2015S

Fountain*, Crain

2015F

2015F1

2017SR1 X1,X2 linear regression

Find the best model for predicting Y based on X1 and X2. Y is the amount of profit that a company makes in a month. X1 is the number of months that the company has been in business. X2 is the amount spent on advertising.

Consider as predictors all possible linear and quadratic terms ($X1$, $X1^2$, $X2$, $X2^2$, and $X1X2$). Consider possible transformations of Y. Include all appropriate diagnostics. When you have found your "best" model, predict a new Y when X1 = 20 and X2 = $1,500, giving a 95% prediction interval. The data set, shown below, appears in "Profits.xlsx".

```
table_2015f1 <- readxl::read_xlsx("qe_lab/Profits_2015f.xlsx")
str(table_2015f1)
library(ggplot2)
ggplot(table_2015f1,aes(X2,Y, color=X1))+labs(x="advertising",y="profit",color="month")+geom_point()+theme_light()
```

```
model_2015f1_1 <- lm(Y^2~X1+X2,table_2015f1)
summary(model_2015f1_1)
anova(model_2015f1_1)
plot(model_2015f1_1)
```

```
model_2015f1_2 <- lm(Y^2~X1*X2,table_2015f1)
summary(model_2015f1_2)
anova(model_2015f1_2)
plot(model_2015f1_2)
```

```
sqrt(predict(model_2015f1_1, newdata=data.frame(X1 = 20 ,X2 =1500), interval="prediction", level=0.95 ))
sqrt(predict(model_2015f1_2, newdata=data.frame(X1 = 20 ,X2 =1500), interval="prediction", level=0.95 ))
```

2015F2

2018F2 5k1p Fractional Factorial Design

A replicated fractional factorial design is used to investigate the effect of five factors on the free height of leaf springs used in an automotive application. The factors are (A) furnace temperature, (B) heating time, (C) transfer time, (D) hold down time, and (E) quench oil temperature. There are 3 observations at each setting.

Write out the alias structure for this design. What is the resolution of this design? Analyze the data. What factors influence the mean free height? The data set appears in the file "Springs.xlsx".

```
table_2015f2 <- readxl::read_xlsx("qe_lab/Springs_2015f.xlsx")
library(tidyverse)
table_2015f2 <- gather(table_2015f2,'Height','...7','...8',key = "1",value = "height" )[,-6]
str(table_2015f2)
kableExtra::kable(table_2015f2)
```

```
library(devtools)
devtools::install_github("tidyverse/tidyr",force=T)
pivot_longer(table_2015f2,-A,-B,-C,-D,values_to = "Height")
```

I=ABCD;

A=BCD; B=ACD; C=ABD; D=ABC; E=ABCDE;

AB=CD; AC=BD; AD=BC; AE=BCDE; BE=ACDE; CE=ABDE; DE=ABCE;

ABE=CDE; BCE=ADE; BDE=ACE;

I=ABCD confounded

Resolution=IV

```
model_2015f2_1 <- aov(height~A*B*C*D*E, table_2015f2)
summary(model_2015f2_1)
library(daewr)
halfnorm(coef(model_2015f2_1)[2:16],alpha=0.05)
library(gghalfnorm)
gghalfnorm(x =coef(model_2015f2_1)[2:16],labs = names(coef(model_2015f2_1)[2:16]) , nlab = 5)+ ggplot2::theme_light
model_2015f2_2 <- lm(height~A+B*E+D, table_2015f2)
summary(model_2015f2_2)
anova(model_2015f2_2)
```

2016S

Fountain, Tableman*

2016S1

2017F1

Find the best model for predicting Y (weight) based on X1 (age), X2 (height), and X3 (indicator for male). Consider as predictors all possible linear and quadratic terms. Consider possible transformations of Y. Include all appropriate diagnostics. When you have found your "best" model, predict a new Y when X1 = 26, X2 = 70, and X3 = 1, giving a 95% prediction interval. The data set, shown below, appears in "RegressionSpr16.xlsx".

```
table_2016s1 <- readxl::read_xlsx("qe_lab/RegressionSpr16.xlsx")[-1,]
table_2016s1$weight <- round(as.numeric(table_2016s1$weight), 2)
table_2016s1$age <- as.factor(table_2016s1$age)
table_2016s1$height <- round(as.numeric(table_2016s1$height), 2)
table_2016s1$male <- factor(table_2016s1$male, labels=c("female","male"))
str(table_2016s1)
```

```
library(ggpubr)
ggline(table_2016s1,"height","weight",add = c("mean","jitter"),color = "age" )
ggline(table_2016s1,"height","weight",add = c("mean","jitter"),color = "male",shape = "male" )
```

```
library(GGally)
ggpairs(table_2016s1)
model_2016s1 <- lm(weight~height*male*age, table_2016s1)
olsrr::ols_step_both_aic(model_2016s1)
```

```
model_2016s1_1 <- lm((weight)~(height):male:age, table_2016s1)
summary(model_2016s1_1)
model_2016s1_2 <- lm(log(weight)~male:age+height:male:age, table_2016s1)
summary(model_2016s1_2)
anova(model_2016s1_2)
plot(model_2016s1_2)
```

2016S2

2017F2

A process engineer is testing the yield of a product manufactured on three specific machines. Each machine can be operated at fixed high and low power settings, although the actual settings differ from one machine to the next. Furthermore, a machine has three stations on which the product is formed, and these are the same for each machine. An experiment is conducted in which each machine is tested at both power settings, and three observations on yield are taken from each station. The runs are made in random order. Analyze this experiment. The data set, shown below, appears in "DesignSpr16.xlsx".

```
DesignSpr16 <- readxl::read_excel("qe_lab/DesignSpr16.xlsx")
library(tidyverse)
table_2016s2 <- gather(DesignSpr16[c(2:4,6:8),c(2:4,6:8,10:12)])
names(table_2016s2) <- c("machine","y")
table_2016s2 <- table_2016s2[c("y","machine")]
table_2016s2$machine <- as.factor(c(rep("machine1",18),rep("machine2",18),rep("machine3",18)) )
table_2016s2$station <- as.factor(rep(c(rep("station1",6),rep("station2",6),rep("station3",6)),3) )
table_2016s2$power <- as.factor(rep(c(rep("power1",3),rep("power2",3)),9) )
str(table_2016s2)
```

```
library(ggpubr)
ggline(table_2016s2,"machine","y",add = c("mean","jitter"),color = "station",shape = "station")
ggline(table_2016s2,"machine","y",add = c("mean","jitter"),color = "power",shape = "power")
```

```
model_2016s2 <- aov(y~machine*power*station, table_2016s2)
summary(model_2016s2)
# anova(model_2016s2)
```

```
library(lme4)
model_2016s2_1 <-lmer(y~machine*station+(1|machine:power)+(1|machine:power:station),table_2016s2)
summary(model_2016s2_1)$varcor
# anova(model_2016s2_1)
# pf(anova(model_2016s2_1)$'F value',df1=anova(model_2016s2_1)$'Df',df2=c(3,6,6), lower.tail = F)
confint(model_2016s2_1)
library(GAD)
table_2016s2$machine_f <- as.fixed(table_2016s2$machine)
table_2016s2$station_f <- as.fixed(table_2016s2$station)
table_2016s2$power_r <- as.random(table_2016s2$power)
model_2016s2_2 <- aov(y~machine_f*station_f+power_r%in%machine_f+power_r%in%machine_f*station_f, table_2016s2)
gad(model_2016s2_2)
```

```
plot(model_2016s2_2)
```

2016F

Jong Sung Kim*, Brad Crain

2016F1

A national insurance organization wanted to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. Data were collected for 1960, 1970, and 1980, but we will focus here on 1970. Using data from 1970, the organization wanted to construct a regression equation that relates statewide cigarette consumption (on a per capita basis) to various socioeconomic and demographic variables, and to determine whether these variables were useful in predicting the consumption of cigarettes. The variables chosen for study are given below. Age, x1: Median age of a person living in the state

Education, x2: Percentage of people over 25 years of age in a state that had completed high school

Income, x3: Per capita personal income for a state (in dollars)

Perblack, x4: Percentage of blacks living in a state

Perfem, x5: Percentage of females living in a state

Price, x6: Average price of a pack of cigarettes in a state (in cents)

Scig, y: Number of packs of cigarettes sold in a state on a per capita basis.

The data on these variables are stored in 8 columns in the same order as listed above; a two-letter alphabetic code is given first, however. The data are saved as "cigcons.xlsx"

Perform a complete regression analysis on these data; including checking of model assumptions and attempting appropriate remedies, if needed. The main objective of the analysis is to find the smallest number of variables that describes the state sale of cigarettes meaningfully and adequately. You might want to consider among others partial regression plot, interaction terms, outliers and influential cases analysis, Box-Cox transformation, and explanation of your final model.

```
table_2016f1 <- readxl::read_xlsx("qe_lab/cigcons.xlsx")
table_2016f1$State <- as.factor(table_2016f1$State)
str(table_2016f1)
```

```
library(GGally)
ggpairs(table_2016f1[,-1])
```

```
model_2016f1 <- lm(Scig~price*perfem*perblack*Income*Education*Age, table_2016f1)
ols_step_both_aic(model_2016f1)
ols_step_both_p(model_2016f1)
```

2016F2

An experiment is conducted to compare the water quality of three creeks in an area. Five water samples are selected from each creek. Each sample is divided into two parts, and the dissolved oxygen content is measured for each part. (Higher dissolved oxygen contents indicate higher water quality.) The results are given as follows:

Creek/Water Sample	1		2	3	4	5
1	5.2,	5.4	5.6, 5.7	5.4, 5.4	5.6, 5.5	5.8, 5.5
2	5.1,	5.3	5.1, 5.0	5.3, 5.2	5.0, 5.0	4.9, 5.1
3	5.9,	5.8	5.8, 5.8	5.7, 5.8	5.8, 5.9	5.9, 5.9

a. Write down an appropriate model with assumptions (including normality).

One-stage nested design

$y = \mu + \tau_i + \beta_{j(i)} + \varepsilon_{k(ij)}, i = 1,2,3; j = 1,2,3,4,5; k = 1,2$

b. Find the ANOVA table for the data.

c. Perform the F-test comparing the creeks using a .05 level.

d. Perform a Tukey multiple comparison on the creeks using a .05 level.

```
creek1 <- c(5.2, 5.4, 5.6, 5.7, 5.4, 5.4, 5.6, 5.5, 5.8, 5.5)
creek2 <- c(5.1, 5.3, 5.1, 5.0, 5.3, 5.2, 5.0, 5.0, 4.9, 5.1)
creek3 <- c(5.9, 5.8, 5.8, 5.8, 5.7, 5.8, 5.8, 5.9, 5.9, 5.9)
library(tidyverse)
table_2016f2 <- gather(data.frame(creek1,creek2,creek3),creek,oxygen)
table_2016f2$creek <- as.factor(table_2016f2$creek)
table_2016f2$sample <- as.factor(c(rep("sample1",2),rep("sample2",2),rep("sample3",2),rep("sample4",2),rep("sample5"
table_2016f2$rep <- as.factor(rep(c("rep1","rep2"),15))
str(table_2016f2)
```

```
library(ggpubr)
ggline(table_2016f2,"creek","oxygen", add = c("mean","jitter"),color = "sample",shape = "sample")
ggline(table_2016f2,"sample","oxygen", add = c("mean","jitter"),color = "creek",shape = "creek")
```

```
model_2016f2_1 <- lm(oxygen~creek/sample,table_2016f2)
anova(model_2016f2_1)
```

```
library(lme4)
model_2016f2_2 <- lmer(oxygen~creek+(1|creek:sample),table_2016f2)
summary(model_2016f2_2)
# anova(model_2016f2_2)
# pf(anova(model_2016f2_2)$'F value',df1=anova(model_2016f2_2)$'Df',df2=12, lower.tail = F)
confint(model_2016f2_2)
```

```
library(GAD)
table_2016f2$creek_f <- as.fixed(table_2016f2$creek)
table_2016f2$sample_r <- as.random(table_2016f2$sample)
model_2016f2_3 <- aov(oxygen~creek_f+sample_r%in%creek_f, table_2016f2)
gad(model_2016f2_3)
plot(model_2016f2_3)
```

```
library(emmeans)
library(kableExtra)
kable(test(lsmeans(model_2016f2_2,~creek,adjust=c("tukey"))))
kable(pairs(lsmeans(model_2016f2_2,~creek,adjust=c("tukey"))))
kable(TukeyHSD(model_2016f2_3,conf.level=0.95)$creek_f)
```

```
# for reference
cre_sam <- pairs(lsmeans(model_2016f2_1,~creek|sample))
sam_cre <- pairs(lsmeans(model_2016f2_1,~sample|creek))
kable(test(rbind(cre_sam,sam_cre),adjust="tukey"),format="latex")%>%kable_styling("condensed",full_width=F,font_size
cre_sam <- pairs(lsmeans(model_2016f2_3,~creek_f|sample_r))
sam_cre <- pairs(lsmeans(model_2016f2_3,~sample_r|creek_f))
kable(test(rbind(cre_sam,sam_cre),adjust="tukey"),format="latex")%>%kable_styling("condensed",full_width=F,font_size
```

2017S

Brad Crain, Jong Sung Kim*

2017SR1

2015F1

Find the best model for predicting Y based on X1 and X2. Y is the amount of profit that a company makes in a month. X1 is the number of months that the company has been in business. X2 is the amount spent on advertising. Consider as predictors all possible linear and quadratic terms ($X1, X1^2, X2, X2^2$, and $X1X2$). Consider possible transformations of Y. Include all appropriate diagnostics. When you have found your "best" model, predict a new Y when X1 = 20 and X2 = /$1,500, giving a 95% prediction interval. The data set, shown below, appears in "Profits.xlsx".

```
table_2017sr1 <- readxl::read_xlsx("qe_lab/Profits_2017s.xlsx")
# table_2017sr1$X1 <- as.factor(table_2017sr1$X1)
str(table_2017sr1)
summary(table_2017sr1)
```

```
library(ggplot2)
ggplot(table_2017sr1, aes(X2,Y,color=X1))+geom_point()+theme_light()
ggplot(table_2017sr1, aes(X1,Y,color=X2))+geom_point()+theme_light()
```

```
model_2017sr1 <- lm(Y^2~X1+X2, table_2017sr1)
# car::vif(model_2017sr1)
summary(model_2017sr1)
anova(model_2017sr1)
```

```
plot(model_2017sr1,c(1,3,5))
residual_2017sr1 <- rstudent(model_2017sr1)
qqnorm(residual_2017sr1)
qqline(residual_2017sr1)
olsrr::ols_plot_resid_hist(model_2017sr1)
hist(residual_2017sr1)
```

```
sqrt(predict(model_2017sr1,newdata = data.frame(X1=20,X2=1500),interval = "prediction", level = 0.95))
```

2017SD1

Review the data provided in 'NBalance.xlsx'. Note, there were nine distinct treatments [Feed Rations] and three distinct animals. An experimental design was used to examine the means differences in the Nitrogen balance in ruminants. Provide the following in your answer

1. Which design was used, include the required parameters of the experimental design $[t; b; k; r; \lambda]$

BIBD
$y = \mu + \tau_i + \beta_j + \varepsilon_{ij} + \varepsilon$ Treatment(Rations)$a = 9$,

Replication$r = 3$,

Block(animals)$b = 3$,

Block size $k = 9$
$\lambda = 3$

2. An appropriate ANOVA
3. A TukeyHSD analysis of the proper means differences
4. Conclusions on the impact of Feed Rations on Nitrogen Balance in Ruminants

Source: J.L. Gill (1978), Design and analysis of experiments in the animal and medical sciences, Vol2. Ames, Iowa: Iowa State University Press

```
library(ggpubr)
ggline(table_2017sd1, "Animal","Nitrogen", add = c("mean","jitter"),color = "Ration",shape = "Ration")
ggline(table_2017sd1, "Ration","Nitrogen", add = c("mean","jitter"),color = "Animal",shape = "Animal")
```

```
model_2017sd1 <- aov(Nitrogen~Animal+Ration, table_2017sd1)
summary(model_2017sd1)
anova(model_2017sd1)
TukeyHSD(model_2017sd1,conf.level = 0.95)
```

2017F

Robert Fountain*, Daniel Taylor-Rodriguez

2017F1

2016S1
Find the best model for predicting Y (weight) based on X1 (age), X2 (height), and X3 (indicator for male). Consider as predictors all possible linear and quadratic terms. Consider possible transformations of Y. Include all appropriate diagnostics. When you have found your "best" model, predict a new Y when X1 = 26, X2 = 70, and X3 = 1, giving a 95% prediction interval. The data set, shown below, appears in "RegressionFall17.xlsx".

```
table_2017f1 <- readxl::read_xlsx("qe_lab/RegressionFall17.xlsx")[-1,]
table_2017f1$weight <- round(as.numeric(table_2017f1$weight),2)
table_2017f1$age <- as.numeric(table_2017f1$age)
table_2017f1$height <- round(as.numeric(table_2017f1$height),2)
table_2017f1$male <- factor(table_2017f1$male, labels = c("female", "male"))
str(table_2017f1)
```

```
library(ggplot2)
ggplot(table_2017f1, aes(height,weight,color=age,shape=male))+geom_point()+theme_light()
library(ggpubr)
ggline(table_2017f1,"height","weight",add=c("mean","jetter"),color="age")
ggline(table_2017f1,"height","weight",add=c("mean","jetter"),color="male",shape = "male")
```

```
model_2017f1 <- lm(weight~height*age*male,table_2017f1)
library(olsrr)
ols_step_both_aic(model_2017f1)
ols_step_both_p(model_2017f1)
model_2017f1_1 <- lm(weight~height+age:male,table_2017f1)
model_2017f1_2 <- lm(log(weight)~height+age:male,table_2017f1)
car::vif(model_2017f1_2)
summary(model_2017f1_2)
anova(model_2017f1_2)
ols_regress(model_2017f1_2)
```

```
plot(model_2017f1_2)
```

```
predict(model_2017f1_2, newdata=data.frame(age= 26, height= 70, male= "male"),interval = "prediction",level = 0.95)
```

2017F2

A process engineer is testing the yield of a product manufactured on three specific machines. Each machine can be operated at fixed high and low power settings, although the actual settings differ from one machine to the next. Furthermore, a machine has three stations on which the product is formed, and these are the same for each machine. An experiment is conducted in which each machine is tested at both power settings, and three observations on yield are taken from each station. The runs are made in random order. Analyze this experiment. The data set, shown below, appears in "DesignFall17.xlsx".

```
DesignFall17 <- readxl::read_excel("qe_lab/DesignFall17.xlsx")
library(tidyverse)
table_2017f2 <- gather(DesignFall17[c(2:4,6:8),c(2:4,6:8,10:12)])
names(table_2017f2)<- c("machine","y")
table_2017f2<- table_2017f2[c("y","machine")]
table_2017f2$machine <- as.factor(c(rep("machine1",18),rep("machine2",18),rep("machine3",18)))
table_2017f2$station <- as.factor(rep(c(rep("station1",6),rep("station2",6),rep("station3",6)),3))
table_2017f2$power <- as.factor(rep(c(rep("power1",3),rep("power2",3)),9))
str(table_2017f2)
```

```
model_2017f2 <- lm(y~power*station*machine, table_2017f2)
summary(model_2017f2)
anova(model_2017f2)
```

```
library(lme4)
model_2017f2_1 <-lmer(y~machine*station+(1|machine:power)+(1|machine:power:station),table_2017f2)
summary(model_2017f2_1)$varcor
# anova(model_2016s2_1)
# pf(anova(model_2017f2_1)$'F value',df1=anova(model_2017f2_1)$'Df',df2=c(3,6,6), lower.tail = F)
pander::pander(confint(model_2017f2_1)[1:4,1:2])
library(GAD)
table_2017f2$machine_f <- as.fixed(table_2017f2$machine)
table_2017f2$station_f <- as.fixed(table_2017f2$station)
table_2017f2$power_r <- as.random(table_2017f2$power)
model_2017f2_2 <- aov(y~machine_f*station_f+power_r%in%machine_f+power_r%in%machine_f*station_f, table_2017f2)
gad(model_2017f2_2)
```

2018S

Robert Fountain*, Daniel Taylor-Rodriguez

2018S1

The data for this problem was obtained from research relating children smoking to pulmonary function. Today it is well established that smoking cigarettes is a very unhealthy habit, especially for children; however, this was not well-known in the past. This data

corresponds to one of the first studies of the effects of smoking on pulmonary (i.e., lung) function, an observational study of 654 youths aged 3 to 19. The variables in the study are displayed in Table 1 below. The outcome variable is volume, which measures the liters of air exhaled by the child in the first second of a forced breath. Some evidence in the literature suggests that children under age 6 may not understand the instructions of the breath exhalation test, so that the quality of volume measurements for those children is suspect. We are interested in the relationship between smoking, gender and the volume of air exhaled. Smoking is expected to impair pulmonary function (i.e., decrease volume).

Find the best model to predict volume considering as predictors all possible linear, quadratic and pairwise interaction terms. Additionally, consider possible transformations of the response (i.e., volume), and include all relevant diagnostic measures. Once you select the best model, write down and test the hypothesis to determine if the volume is influenced by the smoking status in terms of your best model's parameters. Using this same model, predict the volume for a 16-yearold male smoker who is 61 inches high, and provide a 95% prediction interval. A description of the variables is found in the table below, and the data is included in the file Problem1_ChildSmoking.xlsx.

Variable Name and Description

age: age of child in years

volume: volume of air in exhaled breath in liters

height: height of child in inches

male=1 if child is male, and =0 otherwise

smoker=1 if child reports that he or she smokes cigarettes regularly, and =0 otherwise

```
table_2018s1 <- readxl::read_xlsx("qe_lab/Problem1_ChildSmoking.xlsx")
table_2018s1_above6 <- table_2018s1[which(table_2018s1$age>5),]
table_2018s1_above6$age <- factor(table_2018s1_above6$age)
table_2018s1_above6$male <- factor(table_2018s1_above6$male, labels = c("female","male"))
table_2018s1_above6$smoker <- factor(table_2018s1_above6$smoker, labels = c("not regu","regularly"))
str(table_2018s1)
str(table_2018s1_above6)
summary(table_2018s1$height)
```

```
model_2018s1 <- lm(volume~height*age*male*smoker,table_2018s1_above6)
ols_step_both_aic(model_2018s1)
ols_step_both_p(model_2018s1)
```

```
model_2018s1_2 <- lm(log(volume)~log(height):age:male+smoker,table_2018s1_above6)
summary(model_2018s1_2)
anova(model_2018s1_2)
library(olsrr)
ols_regress(model_2018s1_2)
plot(model_2018s1_2)
```

$y = \mu + \beta_1 \ln(H) * Age * Male + \beta_2 Smoker + \varepsilon$

$H_0 : \beta_2 = 0, H_1 : \beta_2 \neq 0$

```
predict(model_2018s1_2, newdata =data.frame(age="16",male="male",smoker="regularly",height=61), interval = "predicti
```

2018S2

[RCBD]

An experiment is conducted to assess the effect of shipping and storage on the acceptability of avocados. Three shipping methods (labeled 1, 2 and 3) and two storage methods (labeled 1 and 2) were considered. Each combination of shipping x storage was applied to a group of four crates. Additionally, three different shipments were made. The experiment's configuration is shown below. Analyze this experiment. The data set can be found in the file Problem2_Avocado.xlsx.

```
library(emmeans)
Blk_Stor <- pairs(lsmeans(model_2018s2,~Block|Storage))
Stor_Blk <- pairs(lsmeans(model_2018s2,~Storage|Block))
Blk_Ship <- pairs(lsmeans(model_2018s2,~Block|Shipping))
Ship_Blk <- pairs(lsmeans(model_2018s2,~Shipping|Block))
Stor_Ship <- pairs(lsmeans(model_2018s2,~Storage|Shipping))
Ship_Stor <- pairs(lsmeans(model_2018s2,~Shipping|Storage))
library(kableExtra)
kable(test(rbind(Blk_Stor,Stor_Blk),adjust="tukey"),format="latex")%>%kable_styling("condensed",full_width=F,font_si
kable(test(rbind(Blk_Ship,Ship_Blk),adjust="tukey"),format="latex")%>%kable_styling("condensed",full_width=F,font_si
kable(test(rbind(Stor_Ship,Ship_Stor),adjust="tukey"),format="latex")%>%kable_styling("condensed",full_width=F,font_
```

2018F

Robert Fountain*, Daniel Taylor-Rodriguez

2018F1

2015F1 [2017S1][]

Find the best model for predicting Y based on X1 and X2. Y is the amount of profit that a company makes in a month. X1 is the number of months that the company has been in business. X2 is the amount spent on advertising. Consider as predictors all possible linear and quadratic terms ($X1, X1^2, X2, X2^2$, and $X1X2$). Consider possible transformations of Y. Include all appropriate diagnostics. When you have found your "best" model, predict a new Y when X1 = 20 and X2 = $1,900, giving a 95% prediction interval. The data set, shown below, appears in "Profits.xlsx".

```
model_2018f1_1 <- lm(Y^2~X1/X2, table_2018f1)
summary(model_2018f1_1)
# library(olsrr)
# ols_regress(model_2018f1_1)
# car::Anova(model_2018f1_1)
# car::vif(model_2018f1_1)
anova(model_2018f1_1)
plot(model_2018f1_1)
```

```
model_2018f1_2 <- lm(-log(Y)~X1/X2, table_2018f1)
summary(model_2018f1_2)
anova(model_2018f1_2)
plot(model_2018f1_2)
```

```
model_2018f1_3 <- lm(Y^(-0.01)~X1/X2, table_2018f1)
summary(model_2018f1_3)
anova(model_2018f1_3)
plot(model_2018f1_3)
```

```
predict(model_2018f1_1, newdata = data.frame(X1=20,X2=1900), interval = "prediction", level=0.95)
```

2018F2

2015F2 [7.4] [8.E.10]

A replicated fractional factorial design is used to investigate the effect of four factors on the free height of leaf springs used in an automotive application. The factors are (A) furnace temperature, (B) heating time, (C) transfer time, and (D) hold down time. There are 3 observations at each setting.

Write out the alias structure for this design. What is the resolution of this design?

I=ABCD, AB=CD, AC=BD, BC=AD; A=BCD, B=ACD, C=ABD, D=ABC; III

Analyze the data. What factors influence the mean free height? The data set appears in the file "Springs.xlsx".

A, B

```
table_2018f2 <- readxl::read_xlsx("qe_lab/Springs_2018f.xlsx")
table_2018f2 <- table_2018f2[order(table_2018f2$D ,table_2018f2$C ,table_2018f2$B,table_2018f2$A),]
str(table_2018f2)
kableExtra::kable(table_2018f2)
```

```
model_2018f2_1 <- aov(Heights~A*B*C*D, table_2018f2)
summary(model_2018f2_1)
library(daewr)
halfnorm(coef(model_2018f2_1)[2:8],alpha=0.4)
library(gghalfnorm)
gghalfnorm(x =coef(model_2018f2_1)[2:8],labs = names(coef(model_2018f2_1)[2:8]) , nlab = 4)+ ggplot2::theme_light()
model_2018f2_2 <- lm(Heights~A+B, table_2018f2)
summary(model_2018f2_2)
```

2019S

Robert Fountain*, Daniel Taylor-Rodriguez

Instructions:

1. Two 8.5" x 11" pages of notes (front and back) are allowed.
2. Perform the statistical analysis in your software of preference for the two problems below. The data sets for each problem are on the flash drive provided. Create a word or pdf document with your findings. Save the document to the flash drive provided with your name as the file name. You may use scratch paper during the exam, but everything you want considered for grading must be included in your document. Additionally, you must copy and paste the code used for the analysis at the end of the word/pdf document you submit.

3. For each question discuss all relevant aspects of your analysis (exploratory and modeling) supporting them with graphical and numerical summaries that are important for communicating results. It should also include a discussion of diagnostics and model adequacy, and rationale for any transformations or other key modeling decisions. The report should include interpretations of the output, written so that a statistically literate person can understand and apply the findings in each case.

2019S1

[4.2.1 PRESS residuals]

The goal of this exercise is to find the best model for predicting (out-of-sample) Y based on the continuous variables x1, x2, x3, and on the binary variables A and B. The data set is in the dataset "ModelBuildingData.xlsx". Consider possible transformations of Y, and for the linear predictor consider 2-way interactions and quadratic terms. Include all appropriate diagnostics, and make any necessary adjustments to the data so model assumptions are met.

Use only the first 250 observations for model training model (i.e., selection, fitting and diagnostics). With your top model, obtain predictions for all 250 remaining observations (the hold-out samples), and their corresponding 95% predictive intervals. Finally, calculate and interpret (in term of the model predictive ability) the Prediction Root Mean Square Error (PRMSE), as follows:

```
table_2019s1 <- readxl::read_xlsx("qe_lab/ModelBuildingData.xlsx")
str(table_2019s1)
dplyr::glimpse(table_2019s1)
table_2019s1_250 <- table_2019s1[1:250,]
table_2019s1_500 <- table_2019s1[251:500,]
str(table_2019s1_250)
str(table_2019s1_500)
```

```
model_2019s1 <- lm(log(y)~ x1*x2*x3*A*B,table_2019s1_250)
car::vif(model_2019s1)
summary(model_2019s1)
library(olsrr)
# ols_plot_diagnostics(model_2019s1_1)
ols_step_both_aic(model_2019s1)
```

```
model_2019s1_3 <- lm(table_2019s1_500,formula=log(y)~ x2+A+B)
summary(model_2019s1_3)
ols_regress(log(y)~ x2+A+B, data = table_2019s1_500)
library(Metrics)
Metrics::rmse(table_2019s1_500$y,exp(predict(model_2019s1_2,table_2019s1_500)))
ols_press(model_2019s1_3)
MPV::PRESS(model_2019s1_3)
sum((residuals(model_2019s1_3)/(1 - lm.influence(model_2019s1_3)$hat))^2)
ols_pred_rsq(model_2019s1_3)
# str(model_2019s1_3)
# From 564-lab caculate prediction power
deviation <- table_2019s1_500$y-mean(table_2019s1_500$y)
SST <- deviation%*%deviation
1-(MPV::PRESS(model_2019s1_3)/SST)
# by definition PRESS
sum((table_2019s1_500$y-exp(model_2019s1_2$fit))^2)
sum((table_2019s1_500$y-exp(predict(model_2019s1_2,table_2019s1_500)))^2)
# one method of RMSE
sqrt(mean(model_2019s1_3$residuals^2))
```

```
# remove outlier
table_2019s1_250[c(189,219,249),]
table_2019s1_250_noouter <- table_2019s1_250[-c(189,219,249), ]
table_2019s1_250_noouter <- table_2019s1_250[-c(113,189,219,249), ]
model_2019s1_noouter <- lm(y ~ sqrt(!is.na(x1))+x2+x3+A+B, data = table_2019s1_250_noouter)
summary(model_2019s1_noouter)
plot(model_2019s1_noouter)
```

a. calculate for each observation the square of the prediction errors,

b. obtain the square root of the average of all squared prediction errors.

https://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables

2019S2

[14.4] [566-fe-4] [Example 8.4]

An experiment was conducted to compare 4 wool fiber treatments (Trt) at 7 dry cycle revolutions (Rev) over 4 experimental runs (Run) (i.e., the blocks). The outcome measured from this experiment was the top shrinkage (Shrink) of the fiber. A restriction on the randomization: within each experimental run (blocks), wool fiber treatments were randomized to whole plots, and within each whole plot, measurements were obtained for all of 7 dry cycle revolutions (split plot treatments). In other words, the experiment was set as a **split-plot** design with:

a. whole plot (wool fiber treatment) treatments: untreated, alcoholic potash 15 Sec, alcoholic potash 4Min, and alcoholic potash 15Min;

b. subplot treatments: dry cycle revolutions (200 to 1400 by 200); and

c. blocks: 4 experimental runs (possibly different days).

Do a full analysis and report your findings for the experiment above (data in "Wool-Shrink.xlsx"), using a split plot design where both Trt and Rev are treated as categorical variables.

```
table_2019s2 <- readxl::read_xlsx("~/qushen26/stat2019_website/static/stat566/qe_lab/WoolShrink.xlsx")
table_2019s2$Run <- factor(table_2019s2$Run,labels=c("Day1","Day2","Day3","Day4"))
table_2019s2$Trt <- factor(table_2019s2$Trt,labels=c("untreated","15 Sec","4Min","15Min"))
table_2019s2$Rev <- as.factor(table_2019s2$Rev)
str(table_2019s2)
```

The above plots show that: There is not much difference in the average shrink from different days. The average shrink are lower when the treatment is longer. The average shrink are higher when the revolutions are faster.

The tables show the same thing with the numerical summaries for each factor level and their combinations.

```
library(GAD)
table_2019s2$Run_r <- as.random(table_2019s2$Run)
table_2019s2$Trt_f <- as.fixed(table_2019s2$Trt)
table_2019s2$Rev_f <- as.fixed(table_2019s2$Rev)
model_2019s2_1 <- aov(formula = Shrink ~ Run_r+Trt_f + Trt_f%in%Run_r+ Rev_f%in%Run_r + Rev_f + Trt_f:Rev_f, data=ta
pander::pander(gad(model_2019s2_1))
```

The results show all the main effects and the interaction effect of Runs and Recolutions are significant at 0.05 significance level (P-value=0.5082).

```
library("lme4")
model_2019s2_2 <- lmer(formula = Shrink ~ (1|Run) + Trt + (1|Run:Trt) + Rev + (1|Run:Rev) + Trt:Rev, data=table_201
summary(model_2019s2_2)$varcor
pander::pander(confint(model_2019s2_2)[1:4,1:2])
```

The results of variance components show the variance of interaction term of Runs and revolutions is negligible and hence dropping interaction term of them.

```
model_2019s2_3 <- aov(formula = Shrink ~ Run_r+Trt_f + Trt_f%in%Run_r+ Rev_f + Trt_f:Rev_f, data=table_2019s2)
pander::pander(gad(model_2019s2_3))
model_2019s2_4<- lmer(formula = Shrink ~ (1|Run)+Trt+Rev+(1|Run:Trt)+Rev*Trt, data=table_2019s2, REML = TRUE)
```

The ANOVA table of new model shows that the interaction effects are significant. This means that the effects of day v.s.revolutions and treatment v.s.revolutions on the shrink are not independent. Hence, the simple effects must be tested.

When the day2, the mean shrinks between the 15-Sec and 4-Min treatment don't have significant difference. For all the rest of days, the mean shrinks are significantly different between any different treatment.

The changes of days for a given treatment don't give consistent results.

For untreated cases, the mean shrinks are not significantly different between 1200 and 1400 revolutions. For all the rest of treatments, the mean shrinks are significantly different between any different revolutions.

For a given revolution, 15-Sec and 4-Min treatment don't have significant differece on the mean shrinks.

- Conclusion

Choosing a higher revolution for a given treatment can get a larger shrink.

In most of the cases, longter alcoholic potash have less shrink. This effect will be more significant when higher revolution.

- Model Adequacy Checking

In the plots of residuals versus predicted value of shrink, there is no significant pattern on this plot. Therefore, the fitted model is good enough to describe the relationship between the mean value of shrink and the days, revolutions, and treatment.

The residuals in this plot are almost symmetrically distributed about zero and hence zero mean assumption is not violated. Further, the vertical deviation of the residuals from zero is about same for each predicted value and hence the constant variance assumption is not violated.

The points are along the straight line in the normal qq plot shown at bottom left and the histogram of residuals shown at the top right is about normal. These plots show no violation of normal distribution assumption of residuals.

2019F

2019F1

2019F2

2020S1

In this exercise you will analyze data from the current COVID-19 pandemic in European countries. Specifically, you are provided with the cumulative number of confirmed cases due to COVID-19 reported on April 23rd 2020 for 50 european countries. This data is in the column confirmed of the file COVID-19.csv. For example, in France, 158,303 cases of COVID-19 were reported from the beginning of the pandemic up to April 23rd 2020. You are also provided with a measurement of the speed at which the pandemic is spreading in each of these countries: the doubling number (labelled confirmed.double in the data set). It is the approximate number of days it took for the confirmed cases to double. For example in France, on April 7th 2020, the number of cumulative confirmed cases was 79,163 (the closest report to 158,303/2 = 79151.5). Thus the doubling number is 16, which is the number of days from April 7th to April 24th. Note that a small doubling number is associated with a fast spreading of the pandemic while a large doubling number is associated with a slow spread of the COVID-19 pandemic. Note also that even if the doubling number is an integer, you will analyze it as if it was a continuous variable.

Your task is to analyze the speed of the spread of the pandemic in each European country, measured by the doubling number, as a function of the demographic, social, and economic variables for each country, available in the file COVID-19.csv and described in the data dictionary. Note that the table contains a few missing values indicated by NA. Please explain how you dealt with them and what you rationale was.

The NA records are concentrated in the countries with top 10 population.

The distribution of each variables: X1, X2, X3, X4, X6 show a exponential distribution. Log transform may help to find some linear relationship.

Linear model with full variables shows X5, X6 have significant effects on the responsor Y, confirmed.double.

```
library(GGally)
ggpairs(table_2020s1,aes(alpha=0.3))+theme_light()
# plot(table_2020s1$population,table_2020s1$land_area_skm,log = "xy")
# plot(table_2020s1$pop_density ,table_2020s1$pop_largest_city,log = "xy") #
```

The 9 missing values for X4 and 2 for X5.

Using data imputation by 'mice' package.

---

pmm any Predictive mean matching midastouch any Weighted predictive mean matching sample any Random sample from observed values cart any Classification and regression trees rf any Random forest imputations mean numeric Unconditional mean imputation norm numeric Bayesian linear regression norm.nob numeric Linear regression ignoring model error norm.boot numeric Linear regression using bootstrap norm.predict numeric Linear regression, predicted values quadratic numeric Imputation of quadratic terms ri numeric Random indicator for nonignorable data logreg binary Logistic regression logreg.boot binary Logistic regression with bootstrap polr ordered Proportional odds model polyreg unordered Polytomous logistic regression lda unordered Linear discriminant analysis 2l.norm numeric Level-1 normal heteroscedastic 2l.lmer numeric Level-1 normal homoscedastic, lmer 2l.pan numeric Level-1 normal homoscedastic, pan 2l.bin binary Level-1 logistic, glmer 2lonly.mean numeric Level-2 class mean 2lonly.norm numeric Level-2 class normal 2lonly.pmm any Level-2 class predictive mean matching

```
# Find the missing values (2 ways)
# summary(table_2020s1)
sapply(table_2020s1,function(x)sum(is.na(x)))
# Imputing Missing Data
library(mice)
md.pattern(table_2020s1,rotate.names=T)
imputed <- mice(table_2020s1,m=6,maxit=30,seed=500,method=c("","","","","cart","mean",""))#
# inspect quality of imputations
stripplot(imputed, pop_largest_city, pch = 19, xlab = "Imputation number")
stripplot(imputed, life_expectancy, pch = 19, xlab = "Imputation number")
stripplot(imputed, pop_largest_city+life_expectancy~confirmed.double, cex=c(1,2), pch=c(1,20),jitter=FALSE,layout=c
table_2020s1_imputed <- complete(imputed)
imputed$imp$pop_largest_city
```

```
# same result
mean(table_2020s1$life_expectancy,na.rm=T)
imputed$imp$life_expectancy
add2life_expectancy <- table_2020s1
add2life_expectancy[c(2,33),6] <- 78.46
# need fill the NA in X5 first
fit_x5<- lm(pop_largest_city~population+land_area_skm+pop_density,add2life_expectancy) # +confirmed.double+life_exp
summary(fit_x5)
predict(fit_x5, add2life_expectancy[is.na(add2life_expectancy$pop_largest_city),-c(1,5:7)], interval="prediction")
```

```
predict(fit_x5, add2life_expectancy[c(2,11,25,29,31,33,36,43,46),], interval="confidence")
# predict(fit_x5,type = "response")[is.na(add2life_expectancy$pop_largest_city)]
# Using linear regression, predicted value exceed the range.
imputed$imp$pop_largest_city[,2]
```

See the correlattion after imputing

Make log transform for X1, X2, X3, X4, X6

Also try to standardize the data. All of the data should have positive values. Thus, set 'center=F'.

```
cor(table_2020s1_imputed)
# y <- table_2020s1_imputed$confirmed.double
# X <- table_2020s1_imputed[,2:7]
# cor(scale(X,center = F, scale = T))
# cov(scale(X,center = F, scale=apply(X,2,sd,na.rm=T)))
table_2020s1_log <- table_2020s1_imputed
table_2020s1_log[,c(2:5,7)] <- log(table_2020s1_imputed[,c(2:5,7)])
table_2020s1_std<- data.frame(scale(table_2020s1_imputed, center=F, scale = T))
ggpairs(table_2020s1_log[,c(2:5,7)],aes(alpha=0.3))+theme_light()
```

Fit different models. Log transform made some large VIF values.

```
fit_2020s1<- lm(confirmed.double~population+land_area_skm+pop_density+pop_largest_city+life_expectancy+gdp_capita,ta
fit_2020s1_log<- update(fit_2020s1,data=table_2020s1_log)
fit_2020s1_std<- update(fit_2020s1,data=table_2020s1_std)
library(car)
vif(fit_2020s1)
vif(fit_2020s1_log)
vif(fit_2020s1_std)
fit_2020s1_log<- update(fit_2020s1_log,. ~ population+pop_density+pop_largest_city+life_expectancy+gdp_capita)
```

Stepwise selection reduce the model to two variables.

log transform gives a complex model.

```
library(MASS)
aic_2020s1 <- stepAIC(fit_2020s1,~.^2,direction = "both")
aic_2020s1_log <- stepAIC(fit_2020s1_log,~.^2,direction = "both")
# aic_2020s1_std <- stepAIC(fit_2020s1_std,~.^2)
summary(aic_2020s1)
summary(aic_2020s1_log)
# summary(aic_2020s1_std)
```

Update the model with two variables.

```
fit_2020s1_2v<- update(fit_2020s1,. ~ pop_largest_city + life_expectancy)
# fit_2020s1_2v_log<- update(fit_2020s1_log,. ~ pop_largest_city + life_expectancy)
# fit_2020s1_2v_std<- update(fit_2020s1_std,. ~ pop_largest_city + life_expectancy)
plot(aic_2020s1)
# plot(aic_2020s1_log)
# plot(aic_2020s1_std)
```

Box-Cox Transformations shows $\lambda = 0.5$ a square root transform may helps stablize the variance.

```
boxcox_2020s1<- boxcox(fit_2020s1)
boxcox_2020s1$x[which.max(boxcox_2020s1$y)]
aic_2020s1_sqrt <- update(aic_2020s1,sqrt(confirmed.double) ~.)
summary(aic_2020s1_sqrt)
plot(aic_2020s1_sqrt)
```

Using glm function with gaussian family and log link,

```
fit_2020s1_glm <- glm(confirmed.double~population+land_area_skm+pop_density+pop_largest_city+life_expectancy+gdp_ca
summary(fit_2020s1_glm)
plot(fit_2020s1_glm)
aic_2020s1_glm <- stepAIC(fit_2020s1_glm,~.^2,direction = "both")
summary(aic_2020s1_glm)
#comp <- fits.compare(fit_2020s1, fit_2020s1_log,fit_2020s1_std)
# comp
# plot(comp)
```