the overall distribution of RBC storage duration if all their transfused units could be loosely characterized as of “younger,” “middle,” or “older” age. Although this approach resulted in the removal of certain patients with wide RBC age distributions, it has the advantage of defining an essentially random and clearly separable exposure.

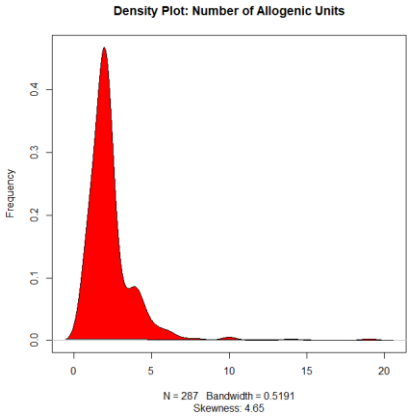
Prostate-specific antigen (PSA) was used as a biochemical marker of prostate cancer recurrence after prostatectomy. A PSA value of at least 0.4 ng/mL (to convert to µg/L, multiply by 1.0) followed by another increase was considered biochemical cancer recurrence.

The initial population consisted of 865 men who had undergone radical prostatectomy and received transfusion during or within 30 days of the surgical procedure at Cleveland Clinic and had available PSA follow-up data. Of these patients, 110 were excluded from the analysis because they received a combination of allogeneic and autologous blood products. Of the remaining 755 patients, 405 (54%) received solely allogeneic and

350 patients (46%) received solely autologous RBC units. Of the 405 patients who received allogeneic RBC transfusion, 89 were excluded because their transfused RBC age distribution included more than one of the terciles. Thus, this dataset consists of the 316 patients who received solely allogeneic blood products and could be classified into an RBC age exposure group.”

So, in summary of the background here, a major part of the study was a test to see if the central measure number of allogenic units of each of the RBC storage groups are all equal to each other, vs. the alternative hypothesis that at least one of the means of allogenic units is different from the others. I tried to replicate this here. Let’s see if an ANOVA test can be completed here. In not all of the assumptions are met, we will do a Krustal-Wallis test.

Before doing any ANOVA procedure, or a similar nonparametric procedure, a test for normality of the number of allogeneic units was conducted. But, just from looking at the histogram of this variable, it appear that there are some outliers that make the data skewed to the right, so the parametric normality test that uses the mean is going to most likely tell us that there is a normality violation. The Shapiro-Wilk normality test does this. Here, the normality test rejects the normality of the data null hypothesis from this test in favor of the alternative hypothesis that the data is not normal, because we have a p-value that is nearly zero.




```
> ##normality tests###
> #normality test#
> shapiro.test(bloodstoragedata$Units)

        Shapiro-Wilk normality test

data:  bloodstoragedata$Units
W = 0.55956, p-value < 2.2e-16
```

So, a nonparametric alternative to ANOVA will be appropriate to telling us if there is a significant difference in the medians of the three groups in RBC storage duration groups on the number of allogenic units. Here, the Krustal Wallis test can be used, because of this violation in the normality assumption of the ANOVA procedure. The ANOVA procedure also requires that the groups being compared have equal variable by the numeric variable. The Kruskal Wallis test does not assume this. It is a nice, although less powerful comparison of the center test.

In addition, the relevant results are those from the Krustal-Wallis rank sum test, according to the published article, my results are correct. At least one of the medians of the Units of the RBC ages groups is different from the rest, because the p-value is significant. (Mine was 0.0097. The study's p-value was 0.02.) That is, we reject the null of the equality of the medians of the three groups from the Krustal Wallis test.

```
> kruskal.test(Units ~ RBC.Age.Group, data = bloodstoragedata)

        Kruskal-Wallis rank sum test

data:  Units by RBC.Age.Group
Kruskal-Wallis chi-squared = 9.2623, df = 2, p-value = 0.009744
```

Now, we may want to know which if the group(s) is different in the median from the others. A nonparametric post-hoc test was completed: the pairwise Wilcoxon rank sum test, which tests each pair of medians versus each other. The null hypothesis is that the two median are equal and the alternative is that the two medians are not equal to each other.

```
> pairwise.wilcox.test(bloodstoragedata$Units, bloodstoragedata$RBC.Age.Group,
+                      p.adjust.method = "BH")

        Pairwise comparisons using Wilcoxon rank sum test

data:  bloodstoragedata$Units and bloodstoragedata$RBC.Age.Group

  1     2
2 0.488 -
3 0.015 0.036

P value adjustment method: BH
```

It appears group 3 (the older group) is different from the groups 1 and 2 because both p-values are significant at the 5% level. Groups 1 and 2 probably have close medians because their p-value from this test is no statistically significant.

R code:

```
bloodstoragedata = read.csv(file="BloodStorage.csv", header=TRUE, sep=",")

ls(bloodstoragedata)

install.packages("MASS")

library(MASS) #in use

bloodstoragedata = na.omit(bloodstoragedata)

RBC.Age.Group <- bloodstoragedata$RBC.Age.Group

RBC.Age.Groupchar <- as.character(RBC.Age.Group)

anovaSupra <- aov(Units ~ RBC.Age.Groupchar, data = bloodstoragedata)

anovaSupra

TukeyHSD(anovaSupra, "RBC.Age.Groupchar")

plot(density(bloodstoragedata$Units), main="Density Plot: Number of Allogenic Units",
ylab="Frequency", sub=paste("Skewness:",
round(e1071::skewness(bloodstoragedata$Units), 2)))

polygon(density(bloodstoragedata$Units), col="red")


#krustall wallis: http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r

#head(bloodstoragedata)

kruskal.test(Units ~ RBC.Age.Group, data = bloodstoragedata)

pairwise.wilcox.test(bloodstoragedata$Units, bloodstoragedata$RBC.Age.Groupchar,

                     p.adjust.method = "BH")

##normality tests###

#normality test#

shapiro.test(bloodstoragedata$Units)


##equality of variance tests##

bartlett.test(Units ~ RBC.Age.Group, data = bloodstoragedata)
```

```
leveneTest(Units ~ RBC.Age.Group, data = bloodstoragedata)

fligner.test(Units ~ RBC.Age.Group, data = bloodstoragedata)
#http://www.sthda.com/english/wiki/compare-multiple-sample-variances-in-r
```

The dataset represents data from the study by Cata et al. “Blood Storage Duration and Biochemical Recurrence of Cancer After Radical Prostatectomy”. Mayo Clin Proc 2011; 86(2): 120-127.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3031436/

4. Regression Diagnostics (Alcohol Metabolism)

Regression diagnostics: What does regression diagnostics mean (i.e. what aspects of a regression model are we trying to diagnose)? For each of them (e.g. violation of the underlying assumptions, leverage and influence of outliers, multicollinearity, etc.), what kind of diagnostic methods are available? Describe them and explain the associated cut-off values. (of the multicollinearity and variance inflation factors (A.K.A.) ratio of the eigenvalues)). Discuss the relationships among them. What are some of the remedial measures when the diagnostic tools detect any problem? Illustrate these diagnostic tools on simulated datasets (just using pictures from sources on from the internet).

What does regression diagnostics mean (i.e. what aspects of a regression model are we trying to diagnose)?

The LINE Assumptions:

The aspects of a logistic regression model that we are trying to diagnose are violations of the linear regression assumptions (LINE: Linearity, Independence of the residuals, Normality of the residuals, and Equality of the variances of the residuals around the line (homoscedasticity); using the residuals to take a look at if there are any influential outliers; and also we want to eliminate any multicollinearity (a phenomenon in which one or more predictor variable(s) in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy), and no autocorrelation. There are hypothesis tests and plots that can be looked at to detect any violation of the LINE assumptions, influential outliers, and multicollinearity. This leads into a bunch of regression diagnostics that seek to validate the regression models. The diagnostic tests and plots for the LINE assumptions are plentiful. It is preferred a sample size of greater than or equal to 30, so that the CLT will apply.

For each of them (e.g. violation of the underlying assumptions, leverage and influence of outliers, multicollinearity, etc.), what kind of diagnostic methods are available?

- ***Describe them and explain the associated cut-off values. Discuss the relationships among them.***
- ***What are some of the remedial measures, when the diagnostic tools detect any problem?***

Violations of the LINE assumptions:

- **Linearity:**

  For linearity, you can just graph the data and see if there is a linear appearance (easy for simple linear regression). You can also make a Residual Plot against the predictor variable (or rather, an observation vs. residual plot), as seen in figure 1. There should not be any systematic

relationship between the predicted and the residual points if the variable if it is linear related. If the data is not obviously going to be easily fit by a regression line, then doing fit a linear regression model on it. Try out another model, such as a quadratic, cubic, or trigonometric model.

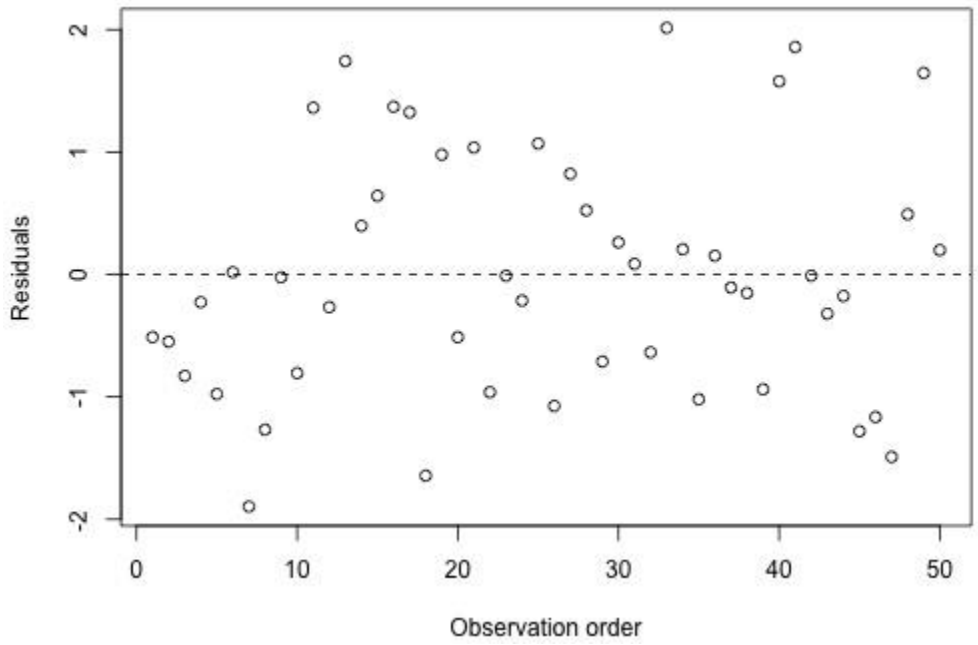
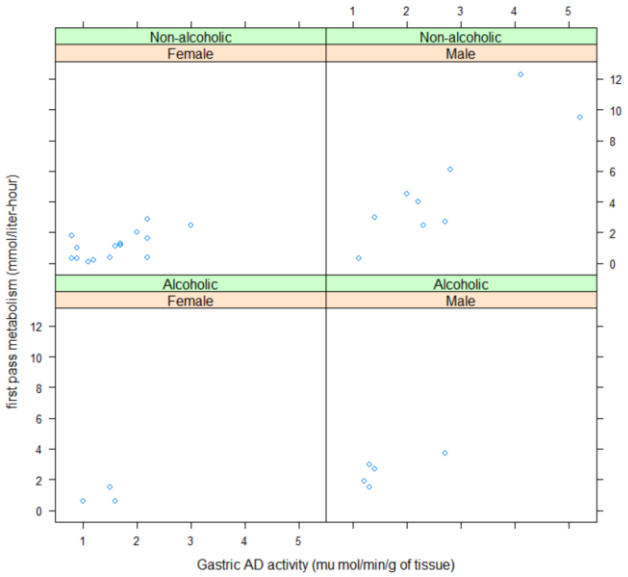


figure 1. Source: https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjTtN-V1PPhAhUJC6wKHSTLA4gQjB16BAgBEAQ&url=https%3A%2F%2Fnewonlinecourses.science.psu.edu%2Fstat501%2Fnode%2F280%2F&psig=AOvVaw3TIaprtVas_oM96rk2xX1k&ust=1556570773668331

I fit a multiple linear regression model on the data set that I will be using in this project, a dataset that I used in my regression class from the Statistical Sleuth Textbook, second or third edition, dataset 11.01. Do men and women metabolism alcohol differently? I fit a full model for estimating metabolism with 32 patients' sex, alcohol consumption (yes or no), and gastric activity.

```
> summary(lm1)

Call:
lm(formula = Metabol ~ Gastric + Sex + Alcohol + Gastric * Sex +
    Sex * Alcohol + Gastric * Alcohol + Gastric * Alcohol * Sex,
    data = case1101)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4286 -0.6189 -0.0466  0.5150  3.6516

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -1.6597     0.9996  -1.660   0.1099
Gastric                           2.5142     0.3434   7.322 1.46e-07 ***
SexFemale                         1.4657     1.3326   1.100   0.2823
AlcoholAlcoholic                  2.5521     1.9460   1.311   0.2021
Gastric:SexFemale                -1.6734     0.6202  -2.698   0.0126 *
SexFemale:AlcoholAlcoholic       -2.2517     4.3937  -0.512   0.6130
Gastric:AlcoholAlcoholic         -1.4587     1.0529  -1.386   0.1786
Gastric:SexFemale:AlcoholAlcoholic  1.1987     2.9978   0.400   0.6928
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.254 on 24 degrees of freedom
Multiple R-squared:  0.8277,    Adjusted R-squared:  0.7774
F-statistic: 16.47 on 7 and 24 DF,  p-value: 9.354e-08
```

The F statistic to test if the linear regression model is significant has a significant p-value at any significance level that is reasonable, so at least one of the Betas in the regression model is not equal to zero. So, a multiple linear regression model is appropriate here.



In addition, I graphed the independent variables against the response variable metabolism (Metabol). This is a plot of each gastric AD activity independent variable and the binary categories of alcohol and sex against the response variable. As you can see, there is a clear linear trend for each of the four plots, so the linearity assumption is satisfied.

- **Independence (of the Residuals): Autocorrelation**

Autocorrelation is A.K.A. serial correlation; often, this is with time series data and models. Autocorrelation can be tested with some plots and hypothesis tests, such as the residuals vs. observation order plot and the Durbin-Watson test of autocorrelation. Beware of time-dependent

regression model, as these almost always have an issue that involves autocorrelation. Accordingly, "Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of y(x+1) is not independent from the value of y(x)." So, this violates the i.i.d. r.v. of the residual errors that we need in order to satisfy the regression model assumptions.

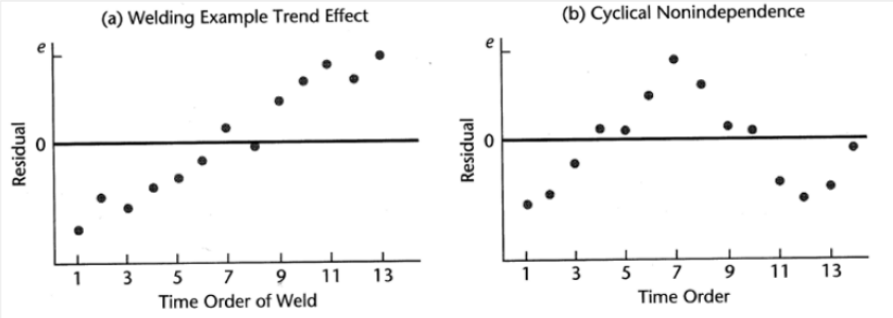


Image from: https://web.njit.edu/~wguo/Math344_2012/Math344_Chapter%203_part1.pdf

If you use the residual vs. observation order and see a clear pattern, such as a linear trend or peaks and valleys to the residuals against the observation order, then you have residuals that are dependent n each other in some way, as shown in the figure below. You can also use the Durbin-Watson hypothesis test, which tests the null hypothesis that the residuals are independent (not auto correlated) vs. the alternative that they are dependent on each other (auto correlated with each other). Accordingly, "As a rule of thumb values of $1.5 < d < 2.5$ show that there is no auto-correlation in the data." The test statistic is the chi-square, as this is a test of independence of the residuals. Again, autocreelation is often found in time series, so the p-value will be significant in these cases.

```
> dwtest(lm1)

        Durbin-Watson test

data:  lm1
DW = 1.8651, p-value = 0.114
alternative hypothesis: true autocorrelation is greater than 0
```

Because there was no time dependence in this model, I hypothesized that the linear model lm1 will not have autocorrelation of the residuals. I used the Durbin Watson test for autocorrelation. The result was insignificant at any reasonable significance level. So, we cannot conclude that the true autocorrelation is greater than 0, and thus it is probably that the residuals are not autocorrelated or dependent on each other.

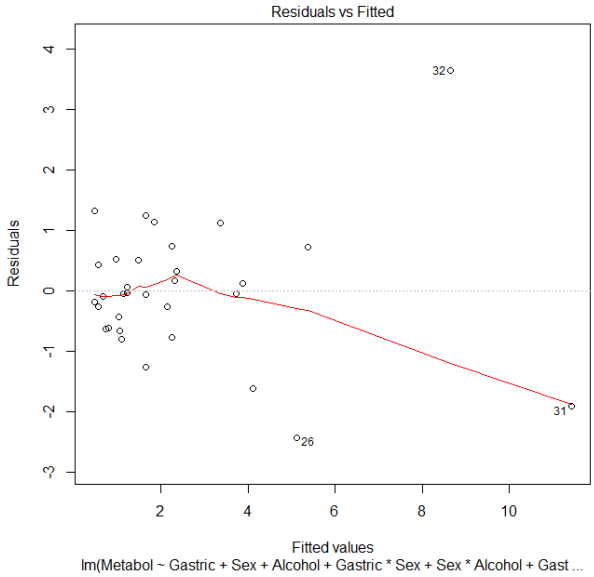
One remedy is to remove this problematic variable(s) from the linear regression model, or to fit another, more appropriate model onto it, such as a time series. Or, you can find another variable that you might have missed to put into the regression model that may help fix this problem and to make the model useful. Another way, if you don't know what variable (if you have a multiple linear regression), or when you find other issues with removing the variable that brought in autocorrelation, is to use the HAC (heteroscedasicity and autocorrelated errors) estimator.

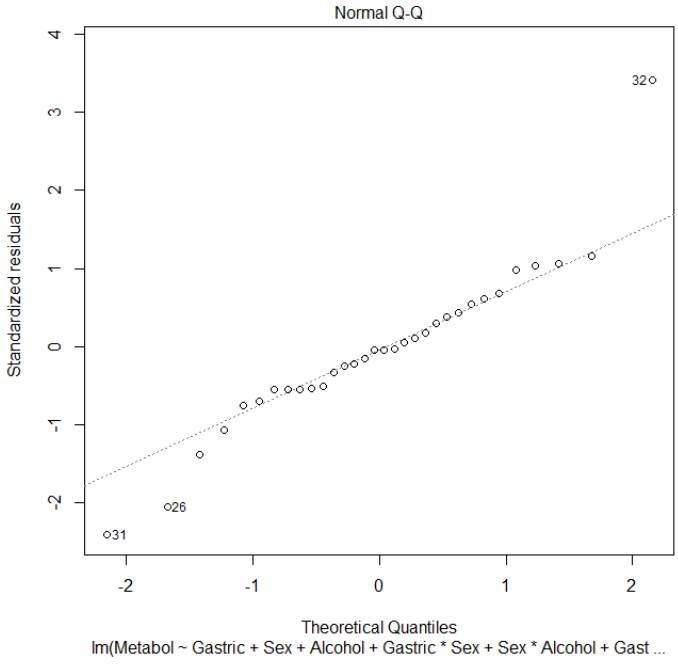
- **Normality (of the Residuals):**

The residuals should be normal and randomly distributed around the regression line. We can look at a normal QQ plot of the residuals, or a normal probability plot (which consists of plotting

the observed of the k-th smallest observation against the expected value k-th smallest observation), a boxplot of the residuals, or we can compare the frequencies (68 percent of the residuals fall between $\sqrt{MSE}$ or about 90 percent between $1.645\pm\sqrt{MSE}$). If the data is not normal, a transformation on the observed data points, such as a log or a cubic transformation to the predictor(s), may be appropriate, in order to fit a linear regression model.

When I fit the above multiple regression model on the data, I looked at the asusmptions. There was a violation in the normality assumption, as shown below. Later, we will resolve this issue by removing the outliers.





The red line above is curved due to the outlier observation 31 and 32 in the residuals vs. fitted value. Ideally, the red line would be close to the dotted line.

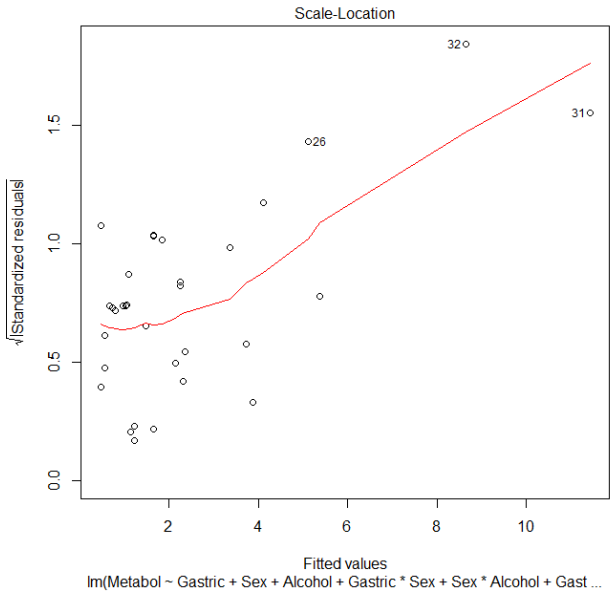
In the normal QQ plot of the residuals, you can see that observations 31 and 32 are for sure outliers in this case, as they make the tails fat in the QQ plot. Perhaps the removal of these outliers will make the normality assumption of the new fitted multiple linear regression model will have a satisfied normally assumption, so that the points will line up on the line.

- **Equal variance (of the residuals):**

We have the predictor variables to be fixed. We want to look at the residuals to see if they are

homoscedastic. This means that the residuals have equal variance around the regression line. That is, the residuals must be equal around the regression line. The residuals should be randomly distributed in equal ways around the regression line. If there is a violation of the equal variance of the residuals assumption, a transformation to the predictor variable(s) can be applied, such as a log transformation, or whatever may be appropriate.

There are several ways to detect if there is a violation of the assumption, or that the data does not violate this assumption. Although a more subjective way of looking to check this assumption of the residuals, looking at a scatterplot, a plot of the residuals vs. the observed values, or a plot of the residuals vs. the fitted values. The residuals should be equally and randomly distributed around the regression line in the scatterplot, or around the constant zero line in the two residual plots. In the above residual vs. fitted plot, the outliers demonstrate that there is a violation of the variance assumption because all of the data points do not fall equally around the dotted "zero" line. Also, notice the corrected standardized residuals vs. fitted values plot below. The points 31 and 32 deviate far from where all of the other data points are. This demonstrates furthermore that there is a normality and equal variance violation.

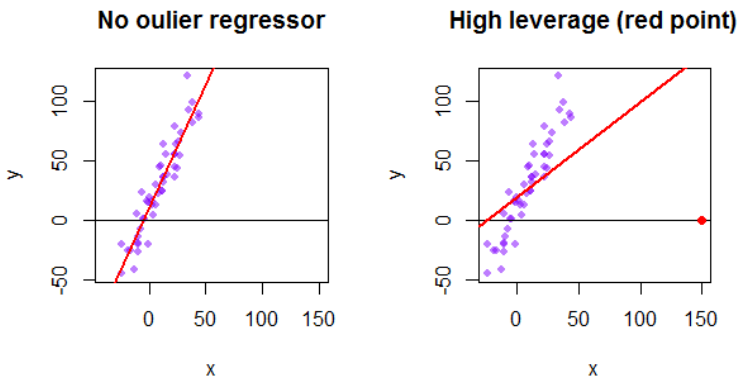


In addition, there are hypothesis tests that can be used to detect a equal residual variance assumption violation. Accordingly, "The Goldfeld-Quandt Test can be used to test for heteroscedasticity. The test splits the data into two groups and tests to see if the variances of the residuals are similar across the groups. If homoscedasticity is present, a non-linear correction might fix the problem." You can also use a couple of other tests for constancy of variance: the Breusch-Pagan test or the Brown-Forsythe.

Leverage and influence of outliers:

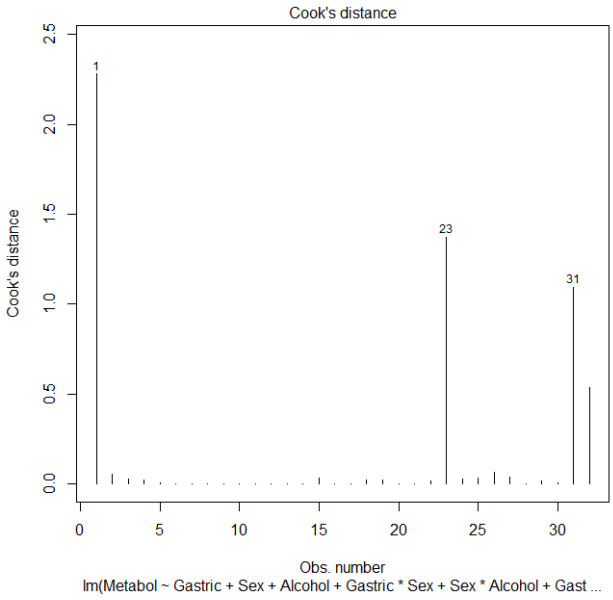
Leverage is defined as the "measure of the distance between its explanatory variable values and the average of the explanatory variable values in the entire data set" (Ramesy). When a case has a high leverage, it has a high possibility (no pun intended) of having a strong influence on the fitted model.

Leverage is often denoted as h_i, where $\frac{1}{n} \leq h_i \leq 1$. The mean of all leverages is $\frac{p}{n}$, with p representing the total number of the regression coefficients (so the mean leverage is $\frac{1}{n}$ in the simple linear regression case). When a residual has low variability, it has high leverage, and it thus dictates where the regression line is going, as seen in the figure below, where the high leverage point is on the mean zero line, and almost all of the data points are far away from the high leverage point. As you can see, this point has an unusually high x value, so it "pulls" the regression line to the right, in this case.



Source: https://stats.stackexchange.com/questions/208242/hat-matrix-and-leverages-in-classical-multiple-regression

Accordingly, "while a large leverage does not necessarily indicate that the case is influential, it does not imply that the case has a high potential for influence" (Ramesy). So, to detect if a value of leverage h_i is large enough to where further attention is needed, see if the observation is twice its mean of $\frac{p}{n}$, or in other words, if it is $> \frac{2p}{n}$. You can also take out the value and fit the regression line on the data without it to see if the regression equation and R^2 and other measures have significantly changed without it, by comparing the model with and without the potential high leverage point.



Also, you can use Cook's distance (typically denoted D_i) to measure the overall influence of a data point. Typically, a D_i close to one indicates a large influence that the data point has on the regression equation. Then, by indicating the potentially influential cases, the model can be refitted with and without such a point to see if a better regression model is found without it. In the Cooks distance chart, we see that observations 23 and 31 have higher leverages than the other observations.

Outliers are similar to high leverage points; they have unusual x and y values. Outliers can be marked as outliers by looking at the studentized residuals, which, accordingly is "a residual divided by its estimated standard deviation." Any studentized residual less than -2 or greater than 2 indicates that its

corresponding observation that could possibly be an outlier and is this worth further investigation. Below, we see that observation 31, 32, and 26 have studentized residuals that show that they are outliers, in the alcohol dataset.

```
> case1101[31,]
   Subject Metabol Gastric  Sex       Alcohol       hat    cooks   studres
31      31     9.5     5.2 Male Non-alcoholic 0.6005996 1.095949 -2.716715
> case1101[32,]
   Subject Metabol Gastric  Sex       Alcohol       hat     cooks  studres
32      32    12.3     4.1 Male Non-alcoholic 0.2699034 0.5365072 4.642475
> case1101[26,]
   Subject Metabol Gastric  Sex       Alcohol       hat      cooks   studres
26      26     2.7     2.7 Male Non-alcoholic 0.1113424 0.06607678 -2.214861
> case1101[23,]
   Subject Metabol Gastric  Sex   Alcohol       hat    cooks    studres
23      23     3.7     2.7 Male Alcoholic 0.9899244 1.371112 -0.3278581
```

Multicollinearity:

Next, multicollinearity can be a major problem to the regression fitting if the data is not tested and fixed against it. Multicollinearity is defined as "when the independent variables are too highly correlated with each other." If there is multicollinearity detected, then the simplest way of riding the model of multicollinearity is to remove the independent variables that have high VIFs (or small Tolerance T). However, another way to address this problem is to center the data, but the former way is more common.

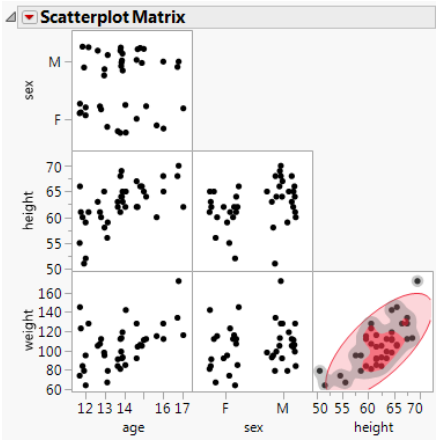


Image from: https://www.jmp.com/support/help/14-2/scatterplot-matrix-2.shtml

There are several ways to detect if there is a multicollinearity issue: a correlation matrix, as seen in the figure above, tolerance, and the variance inflation factor (abbreviated as VIF) between two or more numeric variables. A correlation matrix of the independent variables vs. other independent variables tells us if two of such variables have a high Pearson bivariate correlation coefficient (call this value r; let high correlation be defined as $0.4 < r < 1$). Tolerance (denoted as T) is measurement of one of the independent variable's influence on the other independent variables. As you can see above, for the numeric variables height, weight, and sex, all scatterplots besides height vs. weight, have no clear linear trend to them. This tells us that there is some correlation between height and weight.

This value is a function of the R^2 of the linear regression model. $T = 1 - R^2 = 1 - (SSR/SST) = SSE/SST$. Accordingly, "With $T < 0.1$ there might be multicollinearity in the data and with $T < 0.01$ there certainly is." So, as we will see these cut-off values are related and to the VIF. The Variance Inflation Factor (VIF) is equal to $1/T = 1/(1-R2)$. So, to detect multicollinearity, if $VIF > 1/0.1 = 10$, there is likely multicollinearity, and when $VIF > 1/(0.01) = 100$, there definitely is multicollinearity present among two or more of the independent variables.

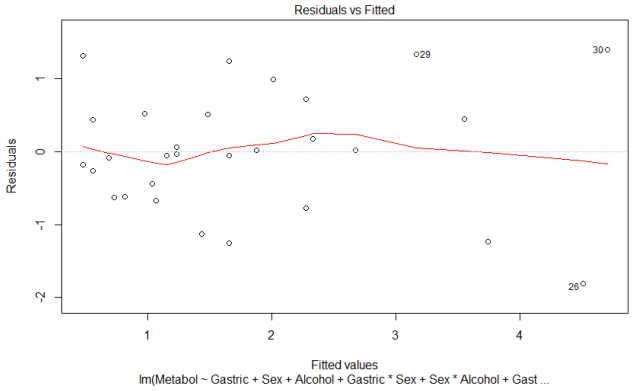
Because there is only one numeric variable here, and two categorical (binary) variables, it is not appropriate to test and look for and multicollinearity violations. If there were two numeric variables, then a correlation matrix would be performed above and the VIFs would be looked at.

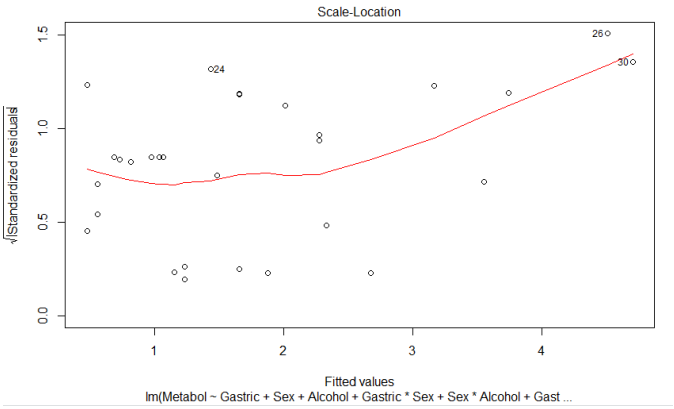
Final, fixed up model and model comparison: How do men and women (differently) metabolize alcohol?

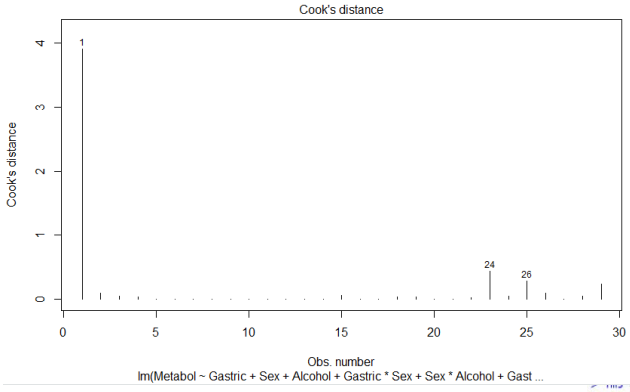
```
> case1101_2 = case1101[-c(31, 32, 23),]
> lm2 = lm(Metabol~Gastric+Sex+Alcohol+Gastric*Sex+Sex*Alcohol+Gastric*Alcohol+Gastric*Alcohol*Sex, data=case1101_2)
> summary(lm2)

Call:
lm(formula = Metabol ~ Gastric + Sex + Alcohol + Gastric * Sex +
    Sex * Alcohol + Gastric * Alcohol + Gastric * Alcohol * Sex,
    data = case1101_2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.80764 -0.61492 -0.03528  0.51250  1.40024

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -0.6797     1.3338  -0.510  0.61564
Gastric                           1.9212     0.6197   3.100  0.00542 **
SexFemale                         0.4858     1.4943   0.325  0.74833
AlcoholAlcoholic                 -2.2453     8.9275  -0.252  0.80387
Gastric:SexFemale                -1.0805     0.7348  -1.471  0.15624
SexFemale:AlcoholAlcoholic        2.5457     9.4218   0.270  0.78965
Gastric:AlcoholAlcoholic          2.0788     6.8085   0.305  0.76313
Gastric:SexFemale:AlcoholAlcoholic -2.3388    7.1386  -0.328  0.74643
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9589 on 21 degrees of freedom
Multiple R-squared:  0.6689,    Adjusted R-squared:  0.5586
F-statistic: 6.061 on 7 and 21 DF,  p-value: 0.0005845
```

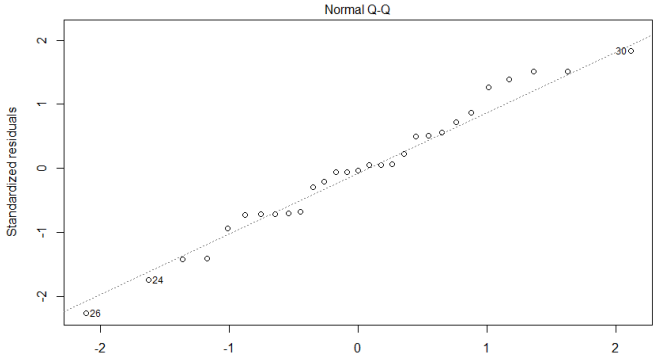
Here, 3 of the points were removed, making the dataset have 29 points. These points were observation 31, 32, and the high leverage point 23. The linear model is significant, according to the F-statistic above.

Let's look at the assumptions of this second linear model. In the residuals vs. fitted plot and studentized residuals vs. fitted values, the red line is closer to the zero line and all of the observations fall equally within the [-2, 2] bounds, so there are not any outliers to be concerned about. The normal QQ plot of the residuals shows that the residuals are normal now, because the observations fall closely on the dotted line. In the cooks distance plot, there are no observations that stand out, so there are not any high leverage points to worry about.









Now, let's fit another model with no alcohol, to see if there is an effect of alcohol on metabolism of males and females by comparing a couple of models. So if we remove alcohol from the model (as seen in lm3) and then run an ANOVA from lm2 and lm3 (both with the modified dataset), we see that the p-value of the F statistic is 0.9725 at the end, so there is no alcohol effect because the two models are probably similar in how they predict Metabol.

```
> summary(lm3)

Call:
lm(formula = Metabol ~ Gastric + Sex + Gastric * Sex, data = case1101_2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77580 -0.59107 -0.02554  0.44402  1.45386

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.12332    0.84682  -0.146 0.885381
Gastric           1.70338    0.44853   3.798 0.000831 ***
SexFemale        -0.07395    1.03251  -0.072 0.943476
Gastric:SexFemale -0.86643    0.57285  -1.512 0.142948
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8891 on 25 degrees of freedom
Multiple R-squared:  0.6612,    Adjusted R-squared:  0.6205
F-statistic: 16.26 on 3 and 25 DF,  p-value: 4.538e-06

> #compare the 2 models lm2 and lm3
> anova(lm3, lm2)
Analysis of Variance Table

Model 1: Metabol ~ Gastric + Sex + Gastric * Sex
Model 2: Metabol ~ Gastric + Sex + Alcohol + Gastric * Sex + Sex * Alcohol +
    Gastric * Alcohol + Gastric * Alcohol * Sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     25 19.761
2     21 19.308  4   0.45325 0.1232 0.9725
> #p-value of the F statistic is 0.9725, so the data is consistant and there is no alcohol effect between men and women.
```

Sources for Regression Diagnostics:

https://www.statisticssolutions.com/assumptions-of-linear-regression/

autocorrelation: https://newonlinecourses.science.psu.edu/stat501/node/280/

https://web.njit.edu/~wguo/Math344_2012/Math344_Chapter%203_part1.pdf

"The Statistical Sleuth: A course in methods of Data Analysis," Ramsey, Schafer AND Dr. Hurtado Rua Linear Regression Lecture Notes, Fall 2017 at Cleveland State University

HAC: https://www.econometrics-with-r.org/15-4-hac-standard-errors.html


```
#There was a very nice data set from a textbook that I used when I was learning regression for the first time.

#It was in the Sleuth3 dataset package that came with the Sleuth textbook, using the mosaic package.

#This data set had data about how men and women process alchohol. case1101 in the Sleuth dataset packages

install.packages("mosaic")

require(mosaic)
```

```
install.packages("Sleuth3")

require(Sleuth3)

summary(case1101)

xyplot(Metabol ~ Gastric | Sex+Alcohol, data = case1101, auto.key =
TRUE, xlab = "Gastric AD activity (mu mol/min/g of tissue)",

       ylab = "first pass metabolism (mmol/liter-hour)")

case1101 = transform(case1101, Sex = factor(Sex, levels =
c("Male","Female")))

case1101 = transform(case1101, Alcohol = factor(Alcohol, levels =
c("Non-alcoholic","Alcoholic")))

case1101

lm1 =
lm(Metabol~Gastric+Sex+Alcohol+Gastric*Sex+Sex*Alcohol+Gastric*Alcohol
+Gastric*Alcohol*Sex, data=case1101)

summary(lm1)

install.packages("MASS")

require(MASS)

case1101 = transform(case1101, hat = hatvalues(lm1))

case1101 = transform(case1101, studres = studres(lm1))

case1101 = transform(case1101, cooks = cooks.distance(lm1))

case1101[31,]

case1101[32,]

case1101[23,]

plot(lm1, which=1)

plot(lm1, which=2)

plot(lm1, which=3)

plot(lm1, which=4) #cooks

plot(lm1, which=5) #residuals vs. Leverage

plot(lm1, which=6) #Cooks dist vs Leverage

#durbin waston test for autocorrelation of the residuals

install.packages("lmtest")

require(lmtest)
```

```
dwtest(lm1)

#from diagnostics, appears that obs. 31 and 32 are influenctial
points. refit the full model without 31 and 32.

case1101_2 = case1101[-c(31, 32, 23),] #remove the three influential
observations

lm2 =
lm(Metabol~Gastric+Sex+Alcohol+Gastric*Sex+Sex*Alcohol+Gastric*Alcohol
+Gastric*Alcohol*Sex, data=case1101_2)

summary(lm2)

plot(lm2, which=1)

plot(lm2, which=2)

plot(lm2, which=3)

plot(lm2, which=4) #cooks

#refine the model to use for comparing with the modle with all
potential terms

#remove alcohol from the equation

lm3 = lm(Metabol~Gastric+Sex+Gastric*Sex, data=case1101_2)

summary(lm3)


#compare the 2 models lm2 and lm3

anova(lm3, lm2)

#p-value of the F statistic is 0.9725, so the data is consistent and
there is no alcohol effect between men and women.
```

5. Ordinal Logistic Regression (Liver Fibrosis) and Multiple Logistic Regression (PHR)

Multivariate and ordinal logistic regression and their applications in biostatistics: Under what circumstances are they needed? Write down each model (including the underlying assumptions), describe how to fit the model (i.e. parameter estimates and inference on the parameters). Describe residuals and how to use these residuals. Find at least 2 datasets appropriate for each kind of regression, fit the appropriate regression model on each dataset and discuss the results.

The circumstances under which ordinal logistic regression is needed is when the response variable (the variable that we want to predict) may not just be binary (0 or 1, yes or no), but could be when we want to predict any ordinal response variable, such as yes, maybe, or no, or even more n ordered categories that we want to predict. Ordinal logistic regression “It can be considered as either a generalization of multiple linear regression or as a generalization of binomial logistic regression.” For example, you may want to predict what weight class someone is in: healthy, overweight, or obese. The

potential predictors could be diet type (low fat/low sugar, exercise level (moderate, high intense, little to no), age, gender, income level, etc., depending if you want to build a simple ordinal logistic regression model (just include the most significant variable) or multiple ordinal logistic regression model (include two or more significant predictor variables to get a useful model).

The assumptions of the ordinal logistic regression are that the dependent variable is ordinal, or that there is a rank to the categories in in the response, such as high, medium, and low. However, the independent variables can be ordinal, nominal, or just numerical variable types. Next is the assumption of no multicollinearity between the independent/predictor variables, or that the predictor variables are not highly correlated with each other. According to O'Connell, "Unfortunately, testing for this assumption can require creating dummy variables for your categorical variables (i.e., dummy variables are new variables based on the values of your existing data)." The number of dummy variables depends on the number of categories in the categorical variables that you may have in the predictor variables. For example, if you only have one categorical independent variable with just three groups (e.g., the variable, "exercise level", with three groups: "moderate", "high intense", "little to no"), you will only have to create two dummy variables. Now, if in the case you have more than one categorical variable, there will potentially be a lot of dummy variables (if the categories in some or all of the categorical variables have more than two variables, for example.)

After looking at multicollinearity (if you have a multiple ordinal logistic regression model), you will need to look at proportional odds, which is the assumption where each of the independent variables (in a multiple ordinal logistic regression model), has the same effect at each of the cumulative split(s) if the ordinal dependent variable. This can be tested using a full likelihood ratio test comparing the fitted location model to one with different types of location parameters. If the model is a simple logistic regression model, well, then it is kept simple with regards to the assumptions. There is no concern of multicollinearity and proportional odds.

Next, here is the model of the ordinal logistic regression model, for *k* ordered categories in the predictor variable:

$$\text{logit}(p_1) \equiv \log \frac{p_1}{1 - p_1} = \alpha_1 + \beta' x$$

$$\text{logit}(p_1 + p_2) \equiv \log \frac{p_1 + p_2}{1 - p_1 - p_2} = \alpha_2 + \beta' x$$

$$\vdots$$

$$\text{logit}(p_1 + p_2 + \cdots + p_k) \equiv \log \frac{p_1 + p_2 + \cdots + p_k}{1 - p_1 - p_2 - \cdots - p_k} = \alpha_k + \beta' x$$

$$\text{and } p_1 + p_2 + \cdots + p_{k+1} = 1$$

Image from http://staff.washington.edu/glynn/olr.pdf

So, in other words, for our example of the weight class predicted variable above, we would have $\ln(y_{k=3}') = \text{logit}(\hat{\pi}(x)) = \text{logit}(p_1+p_2+p_3) = \log(\frac{p_1+p_2+p_3}{1-(p_1+p_2+p_3)}) = \alpha_3 + \beta'x$, if we have a simple ordinal logistic regression model, with x being the most significant predictor variable available. $\ln(y_{k=3}') = \text{logit}(\hat{\pi}(x))=\ln(\pi_k(x))/(1-\pi_k(x)))= \text{logit}(p_1+p_2+p_3) = \log(\frac{p_1+p_2+p_3}{1-(p_1+p_2+p_3)}) = \alpha_3 + \beta_1'x_1 + \beta_2'x_2$ is the model for

a ordinal logistic regression model with two predictors, and so on. α_3 (the intercept) is the threshold for each particular split to the dataset. The above models all predict the expected logit in its category k, conditioned on the predictor(s). $y_{k=3}'$ is the odds of being placed in the higher proficiency categories. Now, if first the predicted logits are transformed into the odds, then they can be transformed into the approximate probability $P(y \leq k) = 1 - \frac{e^{\ln(y_3')}}{1+e^{\ln(y_3')}} = \frac{1}{1+e^{\ln(y_3')}}$.

The next discussion for ordinal logistic regression is how to deal with and interpret the model's residuals. Logistic regression models are heteroscedastic (meaning that there are sub-populations that have different variabilities from other: variation = $\pi_i(1-\pi_i)$. Logistic regression models use the Maximum likelihood procedures to obtain coefficients estimates of which have the greatest likelihood of predicting the outcome variable categories, instead of using the resisuals in a least squares of residual approach that is used in linear regression.

A far as residuals for logistic regression, if you want to peak at them to try to see if there are any outliers, it is common to use the Pearson residual, the deviance residual, and/or the Pearson leverages. Pearson residuals = $r_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i}(1-\hat{\pi}_i)}$, with y_i being the observed number of success, n_i is the number of observations that have the number of observations with x_i, $\hat{\pi}_i$ is estimated probability at the observation x_i. These are the main residuals that I will be using. There was not an option in SPSS to save the residuals, so I did not do any analysis on them. Residuals, in general, are not very useful in logistic regression models. There are not any assumptions of residuals that need to be met like in linear regression diagnostics and t method of estimating the coefficients of the terms is not based on the residuals, but on the maximum likelihood.

LIST OF VARIABLES:

Name	Codes/Values	Abbreviation
Age	years	Age
Gender	1 = male; 2 = female	Sex
Height	cm	Height
Weight	kg	Weight
Body mass index	kg/mg^2	BMI
Duration of obesity	year	Obesity Duration
Diabetes	0 = no; 1 = yes	DM
Metabolic syndrome	0 = no; 1 = yes	MET Syndrome
Hypertension	0 = no; 1 = yes	HTN
Hyperlipidemia	0 = no; 1 = yes	HPL
Plasma triglycerides	%	TG
Cholesterol	mg/dL	CHOL
High-density lipoprotein cholesterol	mg/dL	HDL
Low-density lipoprotein cholesterol	mg/dL	LDL
Very-low-density lipoprotein cholesterol	mg/dL	VLDL
Aspartate aminotransferase	U/L	AST
Alanine aminotransferase	U/L	ALT
Nonalcoholic fatty liver disease activity score	range from 0 to 8	NAS
Fibrosis	0 = none; 1 = perisinusoidal or periportal; 2 = perisinusoidal and portal/periportal; 3 = bridging fibrosis; 4 = cirrhosis	Fibrosis
Positive liver steatosis by ultrasound	0 = no; 0.5 = inconclusive; 1 = yes	LS+ US
Positive liver steatosis by biopsy	0 = no; 0.5 = inconclusive; 1 = yes	LS+ Biopsy

Now, here is the biology and background, and why I am trying to fit the model on the dataset fibrosis, from the Cleveland Clinic. I am going to perform a simple ordinal logistic regression model, with the response: liver steatosis by ultrasound (LSUS) (No, inconclusive, yes are the levels), and the Explanatory variable being Nonalcoholic fatty liver disease activity score (NAS) numeric variable (score of 0 to 8). The higher levels of NAS being more severe forms of steatosis of the liver, and zero being the nonalcoholic fatty liver disease activity score (NAS) is known to be a good indicator of liver steatosis. My model looks like this: $\ln(y_{k=3}') = \text{logit}(\hat{\pi}(x)) = \text{logit}(p_1 + p_2 + p_3) = \log(\frac{p_1+p_2+p_3}{1-(p_1+p_2+p_3)}) = \alpha_3 + \beta'x$, where y is

liver steatosis by ultrasound status and x is NAS. Note the NAS is an ordinal numeric predictor variable, like the LSUS.

I wanted to see if a person with a low value of NAS will have a lower probably of testing positive to liver steatosis by ultrasound, and to see if its results matched up well will detecting liver steatosis by ultrasound. This, so I used the discrete numerical NAS score to predict this ordinal 3 category response variable. According to the article where the liver fibrosis dataset came from, diagnosis of liver steatosis by ultra sound is the "golden standard;" it is nearly always right. Predicts more accurately out of all the diagnosis measures by looking inside the patients.

I did this analysis in SPSS. To help with the analysis of the output, I read through and used this source: https://stats.idre.ucla.edu/spss/dae/ordinal-logistic-regression/

```
DATASET ACTIVATE DataSet1.

SAVE OUTFILE='E:\Grad School 2018-2019\Spring 2019\Biostatistics Independent
'+
    'Study\Projects\liversteatosisordinallogistic.sav'
  /COMPRESSED.
PLUM LSUS BY NAS
  /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
  /LINK=LOGIT
  /PRINT=FIT PARAMETER SUMMARY TPARALLEL
  /SAVE=ESTPROB PREDCAT PCPROB ACPROB.
```

PLUM - Ordinal Regression

```
[DataSet1] E:\Grad School 2018-2019\Spring 2019\Biostatistics Independent
Study\Projects\liversteatosisordinallogistic.sav
```

Case Processing Summary

		N	Marginal Percentage
Positive Liver Steatosis by Ultrasound	no	111	27.9%
	inconclusive	3	0.8%
	yes	284	71.4%
NAS	0	109	27.4%
	1	96	24.1%
	2	73	18.3%
	3	43	10.8%
	4	32	8.0%
	5	28	7.0%
	6	15	3.8%
	7	2	0.5%
Valid		398	100.0%
Missing		45	
Total		443	

In the Case Processing Summary table, we see the number and percentage of cases in each level of our response variable. Positive liver steatosis is 71.4% yes, 0.8% inconclusive, and 27.9% no. There are 45 missing observations. These numbers look fine, but we would be concerned if one level had very few cases in it (that is, less than 45). We also see that 398 out of the 443 observations in our data set were used in the analysis. Fewer observations have been used because the variables had missing values. By default, SPSS does a list wise deletion of cases with missing values. We want the LSUS to be zero because then the patient will likely not have liver steatosis.

Next, we see the Model Fitting Information table, which gives the -2 log likelihood for the intercept-only and final models. The -2 log likelihood can be used in comparisons of nested models, but it will not be used here.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	152.108			
Final	32.875	119.233	7	.000

Link function: Logit.

The null hypothesis for the goodness of fit tests are that the model is a good fit. The alternative is that the model is not a good fit. Because the p-values of both the Pearson and Deviance tests are not significant at any reasonable significance level, we do not reject the null hypothesis. This model is probably a good fit.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	5.211	7	.634
Deviance	4.747	7	.691

Link function: Logit.

Pseudo R-Square

Cox and Snell	.259
Nagelkerke	.360
McFadden	.236

Link function: Logit.

Now, taking a look at the Psuedo R^2 table, both the Cox & Snell R square (which takes into account the log likelihood of the final model vs. the log likelihood of the baseline model) and Nagelkerke R square (a modification of the Cox & Snell), and McFadden statistics are shown. These are typically somewhat useful for comparing different logistic regression models. However, these measures are substitutes (or Pseudo R2 measures) two of the true R2 in a least squares regression model, and are not as easy to interpret as other measurements of model evaluation. Because we only have one model, they are not of much use here.

According to the output in the parameter estimate table, the model predictor NAS is a good predictor of Liver Steatosis status by ultrasound, because all of the Wald chi-square test statistics are significant, and the Wald chi square test statistics of the 3 levels of the LS US are significant.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[LSUS = 0]	-19.512	.758	663.013	1	.000	-20.997	-18.027
	[LSUS = 1]	-19.459	.757	660.232	1	.000	-20.944	-17.975
Location	[NAS=0]	-20.174	.784	662.386	1	.000	-21.710	-18.637
	[NAS=1]	-18.258	.795	527.500	1	.000	-19.816	-16.700
	[NAS=2]	-17.355	.846	421.302	1	.000	-19.012	-15.698
	[NAS=3]	-17.411	.896	377.412	1	.000	-19.168	-15.655
	[NAS=4]	-17.520	.926	357.896	1	.000	-19.335	-15.705
	[NAS=5]	-4.523E-5	3176.658	.000	1	1.000	-6226.135	6226.135
	[NAS=6]	-17.594	.000	.	1	.	-17.594	-17.594
	[NAS=7]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

In this Parameter Estimates table we see the coefficients, their standard errors, the Wald test and associated p-values (Sig.), and the 95% confidence interval of the coefficients. The first five levels of NAS are statistically significant. So, for NAS = 0, we expect a decrease in 20.174 in the ordered log odds of being in a higher level of LSUS (going from 0 to 1), given all of the other values of NAS in the model are held constant. When NAS = 1, we expect a decrease in 18.258 in the ordered log odds of being in a higher level of LSUS (going from 0 to 1), given all of the other values of NAS in the model are held constant. The 95% confidence interval for when NAS =1 is as follows; when NAS =1, we expect a decrease in 19.816 to 16.7 in the ordered log odds of being in a higher level of LSUS (going from 0 to 1), given all of the other values of NAS in the model are held constant. Similarly, when NAS = 2, we expect a decrease in 17.355 in the ordered log odds of being in a higher level of LSUS (going from 0 to 1), given all of the other values of NAS in the model are held constant. And so on, until when NAS = 5.

At the higher levels of NAS scoring (NAS ≥ 5), the estimates are not significant (or recorded) in predicting whether there is a decrease in the ordered log odds of being in higher level of LSUS (that is, that the patient is diagnosis formally with the Liver Steatosis disease by ultrasound). So, as the values of NAS increase, you would expect a larger chance of the order log odds of being in the higher level of LSUS.

Sources and Resources Used for Ordinal Logistic Regression:

https://statistics.laerd.com/spss-tutorials/ordinal-regression-using-spss-statistics.php

http://staff.washington.edu/glynn/olr.pdf

https://newonlinecourses.science.psu.edu/stat501/node/374/

O'Connell, Ann A. and Liu, Xing (2011) "Model Diagnostics for Proportional and Partial Proportional Odds Models," Journal of
Modern Applied Statistical Methods: Vol. 10 : Iss. 1 , Article 15.
DOI: 10.22237/jmasm/1304223240

https://pdfs.semanticscholar.org/68bd/00550e88ed8dc5850821692ea3ba8f7c67d0.pdf

https://journal.r-project.org/archive/2018/RJ-2018-004/RJ-2018-004.pdf

The dataset represents data from the study by Wu et al. “Prevalence of Liver Steatosis and Fibrosis and the Diagnostic Accuracy of Ultrasound in Bariatric Surgery Patients”. ObesSurg 2012; 22: 240-247.

Multiple logistic regression

Next, we will look at multiple logistic regression is when we may have a categorical (typically nominal, binary) response variable, but we have many (two or more) potential predictor variables that we want to use to predict the response variable. These predictors can be numerical or categorical predictors. For example, you may want to predict whether a group of diabetic patients use a personal health record (PHR) system or not, based on a series of health and demographic data. The outcome variable in this case would be the use of the PHR system or not (yes or no). Potential predictors could be race, income level, health insurance type, age, and some sort measurement of the amount of glucose or insulin in the patient's blood (in this case, it will be HBA1C percentage). This is exactly what I will model here to demonstrate the multiple logistic regression model.

LIST OF VARIABLES:

Name	Codes/Values	Abbreviation
Age	years	Age
Gender	1 = female; 0 = male	Female
Race	1 = Caucasian; 0 = other	Caucasian
Insurance Type	1 = Commercial; 0 = other	Insurance
Household Income	US Dollars in thousands	Income
Provider Engagement	percentage of the physician's patients from the study sample who logged in at least 1 day during the study	Engagement
PHR user group	1 = User; 0 = Nonuser	User
Frequency of PHR use	number of logins during study period	Logins
Dilated retinal eye exam	eye exam recorded within study period 1 = yes; 0 = no	Eye Exam
Pneumococcal Vaccination	documented lifetime vaccination 1 = yes; 0 = no	Pneumo Vaccine
Attention to Kidneys	Use of ACEi/ARB and/or test for microalbuminuria within study period 1 = yes; 0 = no	ACE ARB ALB
Attention to Feet	documented foot exam within study period 1 = yes; 0 = no	Foot Exam
Smoking Cessation	documented nonsmoker 1 = yes ; 0 = no	Nonsmoking
HbA1c Test	HbA1c value measured within study period 1 = yes; 0 = no	HBA1C Test
HbA1c value	%, last documented value within study period	HBA1C
Blood pressure	mmHg, last documented value within study period	SBP
Blood pressure	mmHg, last documented value within study period	DBP
Cholesterol	mg/dL, last documented value within study period	LDL
Body Mass Index	kg/m^2, last documented value within study period	BMI

The model that I am looking at is:

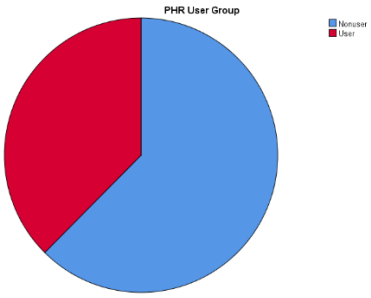
$$\ln[Y/(1-Y)]=a+b_1X_1+b_2X_2+b_3X_3\ldots$$

Where the a and b_i's are found using the maximum likelihood method. The maximum likelihood method basically will give the coefficients of the input variables that will most likelu predict the observed results. Here, for the PHR data set above, my model will be Y = User, $\ln(\frac{Y}{1-Y})$ = a + b_1*Age + b_2*Female + b_3*Caucasian + b_4*Insurance + b_5*Income + b_6*HBA1C + b_7*nonsmoker.

The model can be built with a variety of techniques, such as with stepwise variable selection, or forward or backward likewise variables selection. To determine model fit, the deviance measure (preferably the smaller the better the model), and some "pseudo R^2 values" can be used (the larger these values, the better the fit). Also, the Hosmer and Lemeshow Test, for example, is a test to determine whether a model is a good fit on the data or not. This will be discussed soon.

The observations used in multiple logistic regression must also be independent (here, we are looking at individual and independent patients). In addition, "Multiple logistic regression also assumes that the natural log of the odds ratio and the measurement variables have a linear relationship. It can be hard to see whether this assumption is violated, but if you have biological or statistical reasons to expect a non-linear relationship between one of the measurement variables and the log of the odds ratio, you may want to try data transformations" (according to http://www.biostathandbook.com/multiplelogistic.html). Lastly, one thing that is not assumed in multiple logistic regression is that the numerical predictor variables are normally distributed. So, we do

not need to consider any kind residual analysis analysis here, just like in the ordinal logistic regression case, as discussed above.



Taking a look at the pie chart of the binary response user, Right off the bat, we can see there are many more users than nonusers of the PHR system, as seen in the pie-chart in Figure 2. About two thirds of the diabetic patients use the Personal Health Record (PHR) system. About one third use this online system.

For the multicollinearity assumption, there does not appear to be correlations between the two numerical variables age, income, and HBA1C, according to this correlation matrix. So, the assumption of having no multicollinearity of the variable sis met.

Correlations

		Years of Age	Household income (US dollars)	HBA1C % last documented value within study period
Years of Age	Pearson Correlation	1	.050**	-.175**
	Sig. (2-tailed)		.000	.000
	N	10746	10558	10208
Household income (US dollars)	Pearson Correlation	.050**	1	-.086**
	Sig. (2-tailed)	.000		.000
	N	10558	10558	10032
HBA1C % last documented value within study period	Pearson Correlation	-.175**	-.086**	1
	Sig. (2-tailed)	.000	.000	
	N	10208	10032	10208

**. Correlation is significant at the 0.01 level (2-tailed).

I also imputed the numeric variables to replace the missing values with the Linear trend at point. This was one of the better options that I had in SPSS, (the options being: Series mean, Mean of nearby points, Median of nearby points, Linear interpolation, and Linear trend at point.) I chose linear trend because, accordingly, it "Replaces missing values with the linear trend for that point. The existing series is regressed on an index variable scaled 1 to n. Missing values are replaced with their predicted values." I thought that replacing the missing values with their predicted values would be the best than using the overall mean or the other methods.

Replace Missing Values: Linear Trend At Point

[DataSet1] E:\Grad School 2018-2019\Spring 2019\Biostatistics Independent Study\Projects\phr.sav

Result Variables

	Result Variable	N of Replaced Missing Values	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
			First	Last		
1	Income_1	188	1	10746	10746	SMEAN (Income)
2	Age_1	0	1	10746	10746	SMEAN(Age)
3	HBA1C_1	538	1	10746	10746	SMEAN (HBA1C)

The first time that I ran the logistic regression model, income_1 did not come out to be useful in the model (it was statistically significant, but the coefficient was recorded to be about zero because the sample size is large, so I reran the model without this variable, below are the results for this model. My modified model that was put into the program will be Y = User, $\ln(\frac{Y}{1-Y})$ = a + b_1*Age + b_2*Female + b_3*Caucasian + b_4*Insurance + b_5*HBA1C + b_6*nonsmoker. The method that was applied was the forward likelihood method of variable selection, so perhaps one or more of thse variables will not be in the best model.

The Hosmer and Lemeshow Test tests whether the model is a good fit or not. The null hypothesis is that the model is a good fit and the alternative is that it is not, according to Dr. Fridline's notes on multiple logistic regression. The model with the biggest p-value (= 0.037) is the one from step 4. It's p-value is not significant at the 1% level (but is significant at the 5% level) reasonable significance level, we fail to reject the null in the 1% case, but reject the null in at the 5% significance level. Because the test is a chi-square test, then this conclusion tells us that there is no difference between the observed and the expected, at the 1% level. This implies that the model estimates fit the data at an acceptable level, if we use the 1% significance level. But at the 5% significance level, using similar logic, the model does not fit the data well. This is okay. We will go with the model from step 4. As you will see, this model is the largest model with all of the terms being statistically significant at any reasonable significance level.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	0	.
2	10.233	2	.006
3	20.798	8	.008
4	16.433	8	.037
5	26.513	8	.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13817.728[a]	.037	.051
2	13469.679[a]	.068	.092
3	13379.521[a]	.076	.103
4	13315.453[a]	.081	.111
5	13254.407[a]	.086	.118

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

The Model Summary comparison gives the log likelihood and the pseudo R^2 statistics. These measures are substitutes (or Pseudo R^2 measures) two of the true R^2 in a least squares regression model, and are not as easy to interpret as other measurements of model evaluation. (The specificity, sensitivity, risk.) Both the Cox & Snell R square (which takes into account the log likelihood of the final model vs. the log likelihood of the baseline model) and Nagelkerke R square (a modification of the Cox & Snell) statistics increase as the steps increase (as more variables are added to the model) from 0.037 (step 1) to 0.086 (step 5) and from 0.051 (step 1) to 0.118 (step 5), respectively. The -2 log likelihood ratio also decreases as the number of variables added to the model increases from 13817.73 to 13254.407 in steps 1 and 5, respectively. The variables that were added in each step added significance to the model. Notice that the Pseudo R^2 estimates in steps 4 and 5 are nearly the same, and we should go with the model the fourth one) with the largest Cox & Snell R^2 measurements and smallest log likelihood. Notice that the

Step 4 above gives the best model that will be used here because it has the largest number of predictors of PHR that are statistically and practically significant, when compared to the other models. The logit of this model is logit($\hat{\pi}$) = -1.169 -0.146*HBA1C + 0.951*race+0.836*Insurance+0.538*nonsmoker. All of these terms are significant at the 5% level, according to the Wald tests (which tests the null hypothesis of the true coefficient of a predictor variable to be equal to zero vs. the alternative that it is not equal to zero. It used a chi-square test statistic) for each, as seen in the variables in the equation table above. So, we conclude that each of the predictor variable coefficients do not equal zero, at the 5% level of significance. The standard errors of each of the coefficients of the predictor variables are also relatively small, when compared to the coefficients.

The HBA1C coefficient interpretation is as follows: the adjusted odds ratio is $e^{-0.146} = 0.864$. While adjusted for the other variables in the model, we predict that the odds of a diabetic patient using the PHR system decreases by (((0.864-1) = -0.136) 13.6% for each additional unit of HBA1C. The 95% confidence interval for the odds of HBA1C is, for every increase in one unit of a diabetic patient's age, the predicted odds of using the PHR system decreases by ([(0.837-1) to (0.892-1)], or -0.174 to -0.118) 17.4% to 11.8% for each additional unit of HBA1C.

Note that that Race, Health Insurance type, and nonsmoker all have positive predicted odds (0.951, 0.836, and 0.538, respectively), and odds ratios (2.589, 2.308, and 1.496 respectively), and corresponding odds ratio confidence intervals ((2.343, 2.862), (2.120, 2.512), (1.496,1.960), respectively) that are greater than one. So, all of these variables are positively associated with the patients that use the PHR system. All of these variables are binary, with 1 denoting Caucasian, commercial health insurance, and 1 for being a nonsmoker, respectively.

For patients who are Caucasian, the adjusted odds ratio is $e^{0.951} = 2.589$. While adjusting for the other variables on the model, the predicted odds of using the PHR system is 2.589 times greater for those who are Caucasian than it is for those who are not. The 95% confidence interval for the odds of race is, for patients who are Caucasian, the predicted odds of using the PHR system is 2.343 to 2.862 times greater than those who are not Caucasian.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Race	.962	.050	371.453	1	.000	2.616	2.373	2.885
	Constant	-1.237	.044	777.017	1	.000	.290		
Step 2[b]	Race	.986	.051	379.210	1	.000	2.681	2.428	2.961
	Health Insurance Type	.782	.043	336.418	1	.000	2.186	2.011	2.377
	Constant	-1.725	.053	1044.357	1	.000	.178		
Step 3[c]	SMEAN(HBA1C)	-.151	.016	84.806	1	.000	.860	.833	.888
	Race	.950	.051	347.701	1	.000	2.585	2.339	2.856
	Health Insurance Type	.820	.043	363.172	1	.000	2.270	2.086	2.469
	Constant	-.644	.127	25.607	1	.000	.525		
Step 4[d]	SMEAN(HBA1C)	-.146	.016	78.821	1	.000	.864	.837	.892
	Race	.951	.051	346.981	1	.000	2.589	2.343	2.862
	Health Insurance Type	.836	.043	374.492	1	.000	2.308	2.120	2.512
	documented nonsmoker	.538	.069	61.003	1	.000	1.712	1.496	1.960
	Constant	-1.169	.144	65.445	1	.000	.311		
Step 5[e]	SMEAN(Age)	-.019	.002	61.181	1	.000	.981	.976	.986
	SMEAN(HBA1C)	-.163	.017	95.932	1	.000	.850	.822	.878
	Race	.986	.051	366.556	1	.000	2.680	2.422	2.964
	Health Insurance Type	.616	.051	144.142	1	.000	1.852	1.674	2.047
	documented nonsmoker	.595	.069	73.472	1	.000	1.812	1.582	2.076
	Constant	.165	.223	.548	1	.459	1.179		

a. Variable(s) entered on step 1: Race.
b. Variable(s) entered on step 2: Health Insurance Type.
c. Variable(s) entered on step 3: SMEAN(HBA1C).
d. Variable(s) entered on step 4: documented nonsmoker.
e. Variable(s) entered on step 5: SMEAN(Age).

For patients who have commercial health insurance, the adjusted odds ratio is $e^{0.836} = 1.849$. While adjusting for the other variables on the model, the predicted odds of using the PHR system is 2.308 times greater for those who have commercial health insurance than it is for those who do not. The 95% confidence interval for the odds of health, insurance type is, for patients who have commercial health insurance, the predicted odds of using the PHR system is 2.120 to 2.512 times greater than those who have another type of health insurance.

For patients who do not smoke, the adjusted odds ratio is $e^{0.538} = 1.496$. While adjusting for the other variables on the model, the predicted odds of using the PHR system is 1.496 times greater for

those who do not smoke than it is for those do not. The 95% confidence interval for the odds of nonsmoking status is, for patients who do not smoke, the predicted odds of using the PHR system is 1.496 to 1.960 times greater than those who do smoke.

Classification Table[a]

			Predicted		
			PHR User Group		Percentage Correct
	Observed		Nonuser	User	
Step 4	PHR User Group	Nonuser	3867	2843	57.6
		User	1347	2689	66.6
	Overall Percentage				61.0

a. The cut value is .350

This classification table in Figure 33 has the output for the training and testing datasets. The rate of correct classification by the logistic regression model is 61%. The overall risk estimate = (621+1417)/(1927+1368+621+1417) = 0.3821 = 1 – 0.618 = 1 – response rate (rate of misclassification) for this model. This is seen in the classification table, where the cut value was set to 0.35, close to the percentage of people who were recorded to have used the PHR system in the dataset, according to the pie-chart of PHR users above. The model did about the same job of predicting diabetes and no diabetes of the patients in the data set, as the specificity (57.6% = (3867/(3867+2843))*100%); the percentage that the model correctly predicted that patients do not use the PHR system) and sensitivity (66.6% = (2689/(1347+2689))*100%) the percentage that the model correctly predicted that patients did use the PHR system). These values of specificity and sensitivity are both are close and rather high (above 50%). So, the model predicts correctly which category patients fall in well.

Sources and Resources Used for Multiple Logistic Regression:

Dr. Fridline's Applied Analytics and Decision Trees course notes on logistic regression, Spring 2019.

http://www.biostathandbook.com/multiplelogistic.html

residuals: https://freakonometrics.hypotheses.org/8210

Replacing missing values SPSS:
https://www.ibm.com/support/knowledgecenter/ko/SSLVMB_24.0.0/spss/base/replace_missing_values_estimation_methods.html

The dataset represents data from the study by Tenforde et al. "The Association Between Personal Health Record Use and Diabetes Quality Measures". J Gen Intern Med 2012; 27: 420-24.