Problem Set 7

cdurso

12 March, 2019

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please complete the following tasks, in R where applicable. Please generate a solution document in R markdown and upload the rendered .doc, .docx, or .pdf document. You may add hand computations to a .doc or .docx if you prefer. In the rendered document, please show your code. That is, don't use "echo=FALSE".

In either case, your work should be based on the data's being in the same folder as the R files. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

1.a Let X and Y be independent random variables with χ^2 distributions with n and m degrees of freedom respectively. What is the distribution of $X + Y$?. (10 points)

$X+Y$ has a χ^2 distribution with $n + m$ degrees of freedom.

1.b. Let X and Y be independent random variables with χ^2 distributions with n and m degrees of freedom respectively. What is the distribution of X/Y?.(10 points)

The random variable W defined by $\frac{\frac{X}{n}}{\frac{Y}{m}}$ has an F-distribution with numerator degrees of freedom equal to n and numerator degrees of freedom equal to m, so $\frac{X}{Y} = \frac{m}{n}W$ is a constant multiple of an F-distribution with numerator degrees of freedom equal to n and numerator degrees of freedom equal to m

2. The data in "school_goals.csv" come from http://lib.stat.cmu.edu/DASL/Stories/Students'Goals.html . Students in grades 4-6 in selected schools in Michigan were asked the following question: What would you most like to do at school? a) make good grades, b) be good at sports, c) be popular. The results are recorded in "Goals". Demographic information about the students was also collected.

- 2.a. Please display a table of totals for each Goal for each gender (Gender), and a table of proportions for each Goal for gender. The "spread" function may be helpful in generating a two-way table with rows for gender and columns for priorities. (10 points)

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

```
## Warning: package 'tibble' was built under R version 3.4.4
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
## -- Conflicts ------------------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
dat<-read.csv("school_goals.csv")
dat.tab<-mutate(group_by(dat,Gender),gender.ct=n())
(dat.tab<-summarize(group_by(dat.tab,Gender,Goals),goal.ct=n(),goal.prop=n()/gender.ct[1]))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
## # A tibble: 6 x 4
## # Groups:   Gender [?]
##   Gender Goals   goal.ct goal.prop
##   <fct>  <fct>     <int>     <dbl>
## 1 boy    Grades      117     0.515
## 2 boy    Popular      50     0.220
## 3 boy    Sports       60     0.264
## 4 girl   Grades      130     0.518
## 5 girl   Popular      91     0.363
## 6 girl   Sports       30     0.120
```

```
(two_way_tidy<-dat.tab%>%select(Gender:goal.ct)%>%spread(key=Goals,value=goal.ct))
```

```
## # A tibble: 2 x 4
## # Groups:   Gender [2]
##   Gender Grades Popular Sports
## * <fct>   <int>   <int>  <int>
## 1 boy       117      50     60
## 2 girl      130      91     30
```

```
(two_way_tidy_prop<-dat.tab%>%select(Gender,Goals,goal.prop)%>%spread(key=Goals,
                                                                      value=goal.prop))
```

```
## # A tibble: 2 x 4
## # Groups:   Gender [2]
##   Gender Grades Popular Sports
## * <fct>   <dbl>   <dbl>  <dbl>
## 1 boy     0.515   0.220  0.264
## 2 girl    0.518   0.363  0.120
```

```
# Old school
(two_way<-table(dat$Gender,dat$Goals))
```

```
##
##        Grades Popular Sports
##   boy     117      50     60
##   girl    130      91     30
```

```
two_way/rowSums(two_way)
```

```
##
##           Grades   Popular    Sports
##   boy  0.5154185 0.2202643 0.2643172
##   girl 0.5179283 0.3625498 0.1195219
```

- 2.b. Please do a χ^2 test and a Fisher's Exact Test to determine if the goals are independent of Gender, and compare the results (10 points)

```
table(dat$Gender,dat$Goals)/rowSums(table(dat$Gender,dat$Goals))
```

```
##
##           Grades   Popular    Sports
##   boy  0.5154185 0.2202643 0.2643172
##   girl 0.5179283 0.3625498 0.1195219
```

```
fisher.test(dat$Goals,dat$Gender)
```

```
##
##  Fisher's Exact Test for Count Data
```

```
##
## data:  dat$Goals and dat$Gender
## p-value = 2.088e-05
## alternative hypothesis: two.sided
```

```
chisq.test(dat$Goals,dat$Gender)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dat$Goals and dat$Gender
## X-squared = 21.455, df = 2, p-value = 2.193e-05
```

Both tests perform correctly, and give strong reason to reject the null hypothesis that Goal is independent of Gender.

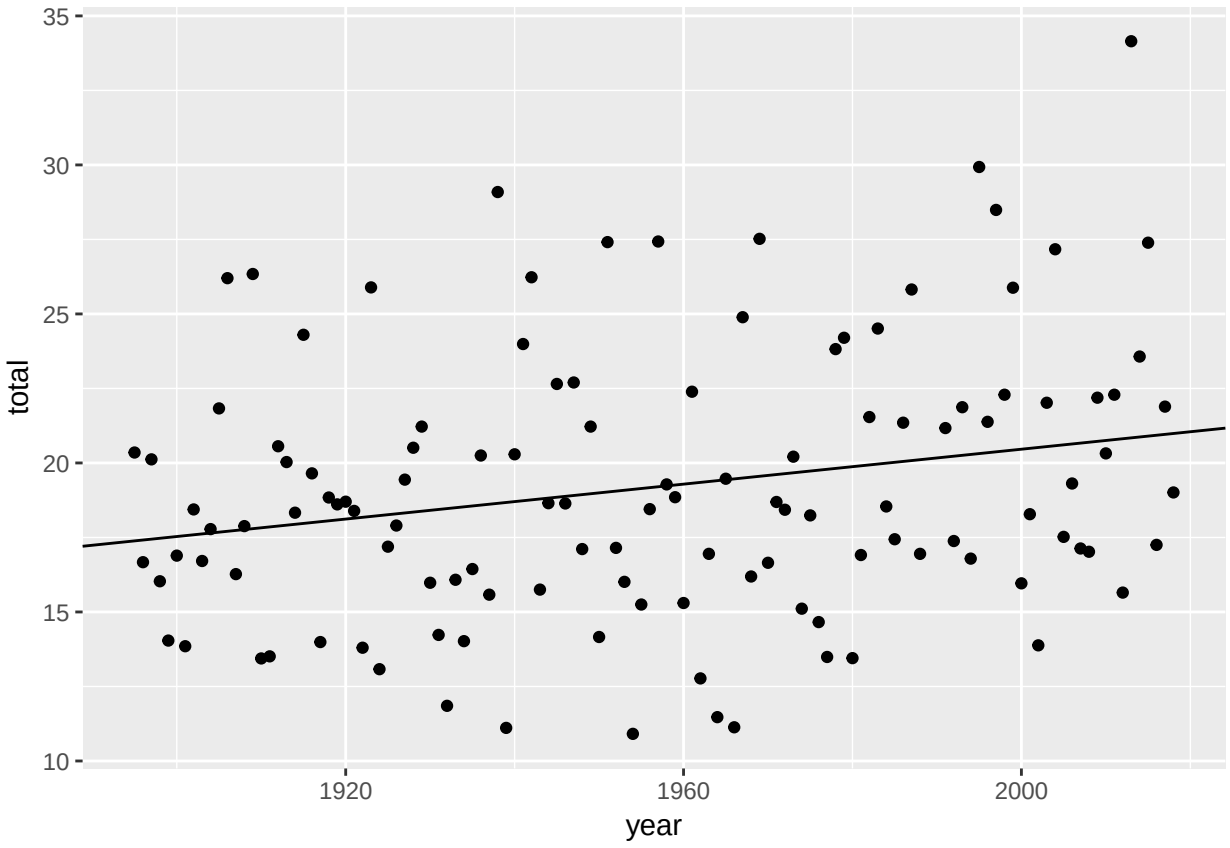
3.a. Returning to the preciptation data, use maximum likelihood regression to model total precipitation (total) as a linear function of year. Show the resulting coefficients and their p-values.

```
dat<-read.csv("dat_ok.csv")
m<-lm(total~year,data=dat)
summary(m)
```

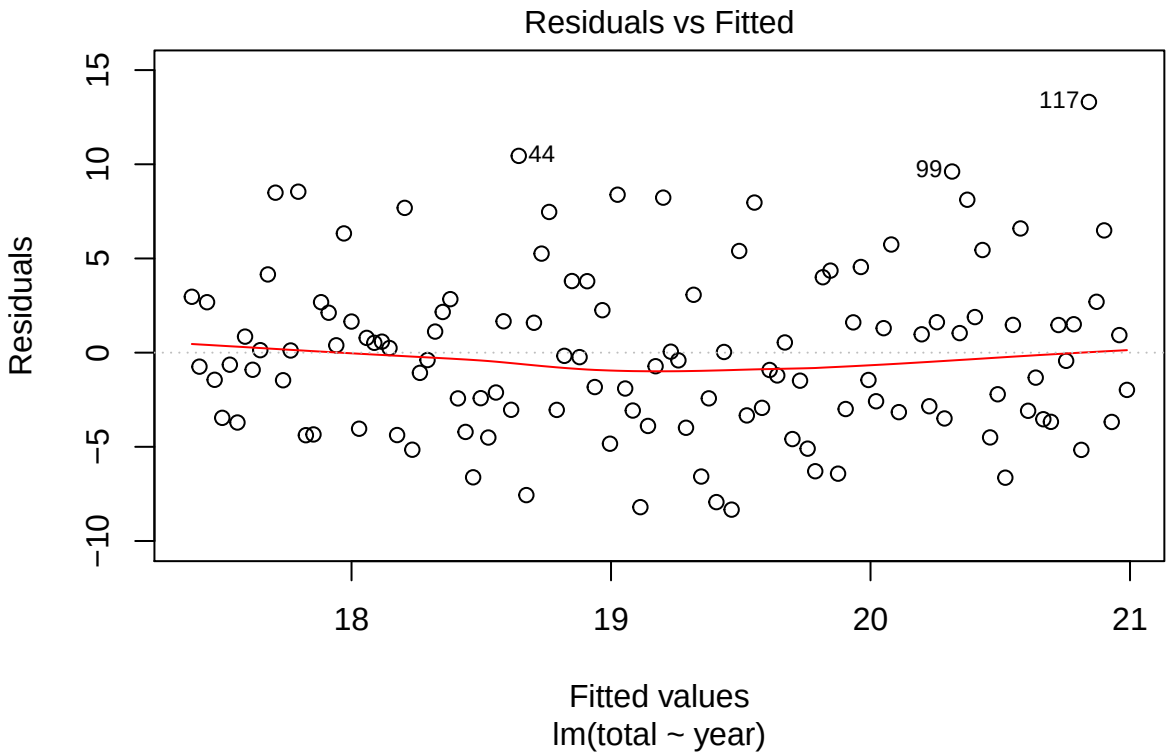
```
##
## Call:
## lm(formula = total ~ year, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3347 -3.1413 -0.4013  2.2296 13.3083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.13305   21.75633  -1.753  0.08220 .
## year          0.02930    0.01112   2.634  0.00954 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.402 on 120 degrees of freedom
## Multiple R-squared:  0.05467,    Adjusted R-squared:  0.04679
## F-statistic:  6.94 on 1 and 120 DF,  p-value: 0.009542
```

3.b. Is a linear model for the relationship between time and total appropriate? Provide any diagnostic plots you feel are relevant to addressing this question.

```
ggplot(dat,aes(x=year,y=total))+geom_point()+
  geom_abline(slope=m$coefficients[2], intercept=m$coefficients[1])
```



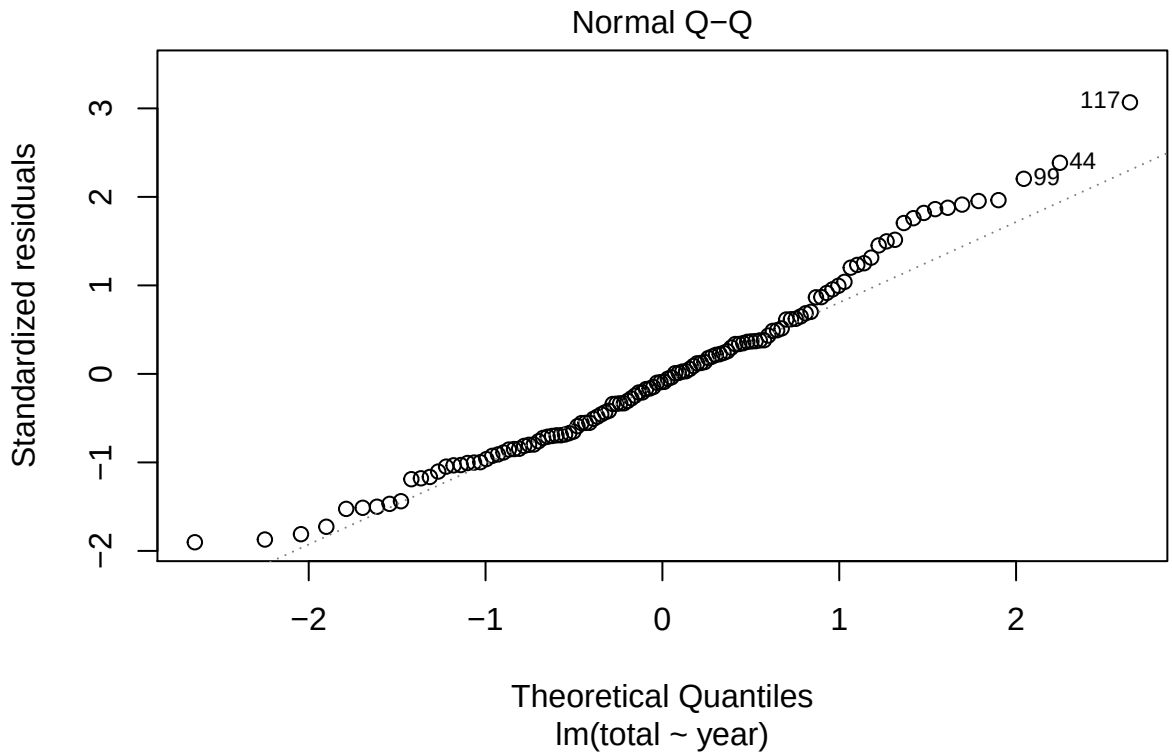

```
plot(m, which=1)
```



The scatterplot shows an upward trend, with no indication of non-linearity. The plot of residuals vs. fitted is consistent with linearity.

3.c. Is the assumption of independent,identically distributed Normal errors justifiable? Please show any relevant plots.

```
plot(m, which=2)
```




```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.4.4
```

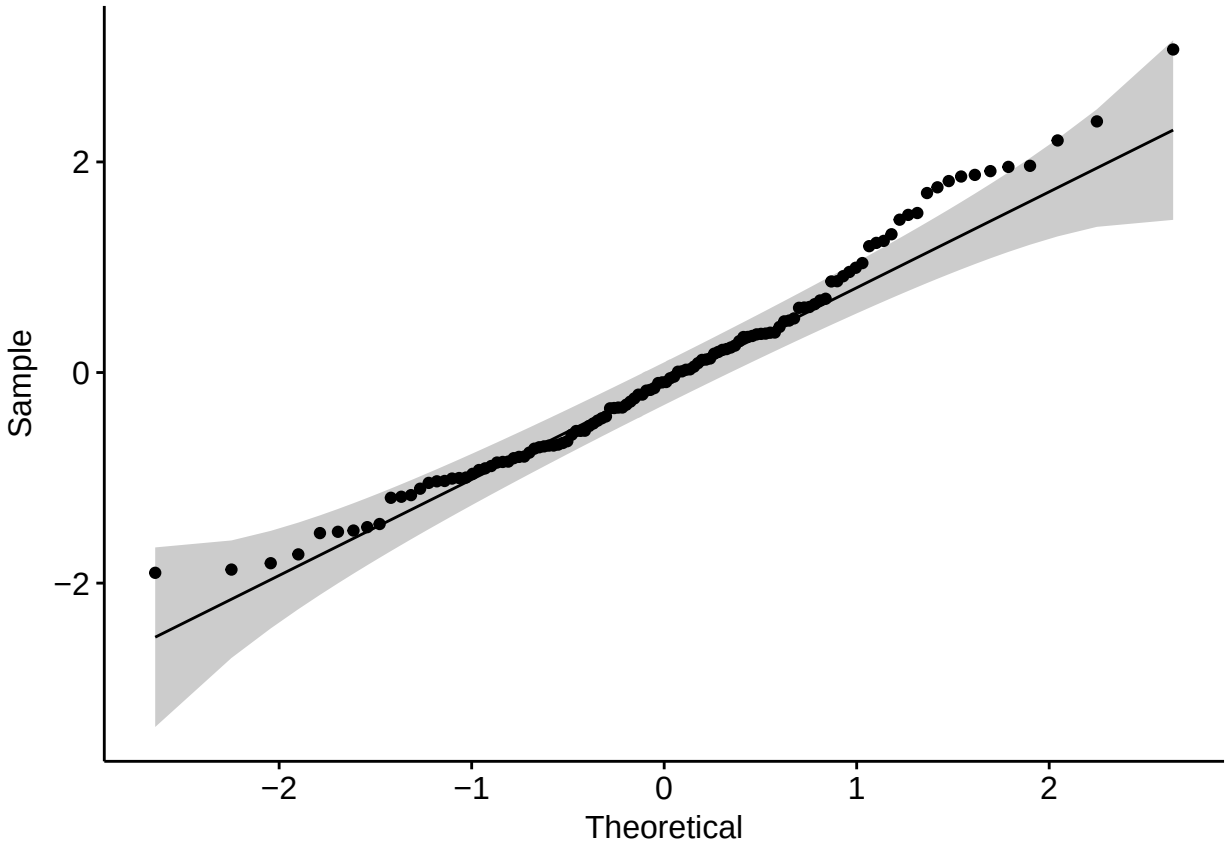
```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

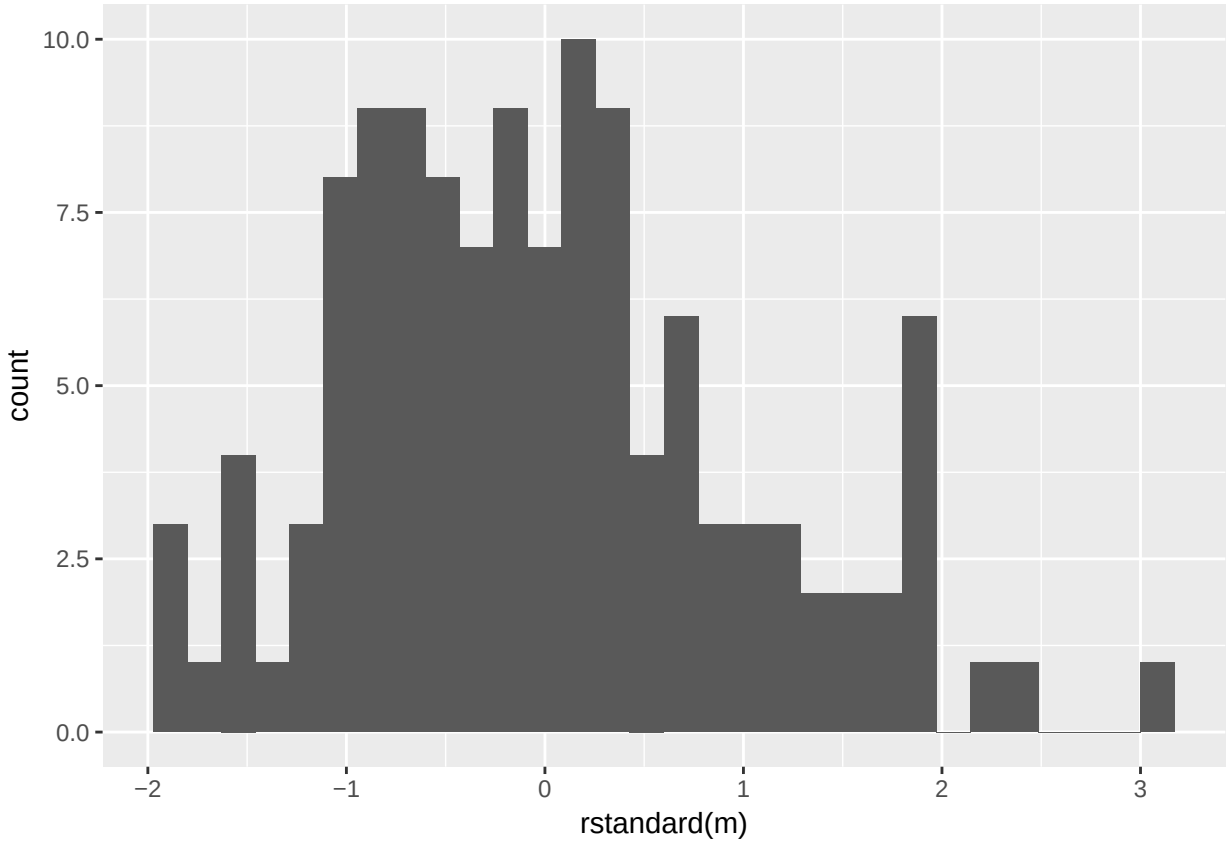
```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
ggqqplot(rstandard(m))
```




```
qplot(rstandard(m))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot of residuals vs. fitted is an even band around the horizontal line at 0, so the variance of the error is approximately constant over time. Independence is also reasonable. Normality isn't perfect, but is acceptable.