MSDS 5043 - Final Assignment

Olivia Samples

10/24/2019

Question 1 (10 points)

Suppose you plan to study whether surgery can prolong life among men suffering from prostate cancer, which typically develops and spreads very slowly. Men diagnosed with prostate cancer will be randomly assigned to either undergo surgery or not. Suppose you believe that approximately 10% of men diagnosed with prostate cancer but do not have surgery will eventually die from prostate cancer, and you want to do the study using a sample size that will retain at least 80% power to detect a drop down to 5%, using a two-sided approach with a 95% confidence level.

What is the smallest number of men (including both the surgery and non-surgery groups) that you will need to enroll in this new study to meet these specifications, using a balanced design? Provide the details of your calculation, and also provide the intepretion of the results.

```
power.prop.test(p1 = .10, p2 = 0.05, sig.level=.05, power=0.80)
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 434.432
##              p1 = 0.1
##              p2 = 0.05
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
434.432 *2
```

```
## [1] 868.864
```

So we have that the smallest number of men for one group is 434.432, or 435 men. In other words, in order to obtain an balanced experiment with 80% power and a 95% confidence interval, we must obtain a sample size of 869 men. This will also ensure that the proportion of men without surgery that will die from prostate cancer is 10%, and the latter is 5%.

Description for Questions 2-6

A study of the effects of carbon monoxide exposure on men with coronary artery disease subjected the patients to several exercise tests. The men involved in the study were recruited from three different medical centers. Before combining the subjects into one large group to conduct the analysis, we need to first examine whether baseline characteristics of the subjects from the three medical centers (21 from Johns Hopkins University, 16 from Rancho Los Amigos, and 23 from St. Louis University) are comparable. Here, we examine pulmonary function at the start of the study, and we've pre-planned pairwise comparisons across all combinations of the three centers.

For each of the 60 subjects, we have a measure of forced expiratory volume in 1 second (FEV1, in liters) stored in the coexpose1.csv and the coexpose2.csv files available on Canvas.

Question 2 (10 points)

The same data appear in the coexpose1.csv and the coexpose2.csv files. What is the difference between the two files, and which of the two files is more useful for fitting an ANOVA to compare the FEV1 means across the three medical centers?

The coexpose1.csv file shows the FEV1 data for each student within three columns corresponding to each college. This data is not as useful as the coexpose2.csv file because it does not have equivalent amount of rows for each column, and you cannot reference each student with an id to compare its college and FEV1 as easily. Hence, coexpose2.csv is the more useful file.

Produce a numerical summary to compare the means across the three centers.

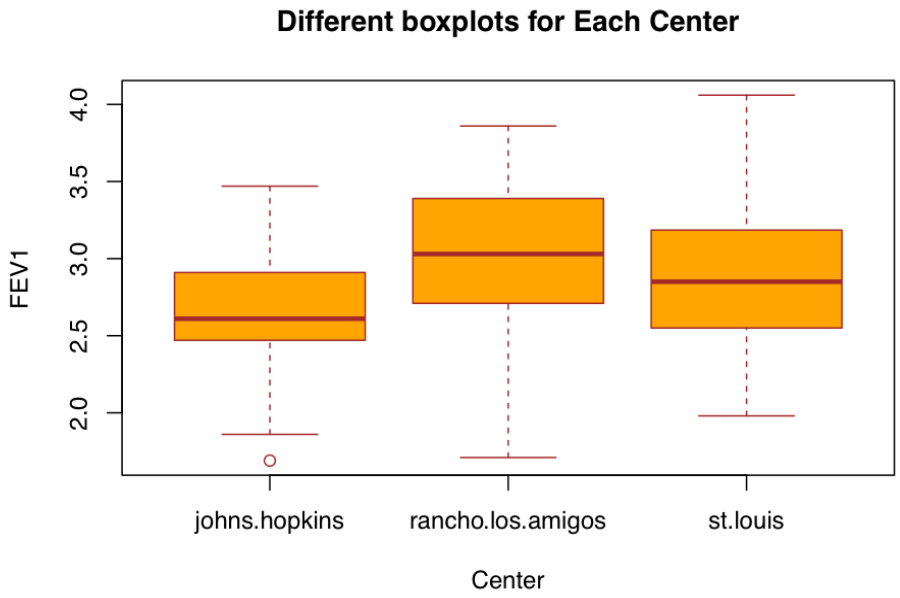
```
coexpose <- read.csv("/Users/osamples/Desktop/MSDS5043/coexpose2.csv", head = TRUE, sep =",")

tapply(coexpose$fev1, coexpose$center, mean)
```

```
##     johns.hopkins rancho.los.amigos          st.louis
##          2.626190          3.032500          2.878696
```

Question 3 (10 points)

Produce a graphical summary to compare the three centers that allows you to assess the Normality and Equal Variances assumptions of an ANOVA to compare the FEV1 means across the three medical centers. What conclusion do you draw about the assumptions in this setting?

```
boxplot(coexpose$fev1 ~ coexpose$center,
main="Different boxplots for Each Center",
xlab="Center",
ylab="FEV1",
col="orange",
border="brown"
)
```



While each center does not necessarily have the same mean, we can see from the graph that the inter-quartile ranges are relatively the same as well as their full range. In other words, we can assume equal variences. Also, none of the centers have extremely skewed box plots to the left or right so we can assume normality.

Question 4 (10 points)

Compare the FEV1 means of the three different medical centers using an analysis of variance. What conclusion do you draw, using a 90% confidence level?

H_0 : the mean is the same across all groups

H_A : at least one mean is different

```
library(pander)
pander(anova(lm(fev1 ~ center, data = coexpose)))
```

Table 1: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
center	2	1.583	0.7914	3.115	0.052
Residuals	57	14.48	0.254	NA	NA

And so we have a p-value of 0.052 and a significant value of 0.10. In other words, our p-value is less than our significant value so that we will reject our null hypothesis in favor of our alternative hypothesis: at least one of the means is different.

Question 5 (10 points)

Specify the linear model regression equation used to predict our FEV1 outcome on the basis of medical center. What fraction of the variation in FEV1 levels is explained by the medical center?

```
pander(summary(lm(fev1 ~ center, data = coexpose)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.626	0.11	23.88	2.29e-31
centerrancho.los.amigos	0.4063	0.1673	2.429	0.0183
centerst.louis	0.2525	0.1521	1.66	0.1024

Table 3: Fitting linear model: fev1 ~ center

Observations	Residual Std. Error	R^2	Adjusted R^2
60	0.504	0.09854	0.06691

We have that our linear model regression equation is as follows: $y = 2.626 + 0.4063x_1 + 0.2525x_2$, where x_1 represents the variable for Rancho Los Amigos, x_2 represents the variable for St. Louis, and where 2.626 is our intercept. We have that our R^2 quantifies the fraction of variation in FEV1 levels that is explained by the medical center. Our R^2 here is 0.09854, and so 9.85% of the variation can be explained by this model.

Question 6 (10 points)

This is a pre-planned comparison, but the sample sizes differ across the groups being compared. Obtain the results from a Bonferroni approach for pairwise comparisons of the population FEV1 means using a 90% confidence level. What is your conclusion?

```
pairwise.t.test(coexpose$fev1, coexpose$center, p.adjust.method = 'bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  coexpose$fev1 and coexpose$center
##
##                   johns.hopkins rancho.los.amigos
## rancho.los.amigos 0.055         -
## st.louis          0.307         1.000
##
## P value adjustment method: bonferroni
```

```
#calculate the adjusted sigma with K = k(k-1)/2
sigma <- .10/3
sigma
```

```
## [1] 0.03333333
```

Here we have an adjusted significance level of 0.0333. We would have had a statistically significant p-value of 0.055 for the comparison between Rancho los Amigos and Johns Hopkins without the Bonferroni adjustement. However, we do not have enough evidence to reject the null hypothesis which means that there is no difference in FEV1 mean comparisons between different groups.

Description for Questions 7-10

Low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have

investigated the determinants of plasma concentrations of these micronutrients. A cross-sectional study was designed to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol and/or beta-carotene. Study subjects (n = 300) were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous.

Variable Description

id - Subject identification number - just a code age - Subject's age (in years) sex - Subject's gender (1 = male, 2 = female) smoking - Smoking Status (1 = never, 2 = former, 3 = current) bmi - Body-Mass Index (weight in kilograms / [height in meters]2) vitamin - Vitamin Use (1 = Yes, fairly often, 2 = Yes, not so often, 3 = No) calories - Number of Calories consumed (per day) fat - Number of grams of fat consumed (per day) fiber - Number of grams of fiber consumed (per day) alcohol - Number of alcoholic drinks consumed (per week) cholesterol - Number of milligrams of cholesterol consumed (per day) betadiet - Number of micrograms of dietary beta-carotene consumed (per day) retdiet - Number of micrograms of dietary retinol consumed (per day) betaplasma - Plasma beta-carotene (in ng/ml) retplasma - Plasma retinol (in ng/ml) holdout - Explained below (1 = hold out, 0 = do not hold out)

The plasma.csv data set is available on canvas. It contains 300 observations. You will use a subset of 275 of those observations (those for which the holdout variable is equal to 0) to fit your models, and a separate sample of the remaining 25 observations (those for which holdout = 1) to validate your model selection.

```
mydata <- read.csv("/Users/osamples/Desktop/MSDS5043/plasma.csv", head = TRUE, sep = ",")

subdata <- mydata[1:275, ]
```

The essential conclusion we are looking to make (if it is true) in the context of these data is as follows:

We conclude that there is wide variability in plasma concentrations of these micronutrients in humans, and that much of this variability is associated with dietary habits and personal characteristics. Your fundamental task is to produce and interpret a series of appropriate statistical models to help us decide whether or not these conclusions are reasonable, in the case of plasma retinol, in particular. The most important thing is to accurately reflect the data.

Question 7 (10 points)

Build and specify a model for plasma retinol. Select predictors from the demographic, behavioral and relevant dietary factors described in the data (i.e. not including id, betadiet, betaplasma or holdout.) Motivate your choice of predictors, including an assessment of the impact of collinearity

```
#model 1
m1 <- lm(retplasma ~ age + sex + smoking + bmi + vitamin + calories + fat + fiber + alcohol + cholester

summary(m1)
```

```
##
## Call:
## lm(formula = retplasma ~ age + sex + smoking + bmi + vitamin +
##     calories + fat + fiber + alcohol + cholesterol + retdiet,
##     data = subdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -468.5 -127.1  -32.5  107.3 1011.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 659.798219 145.064070   4.548 8.26e-06 ***
## age           3.060461   0.986180   3.103  0.00212 **
## sex         -90.525533  41.754918  -2.168  0.03105 *
## smoking      18.121401  19.830448   0.914  0.36165
## bmi           0.400376   2.156347   0.186  0.85284
## vitamin      -6.365070  15.522417  -0.410  0.68210
## calories      0.058274   0.067436   0.864  0.38830
## fat          -0.883465   1.067215  -0.828  0.40852
## fiber        -4.488156   3.477723  -1.291  0.19800
## alcohol      -1.447460   1.593960  -0.908  0.36466
## cholesterol  -0.139234   0.142365  -0.978  0.32897
## retdiet      -0.009902   0.023742  -0.417  0.67697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 211.1 on 263 degrees of freedom
## Multiple R-squared:  0.08898,    Adjusted R-squared:  0.05087
## F-statistic: 2.335 on 11 and 263 DF,  p-value: 0.009358
```

```
#calculate VIF values
library(car)
```

```
## Loading required package: carData
```

```
vif(m1)
```

```
##         age         sex     smoking         bmi     vitamin    calories
##    1.293682    1.252684    1.155089    1.067078    1.099029   13.479602
##         fat       fiber     alcohol cholesterol     retdiet
##    8.333067    2.087698    2.633072    2.314477    1.300960
```

```
#calculate correlation to plasma retinol
cor(subdata$retplasma, subdata[sapply(subdata, is.numeric)])
```

```
## Warning in cor(subdata$retplasma, subdata[sapply(subdata, is.numeric)]):
## the standard deviation is zero
```

```
##              id       age        sex    smoking           bmi     vitamin
## [1,] 0.04327865 0.2333212 -0.1689916 0.03814855 -0.0005006015 -0.01267159
##         calories         fat       fiber     alcohol cholesterol
## [1,] -0.09422908 -0.09725156 -0.07447097 0.006445016 -0.09410956
##         betadiet     retdiet betaplasma retplasma holdout
## [1,] -0.02703203 -0.06832194 0.08544743         1      NA
```

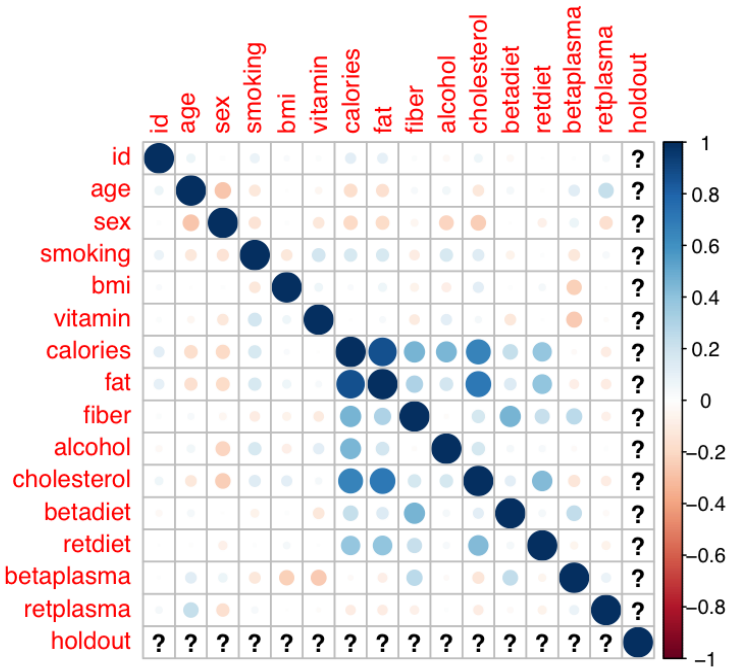
```
#variable correlation plot
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
M <- cor(subdata)
```

```
## Warning in cor(subdata): the standard deviation is zero
```

```
corrplot(M, method = "circle")
```



We say the predictor variables are collinear when they are correlated or when they have high VIF values. We have shown the collinearity between every single data item compared to plasma retinol. It is clear that the variable for calories has a rather high VIF, so we will remove them from the linear model and compare if it impacts the model.

```
#model 2
m2 <- lm(retplasma ~ age + sex + smoking + bmi + vitamin + fat + fiber + alcohol + cholesterol + retdie
summary(m2)
```

```
##
## Call:
## lm(formula = retplasma ~ age + sex + smoking + bmi + vitamin +
##     fat + fiber + alcohol + cholesterol + retdiet, data = subdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -473.78 -127.33  -35.72  112.30 1005.05
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 681.623526 142.779908   4.774    3e-06 ***
## age           2.820350   0.945772   2.982  0.00313 **
## sex         -90.661708  41.734587  -2.172  0.03072 *
## smoking      19.153917  19.784921   0.968  0.33388
## bmi           0.520032   2.150864   0.242  0.80914
```

```
## vitamin      -8.546287  15.308462  -0.558  0.57713
## fat          -0.098764   0.560396  -0.176  0.86024
## fiber        -2.506656   2.613418  -0.959  0.33836
## alcohol      -0.401441   1.036540  -0.387  0.69885
## cholesterol  -0.121504   0.140811  -0.863  0.38898
## retdiet      -0.008627   0.023685  -0.364  0.71596
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 211 on 264 degrees of freedom
## Multiple R-squared:  0.08639,    Adjusted R-squared:  0.05178
## F-statistic: 2.496 on 10 and 264 DF,  p-value: 0.007062
```

Here, the adjusted R^2 is larger than that of the first model. And so, this is the model we will start with.

Question 8 (10 points)

Conduct model selection, what is your final model? Specify and demonstrate the impact of your model selection algorithm. use only the 275 observations where holdout is 0.

We will use the Stepwise Algorithm by AIC.

```
step(m2)
```

```
## Start:  AIC=2954.32
## retplasma ~ age + sex + smoking + bmi + vitamin + fat + fiber +
##     alcohol + cholesterol + retdiet
##
##               Df Sum of Sq      RSS    AIC
## - fat          1      1383 11755951 2952.3
## - bmi          1      2603 11757171 2952.4
## - retdiet      1      5908 11760476 2952.5
## - alcohol      1      6678 11761247 2952.5
## - vitamin      1     13877 11768445 2952.6
## - cholesterol  1     33152 11787720 2953.1
## - fiber        1     40961 11795530 2953.3
## - smoking      1     41730 11796298 2953.3
## <none>                     11754568 2954.3
## - sex          1    210116 11964684 2957.2
## - age          1    395946 12150514 2961.4
##
## Step:  AIC=2952.35
## retplasma ~ age + sex + smoking + bmi + vitamin + fiber + alcohol +
##     cholesterol + retdiet
##
##               Df Sum of Sq      RSS    AIC
## - bmi          1      2464 11758415 2950.4
## - retdiet      1      6691 11762643 2950.5
## - alcohol      1      7300 11763251 2950.5
## - vitamin      1     13840 11769791 2950.7
## - smoking      1     40553 11796505 2951.3
## - fiber        1     48396 11804347 2951.5
## - cholesterol  1     65651 11821603 2951.9
## <none>                     11755951 2952.3
## - sex          1    209103 11965054 2955.2
## - age          1    409471 12165423 2959.8
```

```
##
## Step:  AIC=2950.41
## retplasma ~ age + sex + smoking + vitamin + fiber + alcohol +
##     cholesterol + retdiet
##
##               Df Sum of Sq      RSS    AIC
## - retdiet      1      6710 11765125 2948.6
## - alcohol      1      8208 11766622 2948.6
## - vitamin      1     12942 11771357 2948.7
## - smoking      1     38525 11796940 2949.3
## - fiber        1     51134 11809549 2949.6
## - cholesterol  1     63338 11821753 2949.9
## <none>                     11758415 2950.4
## - sex          1    208538 11966953 2953.2
## - age          1    410679 12169094 2957.8
##
## Step:  AIC=2948.57
## retplasma ~ age + sex + smoking + vitamin + fiber + alcohol +
##     cholesterol
##
##               Df Sum of Sq      RSS    AIC
## - alcohol      1      7768 11772893 2946.8
## - vitamin      1     13235 11778360 2946.9
## - smoking      1     39329 11804453 2947.5
## - fiber        1     58838 11823963 2947.9
## <none>                     11765125 2948.6
## - cholesterol  1     98301 11863426 2948.8
## - sex          1    212508 11977633 2951.5
## - age          1    405993 12171118 2955.9
##
## Step:  AIC=2946.75
## retplasma ~ age + sex + smoking + vitamin + fiber + cholesterol
##
##               Df Sum of Sq      RSS    AIC
## - vitamin      1     14801 11787694 2945.1
## - smoking      1     35741 11808634 2945.6
## - fiber        1     57784 11830677 2946.1
## <none>                     11772893 2946.8
## - cholesterol  1    107272 11880165 2947.2
## - sex          1    205434 11978327 2949.5
## - age          1    401266 12174158 2954.0
##
## Step:  AIC=2945.09
## retplasma ~ age + sex + smoking + fiber + cholesterol
##
##               Df Sum of Sq      RSS    AIC
## - smoking      1     29964 11817658 2943.8
## - fiber        1     53061 11840755 2944.3
## <none>                     11787694 2945.1
## - cholesterol  1    107297 11894990 2945.6
## - sex          1    195738 11983432 2947.6
## - age          1    410159 12197853 2952.5
##
## Step:  AIC=2943.79
```

```
## retplasma ~ age + sex + fiber + cholesterol
##
##               Df Sum of Sq      RSS    AIC
## - fiber        1     63867 11881525 2943.3
## <none>                     11817658 2943.8
## - cholesterol  1     97943 11915601 2944.1
## - sex          1    226090 12043748 2947.0
## - age          1    386260 12203918 2950.6
##
## Step:  AIC=2943.27
## retplasma ~ age + sex + cholesterol
##
##               Df Sum of Sq      RSS    AIC
## <none>                     11881525 2943.3
## - cholesterol  1    131504 12013029 2944.3
## - sex          1    228985 12110510 2946.5
## - age          1    369448 12250973 2949.7
##
## Call:
## lm(formula = retplasma ~ age + sex + cholesterol, data = subdata)
##
## Coefficients:
## (Intercept)          age          sex  cholesterol
##    683.9873       2.6531     -91.8560      -0.1692
```

And so, we have that the most effective variables for predicting plasma retinol are age, sex, and cholesterol.

Question 9 (10 points)

Summarize the findings in a clear presentation of your final model, including a short recap of the steps you took to produce it. Demonstrate the utility of the final model, including summaries based on R2 and significance testing. Use only the 275 observations where holdout is 0.

After a thorough analysis of correlation and performing a stepwise algorithm using AIC, we were able to eliminate every variable except for age, sex, and cholesterol in order to predict plasma retinol concentrations in patients. Our final model is as follows:

```
fm <- lm(retplasma ~ age + sex + cholesterol, data = subdata)
summary(fm)
```

```
##
## Call:
## lm(formula = retplasma ~ age + sex + cholesterol, data = subdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444.32 -127.32  -33.61  114.78 1017.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 683.98730  109.99444   6.218 1.89e-09 ***
## age           2.65314    0.91398   2.903   0.0040 **
## sex         -91.85600   40.19348  -2.285   0.0231 *
## cholesterol  -0.16917    0.09768  -1.732   0.0844 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.4 on 271 degrees of freedom
## Multiple R-squared:  0.07652,    Adjusted R-squared:  0.0663
## F-statistic: 7.485 on 3 and 271 DF,  p-value: 7.85e-05
```

This gives us an adjusted $R^2 = 0.0663$ greater than that of any other models we had. We have that age and sex have p-values < 0.05 which means that are statistically significant at a confidence level of 95%. Cholesterol is also significant when comparing it at a confidence level of 90%. And so, we have the strongest R^2 values and the statistically significant p-values. Hence, our final model is strong, and can be written as follows: $\hat{y} = 683.9873 + 2.6531x_1 - 91.8560x_2 - 0.1692x_3$, where \$x_1 = \$ age, x_2 =sex, and \$x_3 = \$ cholesterol.

Question 10 (10 points)

Validate your choice of model for plasma retinol level by using your final model to predict data for the 25 cases that you have withheld from the data, comparing your final model to these two other models:

A model using age and sex alone. A model that uses the entire set (kitchen sink) of possible predictors, i.e. everything but id, betadiet and betaplasma. Which model looks best in your comparison? Justify your response.

First, we will use our final model to predict the data for the test data.

```
testdata <- mydata[276:300, ]
pred1 <- predict(fm, testdata)
```

Now we will predict data with a model using age and sex alone.

```
pred2 <- predict(lm(retplasma ~ age + sex, data = subdata), testdata)
```

Now we will predict data with a model using the entire set (kitchen sink) of possible predictors, i.e. everything but id, betadiet and betaplasma.

```
pred3 <- predict(lm(retplasma ~ age + sex + smoking + bmi + vitamin + calories + fat + fiber + alcohol
```

We will provide the summary statistics for each prediction and the actual test data for plasma retinol.

```
#prediction from final linear model
summary(pred1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   495.8   568.6   609.5   610.6   636.2   730.6
```

```
#prediction from model using age and sex alone
summary(pred2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   501.9   558.5   591.2   606.2   638.9   737.5
```

```
#prediction from model using kitchen sink method
summary(pred3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   505.4   582.4   607.4   611.0   649.8   746.0
```

```
#prediction from real test data for plasma retinol
summary(testdata$retplasma)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   378.0   562.0   602.0   627.5   709.0   927.0
```

We will use the MSE and SSE metrics to compare models.

```
library(Metrics)
```

```
#Mean Squared Error
avp1 <- mse(testdata$retplasma, pred1)
avp1
```

```
## [1] 21185.76
```

```
avp2 <- mse(testdata$retplasma, pred2)
avp2
```

```
## [1] 20921.7
```

```
avp3 <- mse(testdata$retplasma, pred3)
avp3
```

```
## [1] 20110.69
```

```
#Sum of Squared Errors
savp1 <- sse(testdata$retplasma, pred1)
savp1
```

```
## [1] 529644.1
```

```
savp2 <- sse(testdata$retplasma, pred2)
savp2
```

```
## [1] 523042.4
```

```
savp3 <- sse(testdata$retplasma, pred3)
savp3
```

```
## [1] 502767.3
```

We also have the following accuracy results.

```
library(forecast)
```

```
##
## Attaching package: 'forecast'
```

```
## The following object is masked from 'package:Metrics':
##
##     accuracy
```

```
accuracy(pred1, testdata$retplasma)
```

```
##                ME     RMSE      MAE       MPE     MAPE
## Test set 16.92808 145.5533 109.9879 -2.547442 18.45879
```

```
accuracy(pred2, testdata$retplasma)
```

```
##               ME     RMSE      MAE      MPE     MAPE
## Test set 21.3274 144.6433 108.0839 -1.64617 17.88973
```

```
accuracy(pred3, testdata$retplasma)
```

```
##                ME     RMSE      MAE       MPE     MAPE
## Test set 16.52798 141.8122 113.5186 -2.453326 18.96497
```

Percentage errors have the advantage of being unit-free, and so are frequently used to compare forecast performances between data sets. We have that our Mean Absolute Percentage error for our final model is the smallest. Also, when comparing the summary statistics, the quartiles seem to follow a closer comparison to the real statistics when using pred1, the final model predictors, rather than pred2 or pred3.

However, our model is the largest when it comes to MSE and SSE. While we do have less variables to keep track of in the final model, it does not provide the most optimal predictions as the other two models. Still, they are each relatively similar. Since the kitchen sink model has so many variables, we will eliminate its usefulness. Also, the model using age and sex alone is not as accurate when comparing percentage errors, so we will eliminate it as well. While no model is a perfect model, we will use our final model because of its relative usefulness and accuracy.