STOR 664: FALL 2020
Final Exam, November 20, 2020

This is an open-book, remote-learning exam. Time limit: 6 hours. Access to course materials and standard computational tools (in particular, R) is allowed; communication with other students or with anybody via the internet, other than the instructor, is not. The university Honor Code is in effect at all times. Answers may be given in any of the following formats (including combinations of more than one): R Markdown, Word, Latex or handwritten pages scanned or photographed; if handwritten, it is recommended you use blue or black ink on plain sheets of white paper. They should then be uploaded in gradescope. The exam is worth 100 points total (30 for questions 1 and 2, 40 for question 3); points for each part-question are stated below. Although the questions are intended to be answered in sequence, you may write out your answers in any order and errors in one part-question will not prevent you gaining full credit in other parts of the same question. Attempt all questions.

X_1	X_2	X_3	X_4	X_5	Y	X_1	X_2	X_3	X_4	X_5	Y
58.8	7107	21	129	52	3067	39.5	14119	20	187	37	3286
65.2	6373	22	141	68	2828	44.5	16691	22	195	42	3542
70.9	6796	22	153	29	2891	43.6	14571	19	206	22	3125
77.4	9208	20	166	23	2994	56.0	13619	22	198	28	3022
79.3	14792	25	193	40	3082	64.7	14575	22	192	7	2922
81.0	14564	23	189	14	3898	73.0	14556	21	191	42	3950
71.9	11964	20	175	96	3502	78.9	18573	21	200	33	4488
63.9	13526	23	186	94	3060	79.4	15618	22	200	92	3295
54.5	12656	20	190	54	3211						

Table 1: Water usage data set.

X_1	Average monthly temperature (^oF)
X_2	Average of production (thousands of pounds)
X_3	Number of operating plant days in the month
X_4	Number of persons on the monthly plant payroll
X_5	Two-digit random number
Y	Monthly water usage (gallons)

Table 2: Variables used in Table 1.

1. A production plant cost-control engineer is responsible for cost reduction. One of the costly items in her plant is the amount of water used by the production facilities each month. She decides to investigate water usage by collecting 17 observations of the plant's water usage and other variables. She had heard about multiple regression, but since she was quite skeptical she added a column of random numbers to the original observations. Then she asked her team's statistician to analyze the data and make recommendations about the control variables

$X_1, \ldots, X_5$. The complete set of data is shown in Table 1. The variables are described in Table 2.

You are the team's statistician. Analyze the data, with appropriate consideration for variable selection, diagnostics, transformations and multicollinearity. The objective is to find the best model for predicting Y as a function of $X_1, ..., X_5$.

Write a report summarizing the results of this exercise. In particular, your report should discuss as many as possible of the following items:

(a) The best regression model with variables selected from $X_1, ..., X_5$; **[7 points]**

(b) The goodness of fit of the model with all five variables included, taking into account the various diagnostics available within R; **[8 points]**

(c) The interpretation of statistics regarding outliers, points of high leverage, influence diagnostics and multicollinearity; **[8 points]**

(d) The final conclusions that the engineer should draw. In particular, state which variables are predictors of water usage, whether the associations are positive or negative, and the degree of confidence you have in each of your conclusions. **[7 points]**

2. A dataset was collected on median housing prices in 506 neighorhoods of a large city, together with a number of potential explanatory variables. The dataset contains the following variables:

- crim: per capita crime rate by town;
- zn: proportion of residential land zoned for lots over 25,000 sq.ft.;
- indus: proportion of non-retail business acres per town;
- chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise);
- nox: nitrogen oxides concentration (parts per 10 million);
- rm: average number of rooms per dwelling;
- age: proportion of owner-occupied units built prior to 1940;
- dis: weighted mean of distances to five Boston employment centres;
- rad: index of accessibility to radial highways;
- tax: full-value property-tax rate per $10,000;
- ptratio: pupil-teacher ratio by town;
- black: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town;
- lstat: lower status of the population (percent);
- medv: median value of owner-occupied homes in $1000s.

The object of the exercise is to predict `mdev` as a function of all the other variables.

For the purpose of the present exercise, the dataset has been split randomly into two parts, `housing_train.csv` which is a training dataset of 400 neighborhoods, and `housing_test.csv` which is a test dataset of 400 neighborhoods. Both are `csv` files and may be read into R with the `read.csv` command.

(a) Based on the training dataset, find the best model by variable selection, and use cross-validation to evaluate its performance (mean square prediction error or MSPE). Then, validate your conclusion by running it on the test dataset and calculating the MSPE for that. **[6 points]**

(b) Repeat the exercise of (a) using lasso regression. **[6 points]**

(c) Repeat the exercise of (a) using ridge regression. **[6 points]**

(d) Repeat the exercise of (a) using elastic net regression with $\alpha = 0.5$. **[6 points]**

(e) Compare and contrast the four methods for finding a predictor of `mdev`. **[6 points]**

3. An experiment was performed to compare two treatments that are used in spinning cotton thread. One is "Flyer" (1 or 2), which represent two different machines for spinning the fiber. The other is "Twist" (1, 2, 3 or 4), which indicates four levels of twisting the fiber. The original intention was to use all eight possible combinations of Flyer and Twist, but it was found that the combinations (Flyer=1, Twist=1) and (Flyer=2, Twist=4) were unstable, so these were not used. The experiment was laid out in 13 blocks, as in Table 3.

		Block												
Flyer	Twist	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13
1	2	6	9.7	7.4	11.5	17.9	11.9	10.2	7.8	10.6	17.5	10.6	10.6	8.7
1	3	6.4	8.3	7.9	8.8	10.1	11.5	8.7	9.7	8.3	9.2	9.2	10.1	12.4
1	4	2.3	3.3	7.3	10.6	7.9	5.5	7.8	5	7.8	6.4	8.3	9.2	12.0
2	1	3.3	6.4	4.1	6.9	6.0	7.4	6.0	7.3	7.8	7.4	7.3	10.1	7.8
2	2	3.7	6.4	8.3	3.3	7.8	5.9	8.3	5.1	6.0	3.7	11.5	13.8	8.3
2	3	4.2	4.6	5.0	4.1	5.5	3.2	10.1	4.2	5.1	4.6	11.5	5.0	6.4

Table 3: Breakage rate of cotton fibers.

A plausible model for this experiment is

$$y_{ijk} \quad = \quad \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}, \tag{1}$$

where y_{ijk} denotes an arbitrary observation with Flyer i, Twist j and Block k, μ is an overall mean, $\alpha_i,\ \beta_j,\ \gamma_k$ are the Flyer, Twist and Block effects, and ϵ_{ijk} is an error term which we assume to be $N[0, \sigma^2]$ independent for each (i, j, k). The range of indexes in (1) is $i = 1, 2,\ j = 1, 2, 3, 4,\ k = 1, \ldots, 13$ where we assume the cases $i = j = 1$ and $i = 2, j = 4$ are not observed. The parameters $\alpha_i,\ \beta_j,\ \gamma_k$ are subject to some constraints to make them well-determined but we shall not need to worry about these in the following analysis.

(a) Fit the model (1) to the data, using the `lm` command in R. Note that you will need to use some format such as `lm(Breakrate~factor(Flyer)+...,data=cotton)` to ensure that all three covariates are treated as factors and not as numerical variables. Based on the results, state an estimate of $\alpha_1 - \alpha_2$ (the mean difference between Flyer 1 and Flyer 2), and its standard error. Also state an estimate of σ. **[7 points]**

(b) Can any of the Flyer, Twist or Block factors be dropped from the model? Refit model (1) with each factor dropped in turn, and state your conclusions. **[6 points]**

(c) The Breakrate data in Table 1 are in the form of a 6×13 matrix. If we write the six row means as $M_1,\ M_2,\ \ldots,\ M_6$, show that a natural estimate of $\alpha_1 - \alpha_2$ is $(M_1 + M_2 - M_5 - M_6)/2$ and derive its (theoretical) variance. Show that the results are consistent with the estimate and standard error you reported in part (a). **[7 points]**

(d) The experimental design used for this study will be called Design 1. An alternative design, if it was feasible, would be to use each of the eight combinations of Flyer and Twist. We call this Design 2. What would be a suitable estimator of $\alpha_1 - \alpha_2$ under Design 2, and what would be its variance? **[7 points]**

(e) Suppose, for some integer k, you had the choice of running Design 1 in $4k$ blocks or Design 2 in $3k$ blocks (the number of blocks is chosen so that the total number of experimental units is the same in both designs). If the objective is still to estimate $\alpha_1 - \alpha_2$, which estimate would you prefer and why? **[7 points]**

(f) After further study of the apparatus, the experimenter determines that the combinations (Flyer=1, Twist=1) and (Flyer=2, Twist=4) are possible, but that the resulting measurement errors will have variances $3\sigma^2$ instead of σ^2 (but the other measurements will still have variance σ^2). Which design do you prefer now? **[6 points]**

Solutions

General Comment on Student Solutions. In general, I felt that many students overinterpreted the questions, looking for things that were not asked and not identifying the things that were. I should probably have given more emphasis to the point that I repeatedly stress with my undergraduate class: *Always answer the question!* As a result, this was a difficult exam to grade, and the scores came out lower than I would usually expect with an exam of this nature. Even with most students getting a boost from their scores on the project, the combined scores for the whole course were lower than I would normally expect, but I allowed for that in setting the grade boundaries. Overall, I felt that most students worked hard on the course and the final grades reflected that; the scores are a reflection of the style of both the midterm and the final exam and should not be interpreted as a criticism of student effort.

In the following solutions, comments that are specifically intended as a comment on student solutions are noted in red.

1. (a) A simple sequence of commands would be a best subset regression followed by AIC calculation, for example, if the data frame is called X,

```
require(leaps)
lm1=regsubsets(Y~.,data=X)
lm1s=summary(lm1)
AIC=17*log(lm1s$rss/17)+(1:5)*2
plot(AIC~I(1:5),ylab='AIC',xlab='Number of Predictors',pch=20)
```

which produces the left-hand plot in Figure 1. To see which variables are included in which stage of the model, `lm1s$which` produces

```
  (Intercept)    X1   X2    X3    X4    X5
1        TRUE FALSE TRUE FALSE FALSE FALSE
2        TRUE FALSE TRUE FALSE  TRUE FALSE
3        TRUE  TRUE TRUE FALSE  TRUE FALSE
4        TRUE  TRUE TRUE  TRUE  TRUE FALSE
5        TRUE  TRUE TRUE  TRUE  TRUE  TRUE
```

showing that the variables are entered in the order X2, X4, X1, X3, X5 (and in sequence, i.e. each model is nested in the one below, which is not guaranteed when using best subsets regression). This suggests that the best model order is 4, i.e. we select the model containing X1–X4 but not X5. An alternative way of selecting the model order, given the nested structure, is through a sequence of hypothesis tests through the `anova` function. In this case, the successive p-values are 0.03 for model 1 against model 2; 0.18 for model 2 against model 3; 0.022 for model 3 against model 4; 0.58 for model 4 against model 5. These results mean that we might be tempted to stop at model 2, but a test of model 2 against model 4 has a p-value of 0.027, in other words, rejecting model 2 in favor of model 4 at a significance level of 0.05. Therefore, both AIC and successive testing suggest model 4 is best. However, you may get different conclusions with other model selection criteria, e.g. C_p, BIC or hypothesis testing with a different significance level than 0.05. [**7 points**]

Some students went beyond the above solution and also looked for transformations using the `boxcox` graphical procedure. Specifically, this looks for a transformation of the form $y \to \frac{y^\lambda - 1}{\lambda}$. If you do this, using the full range of λ and with other features of the model unchanged, you get the bizarre result that the maximum likelihood estimator is $\hat{\lambda} = -3.51$. Kudos to those students who figured that out, and some even continued through the whole analysis based on that transformation, but I would find it very hard to explain a result based on that transformation. There's no obvious "need" for a

transformation based on skewness or other commonly cited criteria. The same applies, I think, to each of the possible covariates.

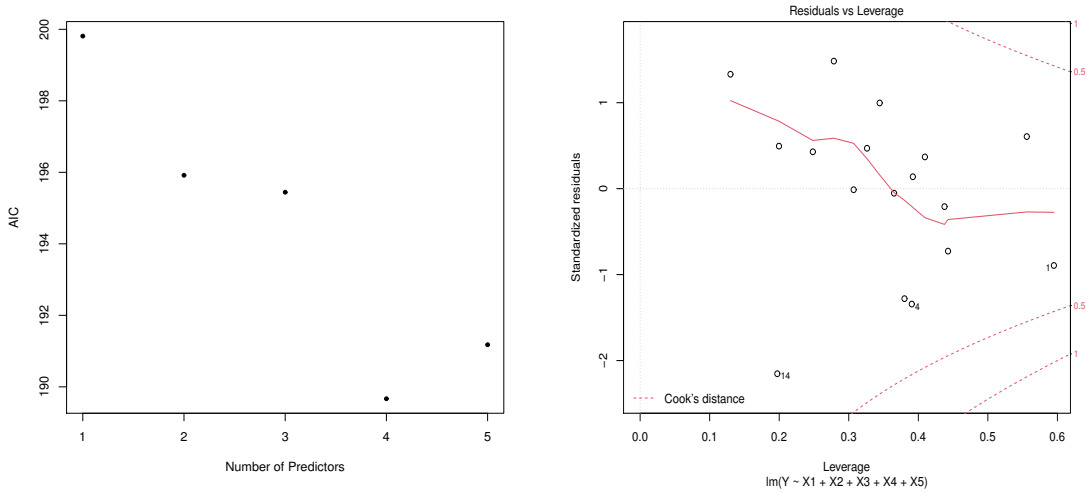

Figure 1: (left) AIC plot in question 1(a). (right) Influence diagram for the model with all five variables.

(b) The standard residual plots based on the five-variable model (e.g. `plot(lm5)`) do not show any extreme outliers, e.g. see Figure 1, right hand side, for the leverage and Cook's distance plot, which does not suggest any extreme data points. Also, the qqplot of residuals (not shown) suggests good fit to a normal distribution. **[8 points]**
If you want to go further than just showing a qqplot of residuals, you could test for normality more formally using the Shapiro-Wilk test, for example

```
> lm5=lm(Y~X1+X2+X3+X4+X5,X)
> shapiro.test(residuals(lm5))
...
W = 0.95481, p-value = 0.5369
```

The p-value is well above 0.05, so the residuals are consistent with a normal distribution.

(c) Outliers: if you use `library(MASS)` followed by `studres(lm5)`, it turns out that observation 14 has studentized residual –2.699, which is potentially problematic (the null distribution is t_{10}, as discussed on p. 178 of Smith and Young; this gives it a two-sided p-value of 0.022). However based on `lm4` (the model with X5 omitted), the studentized residual is –2.11 and a 2-sided p-value of 0.06; this is not a large enough residual (or the p-value not small enough) to be considered problematic.

If you are really astute, you might notice that observation 14 is also the one with the smallest value of X5 (the one variable that we know is not significant), so this contributes to it being an outlier under model `lm5`.

Some students noted the recommendation to use a Bonferroni correction to test for the largest studentized residual in a sample; by this method, even the vlaue –2.699 is not statistically significant. Personally, I would treat this value as somewhat suspicious even after applying the Bonferroni correction, but it is fair comment to note that the residual could have arisen without anything being wrong with the model.

Leverage: Use `influence(lm5)$hat`: the largest hat value is 0.595 in observation 1, but this is less than twice the mean leverage ($p/n = 6/17 = 0.353$), so not considered especially high.

Influence diagnostics: Using `cooks.distance(lm5)`, the largest value is 0.196, less than 0.5, so not a problem.

Multicollinearity: Computing correlations shows that X2 and X4 are highly correlated ($r = 0.918$) and X1 and X3 are moderately correlated ($r = 0.438$); other pairwise correlations are low. Following Faraway, around page 107,

```
library(faraway)
xx=model.matrix(lm5)[,-1]
vif(xx)
sqrt(eigen(t(xx)%*%xx)$val[1]/eigen(t(xx)%*%xx)$val)
xx=scale(xx)
vif(xx)
sqrt(eigen(t(xx)%*%xx)$val[1]/eigen(t(xx)%*%xx)$val)
```

The VIFs, with or without scaling, are 1.26, 6.7, 1.27, 6.74, 1.04, which shows there is some multicollinearity but not enough to be critical (usually defined when VIF$>$ 10). The condition indices are 1, 342,...,7546 without the scaling, but are 1, 1.16, 1.45, 1.89, 5.05 aftr scaling, which does not indicate a serious problem. (Note: some of these answers might have come in part (b), but they will earn credit wherever they appear.) **[8 points]**

(d) The best model seems to be that involving X1–X4 but not X5. It looks as though average monthly temperature and average production tend to increase the water usage, while the number of operating plant days per month and the number of persons on payroll tend to decrease. However, these conclusions might be affected by collinearity, especially between X2 and X4.

[7 points]

2. (a) Variable selection method: an application of `regsubsets` followed by an AIC calculation concludes that the minimum AIC model contains the following 11 covariates: crim, zn, chas, nox, rm, dis, rad, tax, ptratio, black, lstat (the only ones missing are "indus" and "age"). For example,

```
require(leaps)
lm1=regsubsets(medv~crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+black+lstat,nvmax=13,data=houstr)
lm1s=summary(lm1)
lm1s$which
   (Intercept)  crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio black lstat
..
11        TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE    TRUE  TRUE  TRUE
AIC=400*log(lm1s$rss/400)+(1:13)*2
 [1] 1450.717 1377.787 1331.630 1318.651 1297.989 1287.144 1281.123 1277.518 1273.454 1269.870 1264.049 1265.920 1267.881
```

A cross-validation based on leaving out one observation at a time leads to an estimated MSPE of 24.42. For example,

```
MSPE=0
for(i in 1:400){lm1a=lm(medv~crim+zn+chas+nox+rm+dis+rad+tax+ptratio+black+lstat,subset=-i,data=houstr)
MSPE=MSPE+(predict(lm1a,newdata=houstr)[i]-houstr$medv[i])^2}
print(MSPE/400)
```

When this optimal model is applied to the test data, the RMSE turns out to be 20.72, for example,

```
lm1=lm(medv~crim+zn+chas+nox+rm+dis+rad+tax+ptratio+black+lstat,data=houstr)
pr1=predict(lm1,newdata=houste)
sum((pr1-B2$medv)^2)/106
```

[6 points]

(b) There is one point of data management here, which is that if you scale the X variables in the training dataset (here labelled `houstr` as above), then you must make the same scaling in the test dataset (`houste`). The simplest way to achieve this may be to combine both sets of covariates into one, then rescale, and then split them again:

```
X=rbind(as.matrix(houstr[,1:13]),as.matrix(houste[,1:13]))
X=scale(X)
xtr=X[1:400,]
xtest=X[401:506,]
```

We can then proceed with a `glmnet` application, along the following lines:

```
y=as.vector(houstr$medv)
library(glmnet)
fit=glmnet(xtr,y)
cvfit=cv.glmnet(xtr,y,nfolds=400)
plot(cvfit)
cvfit$lambda.min
coef(cvfit, s = "lambda.min")
min(cvfit$cvm)
# [1] 24.70014
which.min(cvfit$cvm)
# 69
cvfit$lambda.min
# 0.01195695
sum(abs(fit$beta[,69]))
# 20.86684
```

This shows that the optimal λ is 0.01195..., which corresponds to column 69 of the output matrix, and achieves a cross-validation score of 24.70. It also achieves an L1 Norm (sum of absolute values of all the regression coefficients, except the intercept) of 20.9 to one decimal place. The cross-validation scores and the optimal model fit are represented on Figure 2.
Finally, using the optimal fit to predict the test values gives MSPE of 20.91 to two decimal places, for example,

```
pr1=predict(fit,xtest)
mean((pr1[,69]-houste$medv)^2)
# 20.90643
```

[6 points]

Many students omitted the data scaling step entirely: the results are quite similar to the above and I gave credit so long as the basic method was correct. Also, some students used alternative measures of fit, such as RMSE rather than MSPE (square roots of all the above MSPEs). Another variant was the some students did not use leave-one-out cross-validation but instead one of the alternative methods were discussed in class, such as a k-fold crossvalidation with, e.g. $k = 5$ or 10. I felt that all these were valid alternatives, but I did not give full credit to those students who omitted cross-validatoin entirely, since the question did ask for that.

(c) Repeat all the above steps with `alpha=0` in the `glmnet` and `cv.glmnet` functions: the key differences are that the minimum cross-validation score is 24.94 (achieved in column 100, i.e. the smallest λ among the 100 values fitted), and the MSPE on the test dataset is 21.89. **[6 points]**

(d) Repeat all the above steps with `alpha=0.5` in the `glmnet` and `cv.glmnet` functions: the key differences are that the minimum cross-validation score is 24.70 (achieved in column 70), and the MSPE on the test dataset is 20.92. **[6 points]**

(e) It looks as though the lasso and elastic net methods perform about the same, with ridge regression a little bit worse but also comparable, while variable selection still comes out best among all the methods, just like several examples in class! **[6 points]**

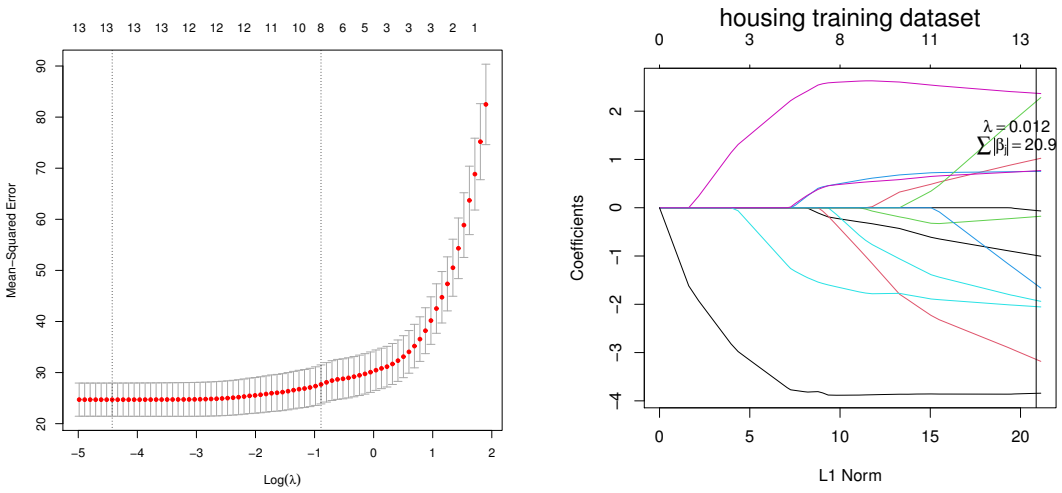

Figure 2: Plots of glmnet output in question 2(b)

3. The original version of the data file `cotton.csv` included `Breakrate` variables that were 10 times too large, an error that was noted during the exam. The solution here assumes the data have been corrected and are the same as given in Table 1.

   (a) The command `lm1=lm(Breakrate~factor(Flyer)+factor(Twist)+factor(Block),Cotton)` followed by `summary(lm1)` leads to (after some editing)

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.9474     1.2433   5.588 5.69e-07 ***
factor(Flyer)2   -3.6692     0.6217  -5.902 1.71e-07 ***
factor(Twist)2    0.3538     0.8224   0.430 0.668520
factor(Twist)3   -1.1231     0.8224  -1.366 0.177072
factor(Twist)4   -3.2385     1.0768  -3.008 0.003822 **
factor(Block)2    2.1333     1.2941   1.648 0.104393
factor(Block)3    2.3500     1.2941   1.816 0.074298 .
factor(Block)4    3.2167     1.2941   2.486 0.015689 *
...
Residual standard error: 2.241 on 61 degrees of freedom
Multiple R-squared:  0.5711,    Adjusted R-squared:  0.4587
F-statistic: 5.077 on 16 and 61 DF,  p-value: 1.646e-06
```

   Here, the output follows the standard R convention that sets $\alpha_1 = \beta_1 = \gamma_1 = 0$, so the estimate of $\alpha_1 - \alpha_2$ is in fact 3.6692, with standard error 0.6217. Also, the `Residual standard error` is $s = 2.241$. **[7 points]**

   (b) Running the same model with `factor(Flyer)` omitted and using `anova` to compare the two models, the p-value is 1.7×10^{-7} which is obviously significant. Similarly, the models with either `Twist` or `Block` omitted from the `lm1` model specification produce p-values

0.0005857 and 0.00281 when tested against the model `lm1`. All three factors are highly significant. **[6 points]**

(c) We have $\mathrm{E}\{M_1\} = \mu + \alpha_1 + \beta_2 + (\sum \gamma_k)/13$, $\mathrm{E}\{M_2\} = \mu + \alpha_1 + \beta_3 + (\sum \gamma_k)/13$, and similar espressions for $M_3, \ldots, M_6$, so the expectation of $(M_1 + M_2 - M_5 - M_6)/2$ is $(\alpha_1 + \beta_2 + \alpha_1 + \beta_3)/2 - (\alpha_2 + \beta_2 + \alpha_2 + \beta_3)/2 = \alpha_1 - \alpha_2$. Moreover, each of $M_1, \ldots, M_6$ has variance $\sigma^2/13$ and they are all independent. Therefore, the variance of $(M_1 + M_2 - M_5 - M_6)/2$ is $4\sigma^2/52 = \sigma^2/13$. The numerical values are $M_1 = 10.8$, $M_2 = 9.277$, $M_5 = 7.085$, $M_6 = 5.654$ which lead to $\hat{\alpha}_1 - \hat{\alpha}_2 = 3.6692$.. with a standard error $s/\sqrt{13} = 0.6215$. In `lm1`, the estimate of `factor(Flyer)2` is that of $\alpha_2 - \alpha_1$, so the values are the same with a change of sign. The slight discrepancy in standard errors is due to rounding error; in fact `summary(lm1)$sigma` gives 2.241464, and this divided by $\sqrt{13}$ is 0.6217 to four decimal places. **[7 points]**

(d) The alternative Design 2 uses treatment combinations (1,1), (1,2), ..., (2,4) and if we call the eight row means $M_1, ..., M_8$, we can again calculate quantities such as $\mathrm{E}\{M_1\} = \mu + \alpha_1 + \beta_1 + (\sum \gamma_k)/13$, $\mathrm{E}\{M_2\} = \mu + \alpha_1 + \beta_2 + (\sum \gamma_k)/13$ and so on. The appropriate unbiased estimator of $\alpha_1 - \alpha_2$ is $(M_1 + M_2 + M_3 + M_4 - M_5 - M_6 - M_7 - M_8)/4$ which has variance $8\sigma^2/(16B) = \sigma^2/(2B)$ where B is the number of blocks. **[7 points]**

(e) If we have $4k$ blocks under Design 1 and $3k$ blocks under Design 2, then the respective variances are $\sigma^2/(4k)$ and $\sigma^2/(6k)$ so Design 2 is better. **[7 points]**

(f) If M_1 and M_8 have variances $3\sigma^2/B$ while the others still have variances σ^2/B, then the variance of $(M_1 + M_2 + M_3 + M_4 - M_5 - M_6 - M_7 - M_8)/4$ is now $12\sigma^2/(16B)$ which reduces to $\sigma^2/(4k)$ when $B = 3k$. Therefore, in this case Design 1 and Design 2 are the same (in the sense that the variance of $\hat{\alpha}_1 - \hat{\alpha}_2$ is the same under both models). **[6 points]**

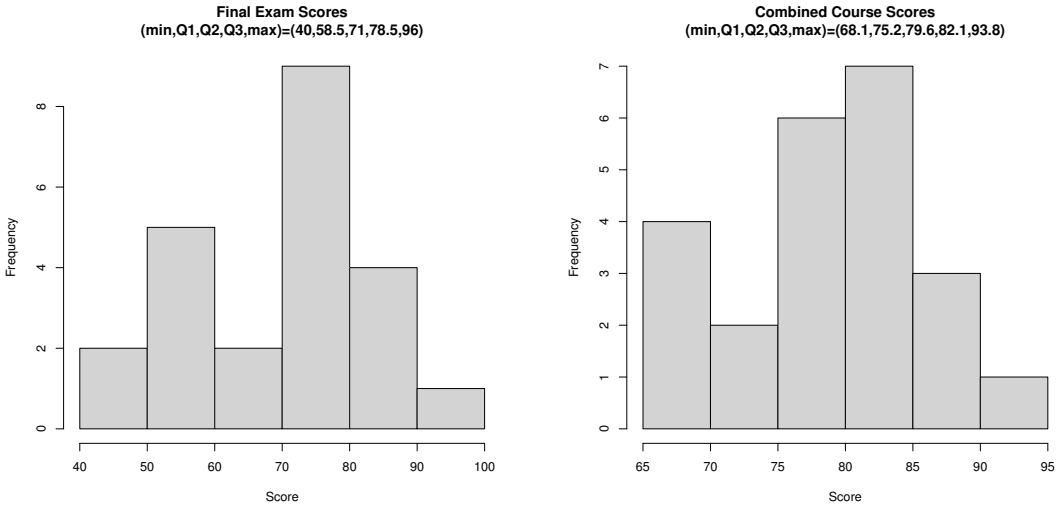

Figure 3: Histogram of Final Exam Scores and Combined Course Scores. The cutoff for H was 79.5.