COMPUTATIONAL ASSIGMENT #3
BRANDON MORETZ

VARIABLES

1.) *For all the categorical variables in the dataset, recode the text-based categories into numerical values that indicate group. For example, for the VITAMIN variable, you could code it so that: 1=regular, 2=occasional, 3=never. Save the categorical variables to the dataset.*

ID	Age	Smoke	Quetelet	Calories	Fat	Fiber	Alcohol	Cholesterol	BetaDiet	RetinolDiet	BetaPlasma	RetinolPlasma	Gender	VitaminUse	PriorSmoke	VitaminCoded	VitaminCoded2	GenderCoded	SmokeCoded	PriorSmokeCoded
1	64	No	21.4838	1298.8	57.0	6.3	0.0	170.3	1945	890	200	915	Female	Regular	2	1	3	0	0	2
2	76	No	23.8763	1032.5	50.1	15.8	0.0	75.8	2653	451	124	727	Female	Regular	1	1	3	0	0	1
3	38	No	20.0108	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721	Female	Occasional	2	2	2	0	0	2
4	40	No	25.1406	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615	Female	No	2	3	1	0	0	2
5	72	No	20.9850	1952.1	82.6	16.2	0.0	170.8	2863	1209	92	799	Female	Regular	1	1	3	0	0	1
6	40	No	27.5214	1366.9	56.0	9.6	1.3	154.6	1729	1439	148	654	Female	No	2	3	1	0	0	2
7	65	No	22.0115	2213.9	52.0	28.7	0.0	255.1	5371	802	258	834	Female	Occasional	1	2	2	0	0	1
8	58	No	28.7570	1595.6	63.4	10.9	0.0	214.1	823	2571	64	825	Female	Regular	1	1	3	0	0	1
9	35	No	23.0766	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517	Female	No	1	3	1	0	0	1
10	55	No	34.9699	1263.6	39.6	15.5	0.0	171.9	3307	493	81	562	Female	No	2	3	1	0	0	2
11	66	No	20.9465	1460.8	58.0	18.2	1.0	137.4	1714	535	184	935	Female	Regular	2	1	3	0	0	2

VITAMIN USE

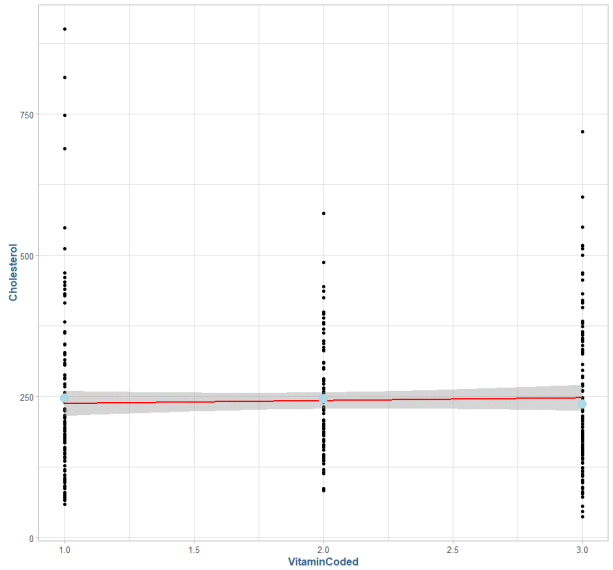
2.) *For the VITAMIN categorical variable, fit a simple linear model that uses the categorical variable to predict the response variable Y=CHOLESTEROL.*

a.) *Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics.*

Cholesterol ~ VitaminCoded

Model 1: 232.634 + 5.001β_1, where β_1 is the level of vitamin use [1=regular, 2=occasional, 3=no]

We note the positive coefficient term in the model, indicating that for each increase in the vitamin coded value, there is an associated positive increase in cholesterol (by approximately 5 points per level).



The R^2 for model 1 is 0.001, suggesting that the amount of variance explained by the model is about .1%, which is almost none. We note in the previous chart where we fitted a linear model to the data using this coded variable, we have a straight line that comes close to the means of each category (the SE does account for the values of the true group means).

Above we have the group cholesterol values, with a blue horizontal bar denoting the sample mean, and the blue dots indicating the individual group means. The amount of variance in these data, and the heavy number of outliers indicate a poor fit.

The null hypothesis in this case would be,

$H_{0:} \beta_1 = 0$

Or that there is no effect on the model using the beta coefficient derived from the vitamin coded variable, against the alternative hypothesis that:

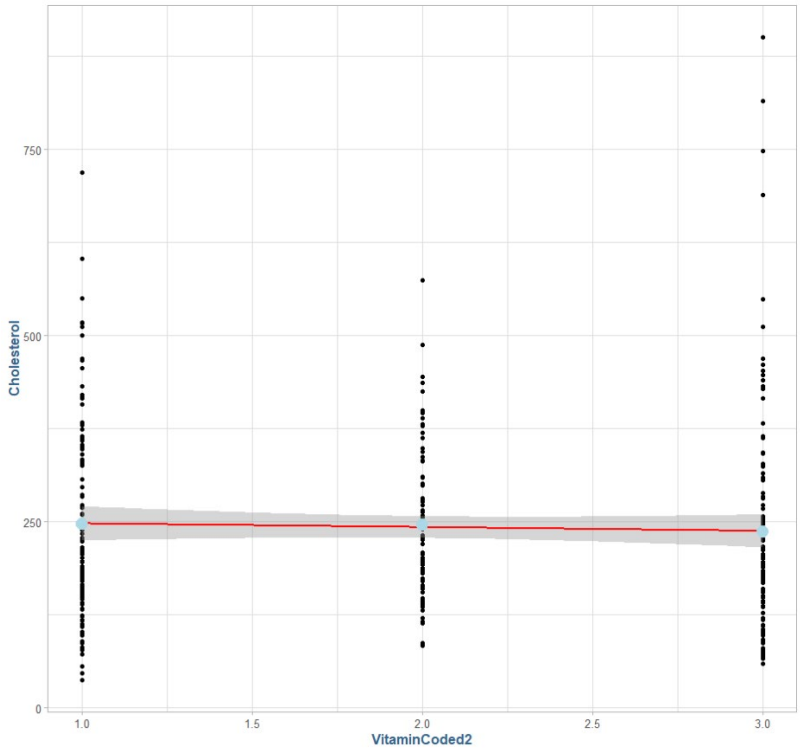
$H_{a:} \beta_1 \neq 0$

Or that there is additional variance explained in the data by including the beta 1 coefficient. In our model summary, the p-value of 0.564 for our vitamin coded variable suggests that there is no statistically significant difference when using the beta coefficient in question.

b.) *Recode the VITAMIN categorical variable so that you have a different set of indicator values. For example, you could code it so that: 1=never, 2=occasional, 3=regular. Re-fit.*

Cholesterol ~ VitaminCoded2

Model 2: 252.637 - 5.001β_1, where β_1 is the level of vitamin use [3=regular, 2=occasional, 1=no]



The model has adjusted for the value encoding, with the intercept value increasing by 20 points and the beta coefficient is now negative, indicating that no vitamin use has a higher cholesterol value, and that for each level of vitamin use (2 , 3), we subtract 5.001 cholesterol points. We can see a negative linear trend in the preceding diagram.

3.) *Create a set of dummy coded (0/1) variables for the VITAMIN categorical variable. Fit a multiple regression model using the dummy coded variables to predict CHOLESTEROL (Y). Remember, you need to leave one of the dummy coded variables out of the equation. That category becomes the "basis of interpretation." Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. Compare the findings here to those in task 2). What has changed?*

Model 3: $\hat{y} = 246.599 - 1.156\beta_1 - 9.908\beta_2$

Here, we see that the intercept term is 246.599, which is the predicted value when all beta coefficient terms are zero. This is identical to the mean of the data set when vitamin use (**VitaminUse**) is equal to zero. The coefficients in this context represent the relative delta in means for each of the vitamin groups:

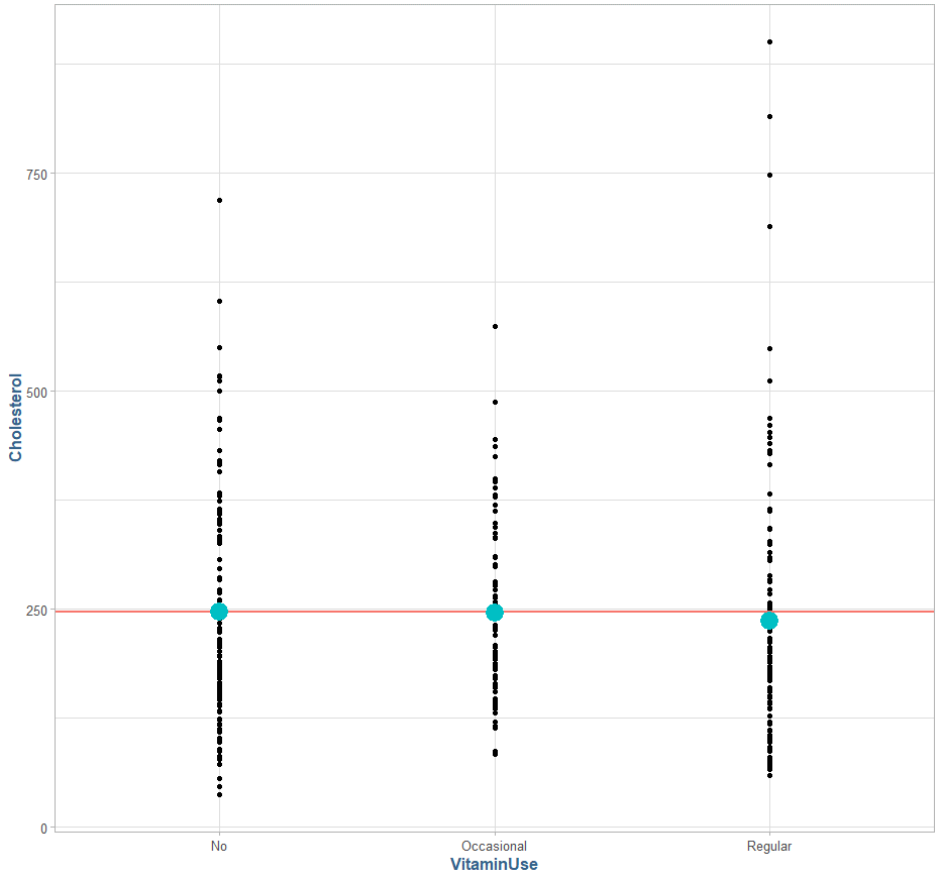
No Vitamin = $\bar{y}$ = 246.599, average cholesterol

Occasional = $\bar{y}$ - 1.156, decreases cholesterol 1.156 points

Regular = $\bar{y}$ - 9.998, decreases cholesterol 9.998 points

Our R^2 for this model is **0.0012**, which indicates that approximately .12% of the variance explained in the data is accounted for by this model.

Visually, we can see a scatterplot of the data to the right, with the red line indicative of the overall mean of the data, agnostic to vitamin use. The blue dots represent the group mean for that category of vitamin use.



4.)

The null hypothesis in this case would be,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu$$

Or that there is no difference in the individual category means compared to the overall (unknown) population mean. Compared to the alternative,

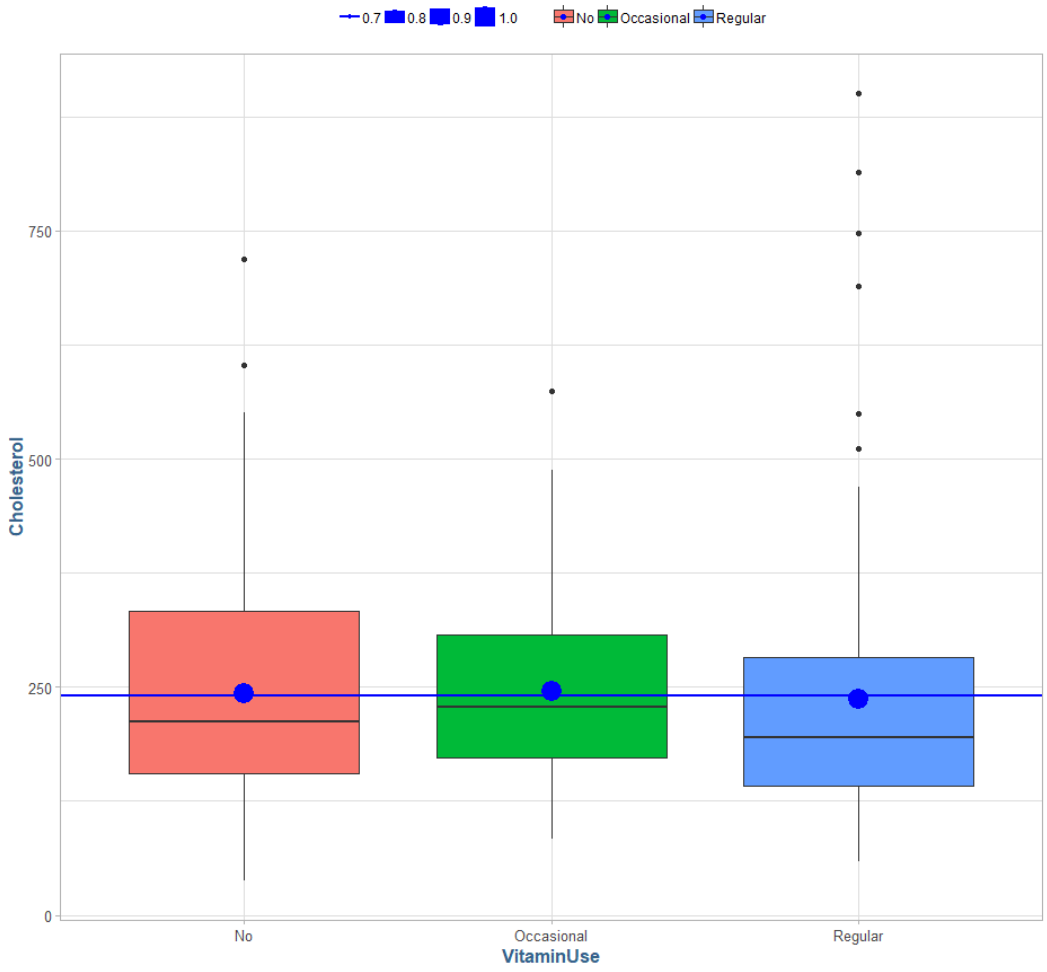
$$H_a: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu$$

Or plainly, at least one of the group means is statistically different than the overall (unknown) population mean. Given the overall variance and heavy presence of outliers in the data across groups, this does not look to be a useful model as it stands.

4.) *For the VITAMIN categorical variable, use the NEVER categorical as the control or comparative group, and develop a set of indicator variables using effect coding. Save these to the dataset. Fit a multiple regression model using the dummy coded variables to predict CHOLESTEROL(Y). Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. Compare the findings here to those in task 3). What has changed? Which do you prefer? Why?*

For effect coding, we will choose the coding scheme [(1,0)=regular, (0,1)=occasional, (-1,-1)=no] since we are controlling for no vitamin use.

Model 4: $242.911 + 2.532\beta_1 - 6.220\beta_2$. where β_1 indicates the difference between the overall mean and occasional vitamin use (the first factor mean), and β_2 indicates the difference between the overall mean and regular vitamin usage (second factor level). Below we can see a boxplot of the three groups, with model estimated mean for each group represented by blue dots inside the box. The R^2 of this model is .1223, or it accounts for about approximately .1% of the overall variance in the data. The blue line in the figure below is representative of the y-intercept in the model, noting it is significantly higher than any of the individual means.



The null hypothesis in this case would be,

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu$ (unknown)

Or that there is no difference in the individual category means compared to the overall (unknown) population mean. Compared to the alternative,

$H_a: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu$ (unknown)

Or at least one of the group means is statistically different than the overall (unknown) population mean. Given the overall variance and heavy presence of outliers in the data across groups, this does not look to be a useful model as it stands.

The main difference here is that the dummy coding gives us the exact means of the respective groups, while the effect coding gives us the both the main effect and the interact effect. Given the more robust interpretation that the effect encoding allows, I prefer this method of encoding.

ALCOHOL

5.) *Discretize the ALCOHOL variable to form a new categorical variable with 3 levels. The levels are:*

- *0 if ALCOHOL = 0*
- *1 if 0 < ALCOHOL < 10*
- *2 if ALCOHOL >= 10*

Use these categories to create a set of indicator variables for ALCOHOL that use effect coding. Save these to your dataset.

```
> data.interaction
     Cholesterol AlcoholHeavy AlcoholModerate VitaminOccasional VitaminRegular HO HR MO MR
  1:       170.3          -1              -1               -1              1  1 -1  1 -1
  2:        75.8          -1              -1               -1              1  1 -1  1 -1
  3:       257.9           1              -1                1             -1  1 -1 -1  1
  4:       332.6          -1               1               -1             -1  1  1 -1 -1
  5:       170.8          -1              -1               -1              1  1 -1  1 -1
 ---
311:       306.5          -1               1               -1             -1  1  1 -1 -1
312:       257.7          -1               1               -1              1  1 -1 -1  1
313:       150.5          -1               1               -1              1  1 -1 -1  1
314:       381.8          -1               1               -1              1  1 -1 -1  1
315:       195.6          -1               1               -1              1  1 -1 -1  1
>
```

6.) At this point, you should have effect coded indicator variables for VITAMIN and 2 effect coded indicator variables for ALCOHOL. Create 4 product variables by multiplying each of the effect coded indicator variables for VITAMIN by the effect coded indicator variables for ALCOHOL. This is all pairwise products of the effect coded variables.

Now, we are going to test for interaction.

Fit an OLS multiple regression model using the 4 VITAMIN and ALCOHOL effect coded indicator variables plus the 4 product variables to predict CHOLESTEROL. Call this the full model:

```
Call:
lm(formula = Cholesterol ~ AlcoholModerate + AlcoholHeavy + VitaminOccasional +
    VitaminRegular + HO + HR + MO + MR, data = data.interaction)

Residuals:
    Min      1Q  Median      3Q     Max
-246.35  -89.87  -35.32   63.46  679.84

Coefficients:
                 Estimate Std. Error t value            Pr(>|t|)
(Intercept)       263.333     19.411  13.566 <0.0000000000000002
AlcoholModerate    10.056     10.120   0.994               0.321
AlcoholHeavy       26.212     20.027   1.309               0.192
VitaminOccasional  -5.448     16.707  -0.326               0.745
VitaminRegular     12.898     17.310   0.745               0.457
HO                 -3.164     17.716  -0.179               0.858
HR                 21.261     17.868   1.190               0.235
MO                 14.990     10.648   1.408               0.160
MR                 15.115      9.345   1.618               0.107

Residual standard error: 132.1 on 306 degrees of freedom
Multiple R-squared:  0.02344, Adjusted R-squared:  -0.002091
F-statistic: 0.9181 on 8 and 306 DF,  p-value: 0.5016
```

For the Reduced model, fit an OLS multiple regression model using only the effect coded variables for VITAMIN and ALCOHOL to predict CHOLESTEROL.

```
Call:
lm(formula = Cholesterol ~ AlcoholModerate + AlcoholHeavy + VitaminOccasional +
    VitaminRegular, data = data.interaction)

Residuals:
    Min      1Q  Median      3Q     Max
-244.04  -90.70  -32.89   69.19  666.43

Coefficients:
                 Estimate Std. Error t value            Pr(>|t|)
(Intercept)      258.47042   15.47103  16.707 <0.0000000000000002
AlcoholModerate    0.40967    8.03356   0.051               0.959
AlcoholHeavy      20.17090   14.53333   1.388               0.166
VitaminOccasional  0.05365    9.64248   0.006               0.996
VitaminRegular    -3.56577    8.75408  -0.407               0.684

Residual standard error: 132.3 on 310 degrees of freedom
Multiple R-squared:  0.008069,  Adjusted R-squared:  -0.00473
F-statistic: 0.6305 on 4 and 310 DF,  p-value: 0.6411
```

Conduct a nested model F-test using the Full and Reduced Models described here. Be sure to state the null and alternative hypothesis, decide regarding the test, and interpret the result.

H_o : $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

H_a : $\beta_j \neq 0$, for at least one value of j (for j in 5, 6, 7, 8)

$$F = [(SSE_R - SSE_C) / (df2 - df1)] / (SSE_C / df1]$$

$$F = ((5{,}426{,}297 - 5{,}342{,}216) / 4) / [5{,}342{,}216 / 306]$$

$$= 21{,}020.3 / 17{,}458.22$$

$$= \mathbf{1.204}$$

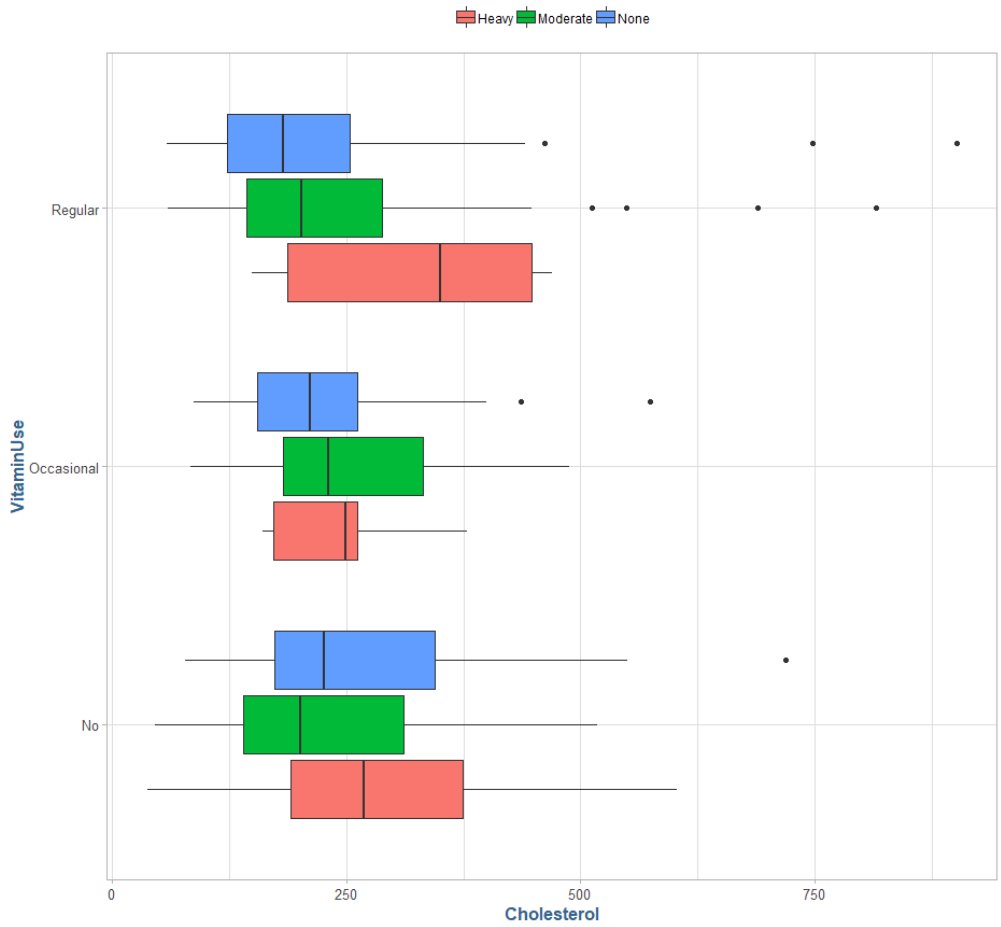
Critical value at 95% confidence ($\alpha = 0.05$), % confidence, = $F_{95,\ 4,\ 306}$= **2.401**

The given F-statistic yielded a value of **1.204** and at 95% confidence, we cannot reject the null hypothesis that the more complex, or complete, model with the additional explanatory variables β_5, β_6, β_7 and β_8 is more powerful than the reduced model. We can also look at the analysis of variance for the two models, summarizing the above results.

```
Analysis of Variance Table

Model 1: Cholesterol ~ AlcoholModerate + AlcoholHeavy + VitaminOccasional +
    VitaminRegular + HO + HR + MO + MR
Model 2: Cholesterol ~ AlcoholModerate + AlcoholHeavy + VitaminOccasional +
    VitaminRegular
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1    306 5342216
2    310 5426297 -4    -84081 1.204 0.3091
>
```

Obtain a means plot to illustrate any interaction, or lack thereof, to help explain the result.



In the preceding plot we can see alcohol use broken out by vitamin use. We can see that in each of the three vitamin groups, the heavy alcohol group has the highest overall mean cholesterol in each category. We also see that each of the alcohol groups cluster together inside their respective vitamin usage categories in terms of means, with the largest outlier being heavy alcohol in the regular vitamin usage category.

GENDER / SMOKE

For the gender and smoke variables we conduct a similar experiment that we did with the alcohol interaction terms, namely we conduct a hypothesis test using full and reduced models.

H_o : $\beta_9 = \beta_{10} = 0$
H_a : $\beta_j \neq 0$, for at least one value of j (for j in 9, 10)

$$F = [(SSE_R - SSE_C) / (df2 - df1)] / (SSE_C / df1]$$

$$F = ((5{,}342{,}216 - 5{,}017{,}925) / 2) / [5{,}017{,}925 / 304]$$

$$= 162{,}145.7 / 16{,}506.33$$

$$= \mathbf{9.8232}$$

Critical value at 95% confidence ($\alpha = 0.05$), % confidence, = $F_{95,\ 2,\ 304}$= **3.0254**

The given F-statistic yielded a value of **9.8232** and at 95% confidence, we should reject the null hypothesis that the more complex, or complete, model with the additional explanatory variables β_9 and β_{10} is more powerful than the reduced model. We can also look at the analysis of variance for the two models, summarizing the above results.

```
Analysis of Variance Table

Model 1: Cholesterol ~ AlcoholModerate + AlcoholHeavy + VitaminOccasional +
    VitaminRegular + HO + HR + MO + MR + Gender + Smoke
Model 2: Cholesterol ~ AlcoholModerate + AlcoholHeavy + VitaminOccasional +
    VitaminRegular + HO + HR + MO + MR
  Res.Df     RSS Df Sum of Sq      F     Pr(>F)
1    304 5017925
2    306 5342216 -2   -324291 9.8232 0.00007345
```

It does not appear that including the gender and smoking interaction terms has an impact to an individual's cholesterol level when accounting for vitamin and alcohol use.

CONCLUSION

In this lab I learned about various coding schemes for categorical variables and how to properly integrate these coded variables into regression models. These effects were primarily concerned with testing the effects of a given attribute has upon the mean of a set of individuals classified in their respective groups using standard regression techniques. We conducted hypothesis tests to confirm the presence (or absence) of these effects on the groups in question. Given that standard linear regression model plots, which plot the response variable vs the independent variable with residuals provide little in terms of value when looking at these categorical values, we devised some boxplot mechanics to help visualize these categorical effects. Finally, we also introduced the concept of introducing arbitrary cut-points in continuous variables in order to deduce new factorized / categorical variables for analysis.