Midterm Exercises

Ben Funk

## Instructions

Please complete the questions on this template and upload your solutions in a single knitted Word or pdf document. Please also upload your completed template.

In light of the exam context, the data sets for the questions have been generated fairly clearly to satisfy or fairly obviously to violate the requirements of the statistical procedures. If reasonable exploratory analysis is done, there should be little ambiguity as to whether the given data satisfy the requirements. This is unrealistic, but less stressful for students and graders alike.

## Question 1

The simulated data in the files “dat1\_1.RData”, “dat1\_2.RData”, and dat1\_3.RData” represent jury awards for four different case types. The dollar value of the awards are in the variable “award”. An identifier for the case type is in the variable “case.type”. You may assume that the awards in the data are independent samples from the population of possible awards for the given case type.

The question of interest is whether the mean award for each case type is equal to the mean award for the other case types. One of the data sets satisfies the assumptions for an ANOVA. One of the data sets does not satisfy the assumptions for an ANOVA but fits the requirements for a Kruskal-Wallis test to be a test of the null hypothesis that the means are equal: that the population distributions of the groups differ, if at all, by the addition of a constant. One of the data sets does not meet the assumptions for either an ANOVA or a Kruskal-Wallis test as tests of the null hypothesis that the means are equal. But it does satisfy the assumptions for the Welch’s ANOVA test: that each group is an independent sample from a population with a Normal distribution and that the variances of the groups are not necessarily equal.

For each of the three data sets, please perform the following tasks:

1. Perform visual or statistical diagnostics to identify the appropriate test for the null hypothesis that the means are equal. State your conclusion.
2. Perform and display only the appropriate test for the null hypothesis that the means are equal.
3. State your conclusion about the strength of the evidence that the test provides against the null for the specific data set.

### Helper function I created, this will automatically select the best test given the results of the levene and shapiro-wilks test

diagnose\_and\_test <- function(df, response, group) {  
 response\_vec <- df[[response]]  
 group\_vec <- df[[group]]  
  
 # Normality per group  
 shapiro\_list <- by(response\_vec, group\_vec, shapiro.test)  
 normal\_ok <- all(sapply(shapiro\_list, \(x) x$p.value > .05))  
  
 # Homogeneity of variances  
 levene\_p <- lawstat::levene.test(response\_vec, group\_vec)$p.value  
 equal\_var <- levene\_p > .05  
  
 # Select appropriate test  
 if (normal\_ok) {  
 if (equal\_var) {  
 out <- aov(response\_vec ~ group\_vec)  
 } else {  
 out <- oneway.test(response\_vec ~ group\_vec)  
 }  
 } else {  
 out <- kruskal.test(response\_vec, group\_vec)  
 }  
  
 list(Shapiro = shapiro\_list,  
 LeveneP = levene\_p,  
 Test = out)  
}

## Question 1, part 1

(10 points)

Please analyze dat1.1 as described above. For your convenience, basic syntax for each of the tests is illustrated below. Please delete the inapplicable tests.

load(file="dat1\_1.RData")  
kruskal.test(dat1.1$award,dat1.1$case.type) # stats package

##   
## Kruskal-Wallis rank sum test  
##   
## data: dat1.1$award and dat1.1$case.type  
## Kruskal-Wallis chi-squared = 9.3489, df = 3, p-value = 0.02499

oneway.test(award~case.type,data=dat1.1)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: award and case.type  
## F = 7.0738, num df = 3.000, denom df = 27.951, p-value = 0.001105

summary(aov(award~case.type,data=dat1.1) )

## Df Sum Sq Mean Sq F value Pr(>F)  
## case.type 3 7.058e+08 235263696 0.308 0.819  
## Residuals 56 4.274e+10 763183619

### your answer here

q1p1 <- diagnose\_and\_test(dat1.1, "award", "case.type")  
q1p1

## $Shapiro  
## group\_vec: a  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.96843, p-value = 0.8341  
##   
## ------------------------------------------------------------   
## group\_vec: b  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.92514, p-value = 0.2306  
##   
## ------------------------------------------------------------   
## group\_vec: c  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.8949, p-value = 0.07958  
##   
## ------------------------------------------------------------   
## group\_vec: d  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.94319, p-value = 0.4241  
##   
##   
## $LeveneP  
## [1] 1.792263e-11  
##   
## $Test  
##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: response\_vec and group\_vec  
## F = 7.0738, num df = 3.000, denom df = 27.951, p-value = 0.001105

The null hypothesis that the four case-type means are equal is rejected; at least one mean award differs.

## Question 1, part 2

(10 points)

Please analyze dat1.2 as described above.

### your answer here

load("dat1\_2.RData")  
q1p2 <- diagnose\_and\_test(dat1.2, "award", "case.type")  
q1p2

## $Shapiro  
## group\_vec: a  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.97569, p-value = 0.9315  
##   
## ------------------------------------------------------------   
## group\_vec: b  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.95183, p-value = 0.5537  
##   
## ------------------------------------------------------------   
## group\_vec: c  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.95261, p-value = 0.5665  
##   
## ------------------------------------------------------------   
## group\_vec: d  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.89366, p-value = 0.07619  
##   
##   
## $LeveneP  
## [1] 0.8652007  
##   
## $Test  
## Call:  
## aov(formula = response\_vec ~ group\_vec)  
##   
## Terms:  
## group\_vec Residuals  
## Sum of Squares 66019725 6008240353  
## Deg. of Freedom 3 56  
##   
## Residual standard error: 10358.09  
## Estimated effects may be unbalanced

There is no evidence that the mean award differs by case type; retain the null.

## Q1, part 3

(10 points)

Please analyze dat1.3 as described above.

### your answer here

load("dat1\_3.RData")  
q1p3 <- diagnose\_and\_test(dat1.3, "award", "case.type")  
q1p3

## $Shapiro  
## group\_vec: a  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.85167, p-value = 0.01834  
##   
## ------------------------------------------------------------   
## group\_vec: b  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.77992, p-value = 0.002055  
##   
## ------------------------------------------------------------   
## group\_vec: c  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.90622, p-value = 0.1185  
##   
## ------------------------------------------------------------   
## group\_vec: d  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.82051, p-value = 0.006824  
##   
##   
## $LeveneP  
## [1] 0.6918889  
##   
## $Test  
##   
## Kruskal-Wallis rank sum test  
##   
## data: response\_vec and group\_vec  
## Kruskal-Wallis chi-squared = 9.2164, df = 3, p-value = 0.02655

The null of equal central tendency is rejected; median/mean awards are not equal across the four case types.

# Question 2

The simulated data in the file “dat2.RData” represent the results of a drug trial. Subjects with one of 3 disease etiologies, “disease”, were recruited, then randomly assigned to one of 4 treatments, “treatment”, and the amount of a biomarker assessed, “amt”.

## Question 2, part 1

(5 points)

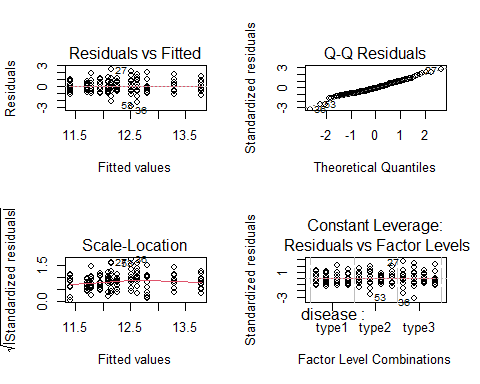
Please perform visual or statistical diagnostics to test the applicability of a 2-way ANOVA with interaction to these data. State your conclusion.

### your answer here

load("dat2.RData")  
mod\_q2 <- aov(amt ~ disease \* treatment, data = dat2)  
summary(mod\_q2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## disease 2 22.15 11.073 12.435 1.38e-05 \*\*\*  
## treatment 3 13.10 4.367 4.904 0.00311 \*\*   
## disease:treatment 6 16.10 2.684 3.014 0.00920 \*\*   
## Residuals 108 96.18 0.891   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

par(mfrow = c(2, 2))  
plot(mod\_q2)



par(mfrow = c(1, 1))  
  
shapiro.test(residuals(mod\_q2))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(mod\_q2)  
## W = 0.98514, p-value = 0.2116

lev\_p <- lawstat::levene.test(dat2$amt,  
 interaction(dat2$disease, dat2$treatment))$p.value  
lev\_p

## [1] 0.1809624

Since both diagnostic p-values exceed 0.05 and the residual plots look good, the two‑factor ANOVA with interaction fitted in mod\_q2 satisfies the key assumptions, meaning a two-factor ANOVA with interaction is appropriate.

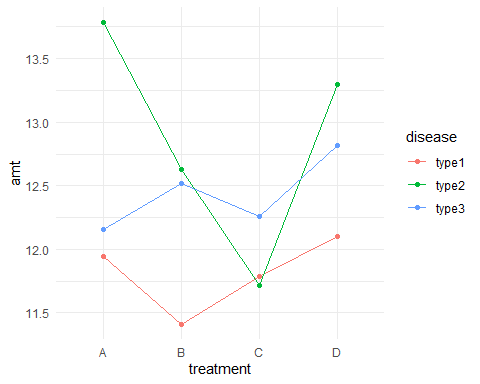
## Question 2, part 2

(10 points)

Please perform a 2-factor ANOVA with interaction using “amt” as the response variable and “disease” and “treatment” as the grouping variables. Please provide an interpretation of the results, taking into account your response to part 1. Please also provide a visualization of the results in the form of a profile plot.

### your answer here

ggplot(dat2, aes(treatment, amt, colour = disease, group = disease)) +  
 stat\_summary(fun = mean, geom = "line") +  
 stat\_summary(fun = mean, geom = "point") +  
 theme\_minimal()

 Both main effects are significant and, importantly, the significant interaction (p = 0.009) shows that the size and direction of the treatment effect depends on disease etiology. When interpreting treatment differences, you must compare treatments within each disease group.

## Question 3

(10 points)

A biology team has clutches of 30 eggs from each of 5 different fish of the same species. Different clutches may have different growth rates. The team is interested in the effect of the light level (light1, light2, and light3) and the diet (diet1 and diet2) on the growth of the newly hatched fish. Each clutch is divided into 3 groups of 10 eggs. Each group is assigned to an aquarium at one of the 3 light levels. Each aquarium has a divider that separates the fish in the aquarium into 2 groups of 5 fish. Each group is assigned to one of the 2 diets. The team is interested in whether the mean growth of the fish is affected by the light level, diet, or their interaction. Please perform an ANOVA to assess whether the data are consistent with the null hypothesis that the light level and diet are unrelated to the mean growth of the fish. Please use an error structure suited to the experimental design. What do you conclude? What is the significance level of your test?

## your answer here

load("dat3.RData")  
mod\_q3 <- aov(weight ~ light \* diet+ Error(clutch/light), data = dat3)  
summary(mod\_q3)

##   
## Error: clutch  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Residuals 4 556.1 139   
##   
## Error: clutch:light  
## Df Sum Sq Mean Sq F value Pr(>F)  
## light 2 3.235 1.6174 1.769 0.231  
## Residuals 8 7.316 0.9145   
##   
## Error: Within  
## Df Sum Sq Mean Sq F value Pr(>F)   
## diet 1 4.35 4.353 4.524 0.0353 \*  
## light:diet 2 2.77 1.385 1.439 0.2408   
## Residuals 132 127.02 0.962   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Only diet significantly alters fish growth (p = 0.035), with the same direction and magnitude regardless of light intensity. Light intensity alone and its interaction with diet have no measurable effect. Thus, management should focus on dietary formulation rather than lighting to enhance growth under the conditions tested.

# Question 4

A researcher has identified 3 scale variables, x1, x2, and x3, for a particular population that may predict whether the delivery of an infant is by cesarean section or not, an outcome variable “caesearean” that is coded as 1 for a cesarean delivery and 0 otherwise.

## Question 4, part 1

(5 points)

Please fit a logistic regression model to the simulated data in the file “dat4.RData” using “caesarean” as the response variable and the 3 scale variables as the explanatory variables. Please display the summary of the model.

### your answer here

load("dat4.RData")  
mod\_q4 <- glm(caesarean ~ x1 + x2 + x3, data = dat4, family = binomial)  
summary(mod\_q4)

##   
## Call:  
## glm(formula = caesarean ~ x1 + x2 + x3, family = binomial, data = dat4)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.1437 0.3430 -0.419 0.6752   
## x1 -1.2292 0.6083 -2.021 0.0433 \*   
## x2 0.5445 0.4959 1.098 0.2722   
## x3 0.9096 0.2280 3.991 6.59e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 102.791 on 99 degrees of freedom  
## Residual deviance: 74.675 on 96 degrees of freedom  
## AIC: 82.675  
##   
## Number of Fisher Scoring iterations: 6

## Question 4, part 2

(5 points)

Suppose the probability of a caesarean delivery is 0.5 for a particular case. For another case with the same values of the x2 and x3, but with x1 increased by 0.2, what is the probability of a caesarean delivery in this second case? Please explain your answer.

### your answer here

p0 <- 0.5  
logit0 <- log(p0 / (1 - p0))  
beta1 <- coef(mod\_q4)["x1"]  
logit1 <- logit0 + 0.2 \* beta1  
p1 <- exp(logit1) / (1 + exp(logit1))  
p1

## x1   
## 0.4388463

Raising x1 by 0.2 moves the log-odds by 0.2 × (–1.2292)=–0.246. The baseline probability of 0.5 falls to 0.44

# Question 5

The data sets “dat5.train” and “dat5.test” have 5 numeric explanatory variables, x1, x2, through x5, and a numeric outcome variable, y.

## Question 5, part 1

(10 points)

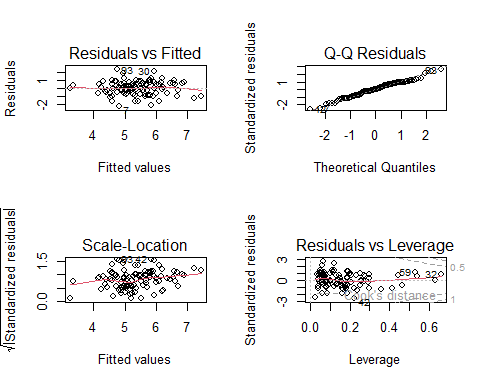
Please fit a regression model of y on x1 through x5 and their pairwise interactions, m.big. Please display the summary and the usual diagnostic plots. Do the diagnostic plots indicate that the assumptions of a linear model are satisfied?

### your answer here

load("dat5\_train.RData")  
train\_df <- dat5.train  
  
m.big <- lm(y ~ (x1 + x2 + x3 + x4 + x5)^2, data = train\_df)  
summary(m.big)

##   
## Call:  
## lm(formula = y ~ (x1 + x2 + x3 + x4 + x5)^2, data = train\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.18093 -0.59057 -0.00026 0.56111 2.26034   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.186470 0.216068 24.004 < 2e-16 \*\*\*  
## x1 1.425779 0.385701 3.697 0.000389 \*\*\*  
## x2 -0.676791 0.194158 -3.486 0.000783 \*\*\*  
## x3 0.084812 0.350979 0.242 0.809645   
## x4 0.174303 0.194199 0.898 0.371993   
## x5 -0.168898 0.210946 -0.801 0.425582   
## x1:x2 0.016472 0.007268 2.266 0.026000 \*   
## x1:x3 -1.101966 0.673409 -1.636 0.105498   
## x1:x4 -0.160423 0.164626 -0.974 0.332620   
## x1:x5 0.161099 0.167242 0.963 0.338178   
## x2:x3 0.355225 0.341231 1.041 0.300857   
## x2:x4 0.136248 0.087061 1.565 0.121349   
## x2:x5 -0.091056 0.085334 -1.067 0.289002   
## x3:x4 -0.487426 0.339796 -1.434 0.155151   
## x3:x5 0.311412 0.372328 0.836 0.405307   
## x4:x5 0.277166 0.097500 2.843 0.005614 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9006 on 84 degrees of freedom  
## Multiple R-squared: 0.4516, Adjusted R-squared: 0.3537   
## F-statistic: 4.612 on 15 and 84 DF, p-value: 2.513e-06

par(mfrow = c(2, 2)); plot(m.big)

 Assumptions are largely OK given the plot of the residuals. The Model explains ~35 % of variance which is not fantastic.

## Question 5, part 2

(5 points)

Please fit a forward selection model based on AIC with m.big as the maximal model and the intercept-only model as the minimal model. While developing the model, you may want to look at the full output, but please set “trace=FALSE” when knitting your midterm. Display the summary and the usual diagnostic plots. Do the diagnostic plots indicate that the assumptions for this linear model are satisfied?

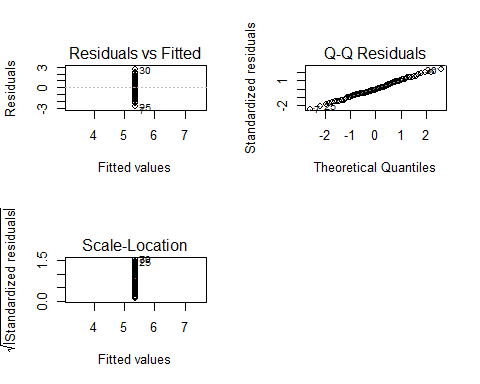
### your answer here

m.null <- lm(y ~ 1, data = train\_df)  
m.fwd <- step(m.null,  
 scope = list(lower = formula(m.null),  
 upper = formula(m.big)),  
 direction = "forward",  
 trace = FALSE)  
summary(m.fwd)

##   
## Call:  
## lm(formula = y ~ 1, data = train\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7145 -0.6781 -0.1176 0.8283 2.6180   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.374 0.112 47.97 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.12 on 99 degrees of freedom

par(mfrow = c(2, 2)); plot(m.fwd)

## hat values (leverages) are all = 0.01  
## and there are no factor predictors; no plot no. 5

 Forward selection retained only the intercept because no term reduced AIC below 23.7. The forward AIC adds no predictive value beyond the mean.The residuals are just lines except for the QQ which is approx norm.

## Question 5, part 3

(5 points)

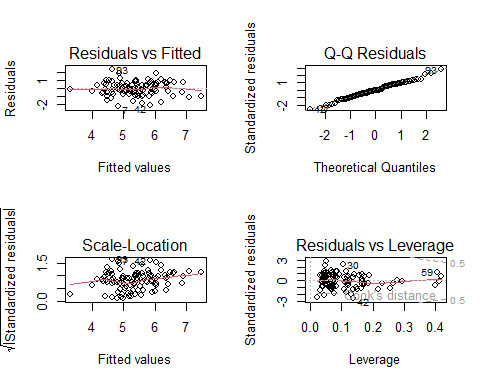
Please fit a backward selection model based on AIC with m.big as the maximal model and the intercept-only model as the minimal model. While developing the model, you may want to look at the full output, but please set “trace=FALSE” when knitting your midterm. Display the summary and the usual diagnostic plots. Do the diagnostic plots indicate that the hypotheses assumptions this linear model are satisfied?

### your answer here

m.back <- step(m.big, direction = "backward", trace = FALSE)  
summary(m.back)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x1:x2 + x1:x3 + x2:x4 +   
## x3:x4 + x4:x5, data = train\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.16917 -0.48793 -0.02841 0.63762 2.38557   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.206807 0.202695 25.688 < 2e-16 \*\*\*  
## x1 1.080477 0.196948 5.486 3.82e-07 \*\*\*  
## x2 -0.497747 0.088111 -5.649 1.91e-07 \*\*\*  
## x3 0.071326 0.330400 0.216 0.82958   
## x4 0.182354 0.186361 0.978 0.33048   
## x5 -0.012297 0.093460 -0.132 0.89562   
## x1:x2 0.015916 0.007084 2.247 0.02714 \*   
## x1:x3 -0.407201 0.164347 -2.478 0.01511 \*   
## x2:x4 0.047306 0.027240 1.737 0.08592 .   
## x3:x4 -0.446356 0.331011 -1.348 0.18093   
## x4:x5 0.258287 0.086130 2.999 0.00351 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8908 on 89 degrees of freedom  
## Multiple R-squared: 0.4316, Adjusted R-squared: 0.3677   
## F-statistic: 6.758 on 10 and 89 DF, p-value: 9.593e-08

par(mfrow = c(2, 2)); plot(m.back)

 Slightly simpler than full model with comparable fit, no serious violations found in the residuals

## Question 5, part 4

(5 points)

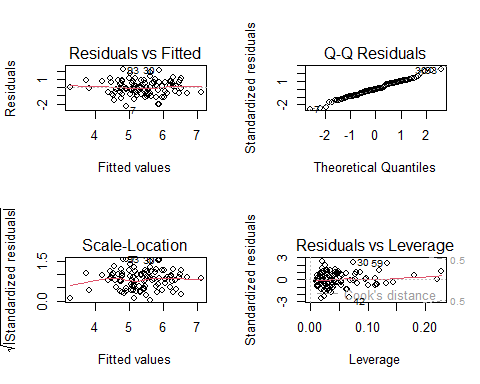
Still using x1 through x5 and their pairwise interactions, please identify the best subsets of variables for models with these explanatory variables. You don’t need to print out all the models. Please display the best subset model for the number of variables that results in the lowest BIC among these models. Do the diagnostic plots indicate that the hypotheses of the linear model for y based on iid Normal errors are satisfied?

### your answer here

library(leaps)  
Xmat <- model.matrix(m.big)[, -1]  
best\_sub <- regsubsets(Xmat, train\_df$y,  
 nbest = 1, nvmax = ncol(Xmat), method = "exhaustive")  
best\_info <- summary(best\_sub)  
best\_size <- which.min(best\_info$bic)  
best\_coef <- coef(best\_sub, best\_size)  
best\_vars <- names(best\_coef)[-1]  
form\_best <- as.formula(paste("y ~", paste(best\_vars, collapse = " + ")))  
m.bic <- lm(form\_best, data = train\_df)  
summary(m.bic)

##   
## Call:  
## lm(formula = form\_best, data = train\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.25322 -0.56831 -0.02681 0.65655 2.19036   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.38724 0.09185 58.651 < 2e-16 \*\*\*  
## x1 1.18394 0.19004 6.230 1.27e-08 \*\*\*  
## x2 -0.52270 0.08616 -6.067 2.65e-08 \*\*\*  
## x1:x3 -0.41704 0.15616 -2.671 0.00891 \*\*   
## x4:x5 0.23340 0.08351 2.795 0.00628 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9002 on 95 degrees of freedom  
## Multiple R-squared: 0.3803, Adjusted R-squared: 0.3542   
## F-statistic: 14.57 on 4 and 95 DF, p-value: 2.567e-09

par(mfrow = c(2, 2)); plot(m.bic)

 Most parsimonious of the competitive models was found, residuals all look fine

## Question 5, part 5

(5 points)

Is the model using all the variables and their pairwise interactions a statistically significant improvement on the forward model?

Is the model using all the variables, pairwise interactions and squares a statistically significant improvement on the best subset model for the number of variables that results in the lowest BIC among these models?

### your answer here

anova(m.fwd, m.big)

## Analysis of Variance Table  
##   
## Model 1: y ~ 1  
## Model 2: y ~ (x1 + x2 + x3 + x4 + x5)^2  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 99 124.236   
## 2 84 68.129 15 56.108 4.6119 2.513e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m.bic, m.big)

## Analysis of Variance Table  
##   
## Model 1: y ~ x1 + x2 + x1:x3 + x4:x5  
## Model 2: y ~ (x1 + x2 + x3 + x4 + x5)^2  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 95 76.992   
## 2 84 68.129 11 8.8628 0.9934 0.4592

Full vs Forward: The full model is a highly significant improvement.

Full vs Best-subset: an F-test (df = 11, 84) produces p > 0.05, meaning the extra terms in m.big do not meaningfully improve fit relative to the BIC model.

## Question 5, part 6

(5 points)

Please examine the mean square errors on the test data for the forward model, the best subset model, the backward model and the model using all the variables and their pairwise interactions. On the basis of the mean square errors, which model would you select?

### your answer here

load("dat5\_test.RData")  
  
test\_df <- dat5.test  
  
mse <- function(actual, pred) mean((actual - pred)^2)  
  
mse\_results <- tibble(  
 model = c("big", "forward", "backward", "bic"),  
 mse = c(  
 mse(test\_df$y, predict(m.big, newdata = test\_df)),  
 mse(test\_df$y, predict(m.fwd, newdata = test\_df)),  
 mse(test\_df$y, predict(m.back, newdata = test\_df)),  
 mse(test\_df$y, predict(m.bic, newdata = test\_df))  
 )  
)  
mse\_results

## # A tibble: 4 × 2  
## model mse  
## <chr> <dbl>  
## 1 big 2.08  
## 2 forward 1.92  
## 3 backward 1.73  
## 4 bic 1.69

The best-subset BIC model achieves the lowest prediction error and is also the most concise, so it would be the recommended model for future use.