NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Computational Assignment #3: OLS Regression Modeling with Categorical Variables
MSDS 410

1) For all of 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.

2) For the VITAMIN categorical variable, fit a simple linear model that uses the categorical variable to predict the response variable Y=CHOLESTEROL. Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. 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 an OLS simple linear model using the new categorization. Report the model, interpret the coefficients, discuss test results, etc. What is going on here?

- Model 1:
  - Y = 5.001*VitaminCode + 232.634
- Interpret the coefficients:
  - 5.001*VitaminCode: For every one numerical increase in Vitamin (ie going from regular to occasional or occasional to never) cholesterol increases 5.001.
  - 232.634: y-intercept value for when Vitamin = 0. There are no values where Vitamin can equal zero so this intercept does not make logical sense and is merely the starting point for the line of best fit.
- Discuss hypothesis test results, goodness of fit statistics:

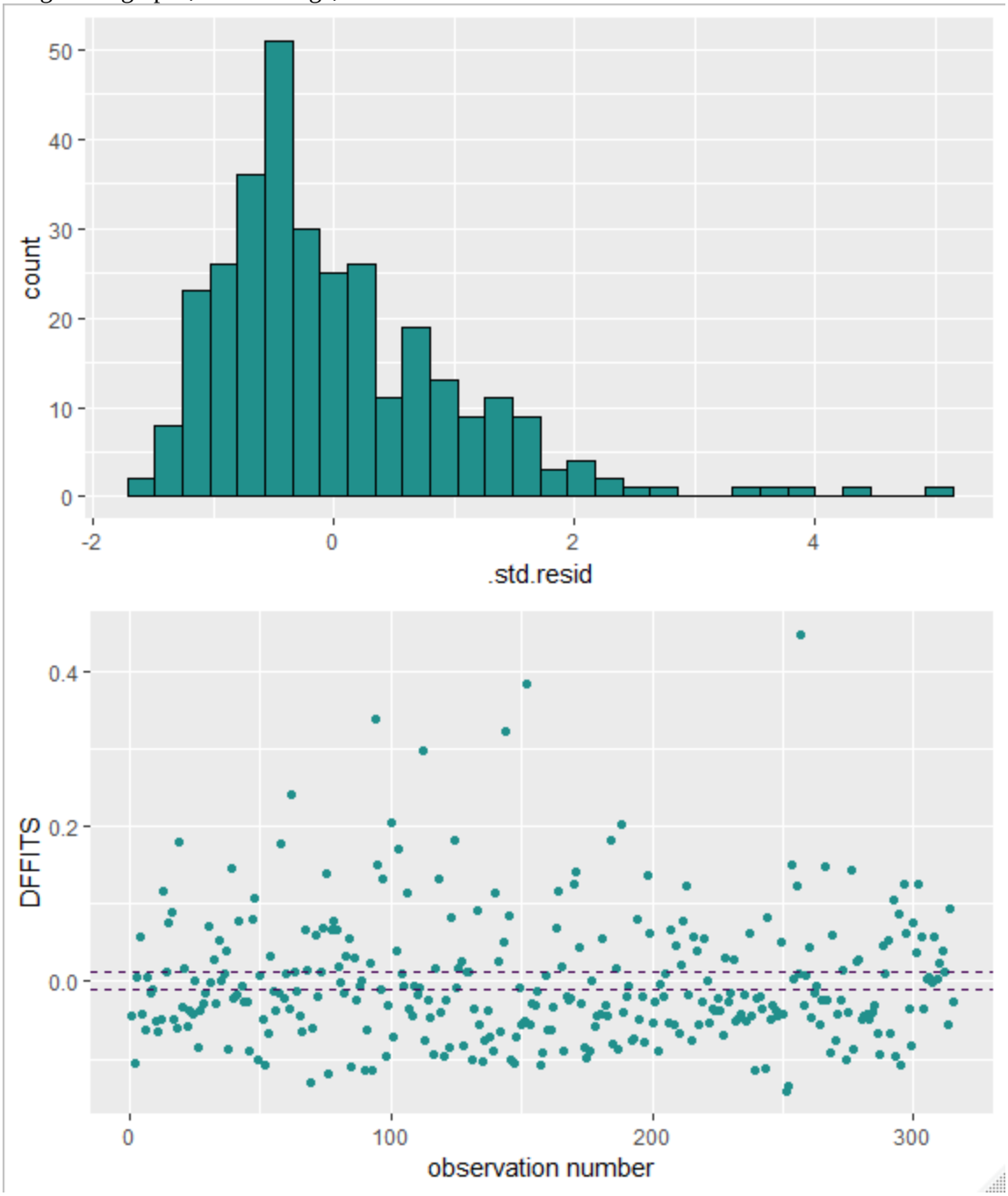
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  232.634     18.581  12.520   <2e-16 ***
VitaminCode    5.001      8.663   0.577    0.564
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 132.1 on 313 degrees of freedom
Multiple R-squared:  0.001063,  Adjusted R-squared:  -0.002128
F-statistic: 0.3332 on 1 and 313 DF,  p-value: 0.5642
```

  - H_0: $\beta_0 = 0$. We reject this null hypothesis and conclude that the y-intercept is not equal to 0. This is logical because a cholesterol value of zero is not only outside the observed dataset but also outside the possible values for humans.
  - H_0: $\beta_1 = 0$. We do not reject the null hypothesis on the Vitamin variable. We cannot conclude that the estimated effect of a change in the Vitamin value is significantly different from 0.

- The R-squared value for this model is 0.001, which means that our model accounts for almost none of the variance in Cholesterol.

- Diagnostic graphs, and leverage, influence and Outlier statistics



As seen in the histogram above, the residuals have a right skew, which violates the assumption of normality. If the DFFITs plot, the majority of the points are beyond the threshold for outliers, which means that this model is a particularly bad fit for the given inputs.

- Model 2:
  - Y = -5.001*VitaminCodeNew + 252.637
- Interpret the coefficients:

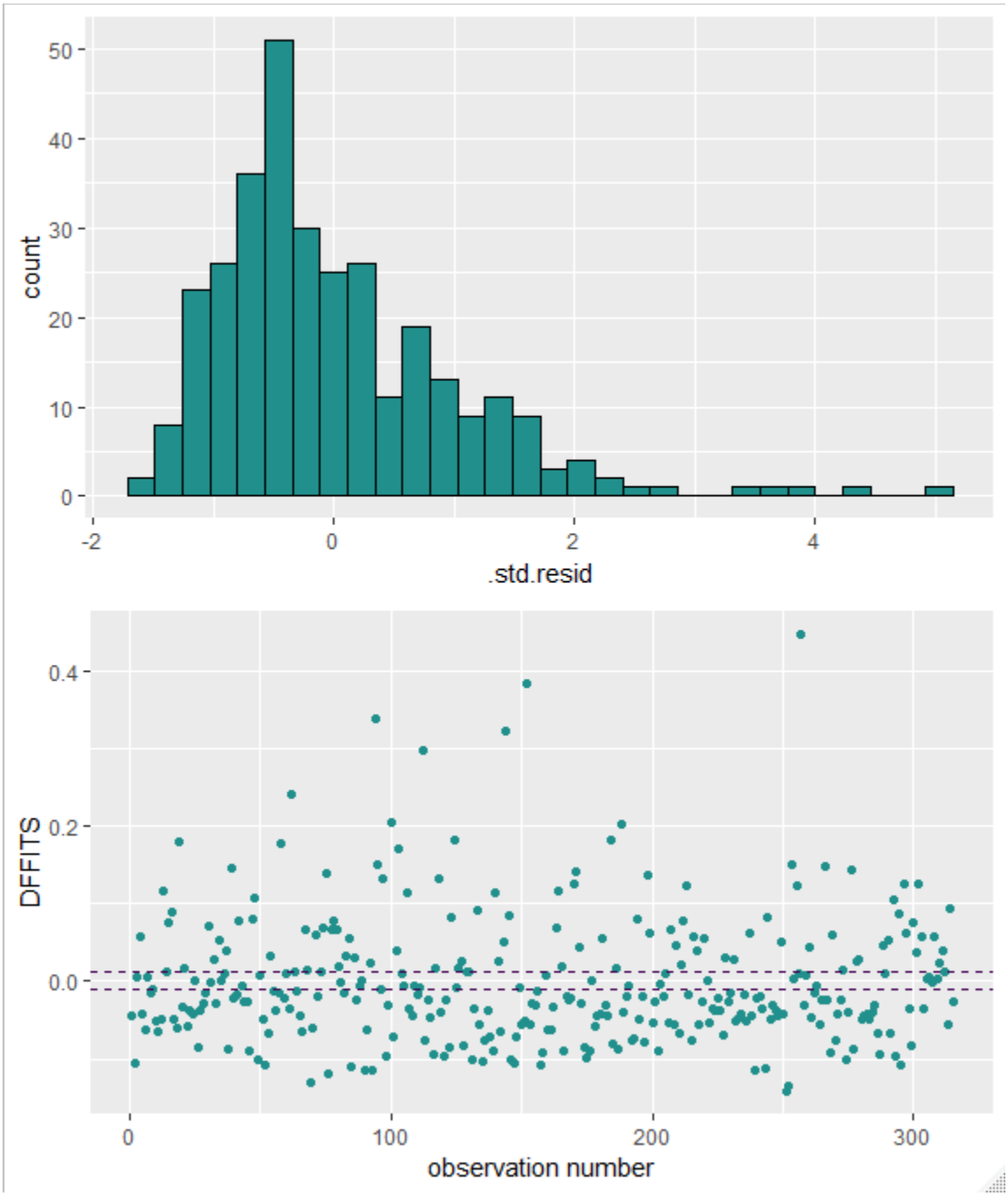
- -5.001*VitaminCode: The reverse of the original model. For every one numerical increase in Vitamin (ie going from never to occasional or occasional to regular) cholesterol decreases 5.001. Essentially the same as model1, just adjusting for the change in the order of the values.
- 252.637: y-intercept value for when Vitamin = 0. There are no values where Vitamin can equal zero so this intercept does not make logical sense and is merely the starting point for the line of best fit.

- Discuss hypothesis test results, goodness of fit statistics:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     252.637     19.137  13.202   <2e-16 ***
VitaminCodeNew   -5.001      8.663  -0.577    0.564
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 132.1 on 313 degrees of freedom
Multiple R-squared:  0.001063,  Adjusted R-squared:  -0.002128
F-statistic: 0.3332 on 1 and 313 DF,  p-value: 0.5642
```

  - H_0: $\beta_0 = 0$. We reject this null hypothesis and conclude that the y-intercept is not equal to 0. This is logical because a cholesterol value of zero is not only outside the observed dataset but also outside the possible values for humans.
  - H_0: $\beta_1 = 0$. We do not reject the null hypothesis on the Vitamin variable. We cannot conclude that the estimated effect of a change in the Vitamin value is significantly different from 0.
  - The R-squared value for this model is 0.001, which means that our model accounts for almost none of the variance in Cholesterol.
- Diagnostic graphs, and leverage, influence and Outlier statistics



We have identical results for the skewed residuals and DFFITs plot as model 1. Our model is still skewed and poorly fit.

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?

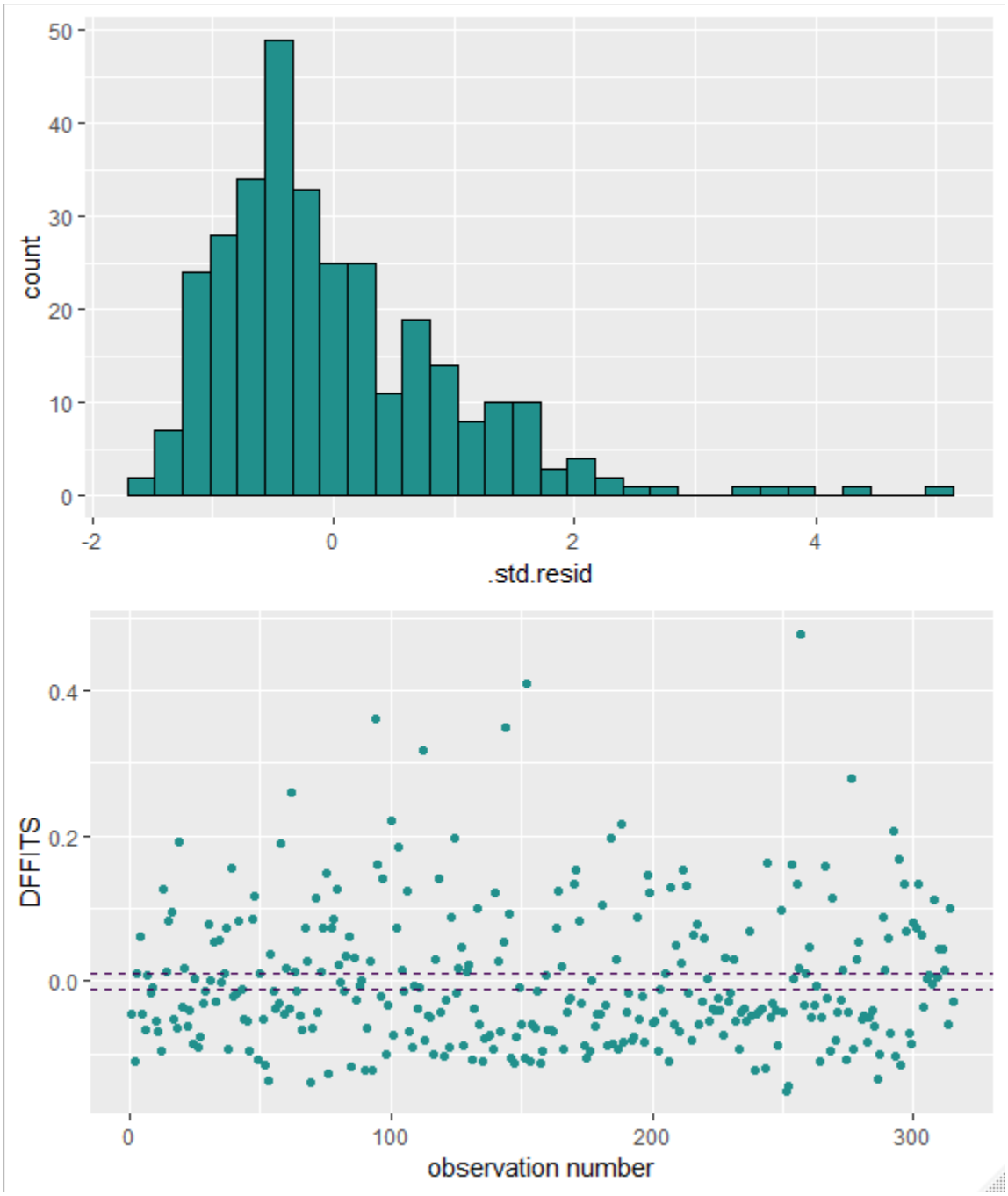
- Model 3:
    - Y = -1.156*x_occasional - 9.908*x_regular + 246.599

- Interpret the coefficients:
  - 246.599: Our y-intercept is now the predicted Cholesterol level when VitaminUse = never.
  - -1.156*x_occasional: The change from our baseline (no vitamin usage, aka the y-intercept) when VitaminUse = occasional. Given this model all occasional vitamin users would have a predicted Cholesterol value of 245.443.
  - -9.908*x_regular: The change from our baseline (no vitamin usage, aka the y-intercept) when VitaminUse = regular. Given this model all regular vitamin users would have a predicted Cholesterol value of 236.691.
- Discuss hypothesis test results, goodness of fit statistics:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   246.599     12.560  19.633   <2e-16 ***
x_occasional   -1.156     19.270  -0.060    0.952
x_regular      -9.908     17.358  -0.571    0.569
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132.3 on 312 degrees of freedom
Multiple R-squared:  0.001223,  Adjusted R-squared:  -0.005179
F-statistic: 0.1911 on 2 and 312 DF,  p-value: 0.8262
```

  - H_0: $\beta_0 = 0$. We reject this null hypothesis and conclude that the y-intercept is not equal to 0. This is logical because a cholesterol value of zero is not only outside the observed dataset but also outside the possible values for humans.
  - H_0: $\beta_1 = 0$. We do not reject the null hypothesis on the occasional Vitamin usage variable. We cannot conclude that the estimated effect of occasionally using Vitamins significantly differs from no vitamin usage.
  - H_0: $\beta_2 = 0$. We do not reject the null hypothesis on the regular Vitamin usage variable. We cannot conclude that the estimated effect of regularly using Vitamins significantly differs from no vitamin usage.
  - The R-squared value for this model is 0.001, which means that our model accounts for almost none of the variance in Cholesterol. The r-squared, F-statistic, and p-value of model 3 is equal to those of model 1 and model 2, meaning that simply rearranging the variables as dummy variables did not add predictive power to the model, it only made the model interpretation more logical.
- Diagnostic graphs, and leverage, influence and Outlier statistics



We are still seeing the same results for the residuals and DFFITs plot as models 1 and 2. Our model is still skewed and poorly fit.

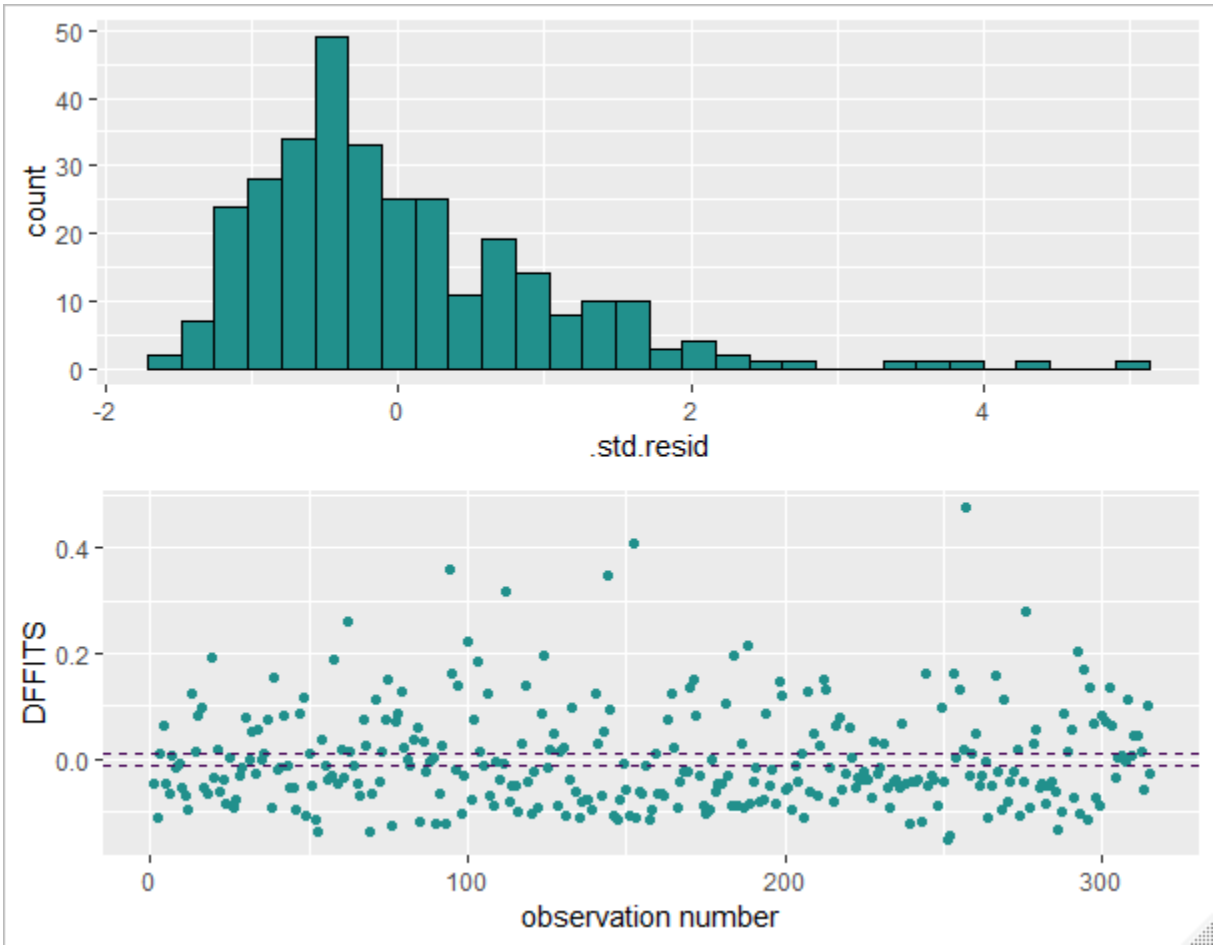
4) For the VITAMIN categorical variable, use the NEVER categorical as the control or comparative group, and develop a set of indicator variables using effect coding. Save these to the dataset. Fit a multiple regression model using the dummy coded variables to predict CHOLESTEROL(Y). Report the model, 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?
   - Model 4:
     - Y = -6.22*regular + 2.532*occasional + 242.911

- Interpret the coefficients:
  - 242.911: The y-intercept is now the grand mean of Cholesterol for the entire sample.
  - -6.22*regular: The difference between the grand mean and the mean of the regular vitamin users.
  - 2.532*occasional: The difference between the grand mean and the mean of the regular vitamin users.
- Discuss hypothesis test results, goodness of fit statistics:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        242.911      7.564  32.116   <2e-16 ***
effect_regular      -6.220     10.250  -0.607    0.544
effect_occasional    2.532     11.331   0.223    0.823
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132.3 on 312 degrees of freedom
Multiple R-squared:  0.001223,  Adjusted R-squared:  -0.005179
F-statistic: 0.1911 on 2 and 312 DF,  p-value: 0.8262
```

  - H0: $\beta_0 = 0$. We reject this null hypothesis and conclude that the y-intercept is not equal to 0. This is logical because a cholesterol value of zero is not only outside the observed dataset but also outside the possible values for humans.
  - H0: $\beta_1 = 0$. We do not reject the null hypothesis on the regular Vitamin usage variable. We cannot conclude that the estimated effect of occasionally using Vitamins significantly differs from no vitamin usage.
  - H0: $\beta_2 = 0$. We do not reject the null hypothesis on the regular Vitamin usage variable. We cannot conclude that the estimated effect of regularly using Vitamins significantly differs from no vitamin usage.
  - The R-squared value for this model is 0.001, which means that our model accounts for almost none of the variance in Cholesterol. The r-squared, F-statistic, and p-value of model 4 is equal to those of all three previous models, meaning that simply rearranging the variables as effect coded variables did not add predictive power to the model.
- Diagnostic graphs, and leverage, influence and Outlier statistics



- We observe the same residual and DFFITs charts as in all previous models, which means we are not changing the model in any significant way.
- The only real change is the way the coefficients and inputs are presented. For this example I prefer using dummy variables. The group that does not use any vitamins makes a natural control that is observable using the y-intercept and then it follows logically for the effect to show the change at each following level. I can see scenarios where there isn't such a clear cut control opportunity where it would be more informative to have the grand mean as the y-intercept. In a case like that I would prefer to use effect coding.

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

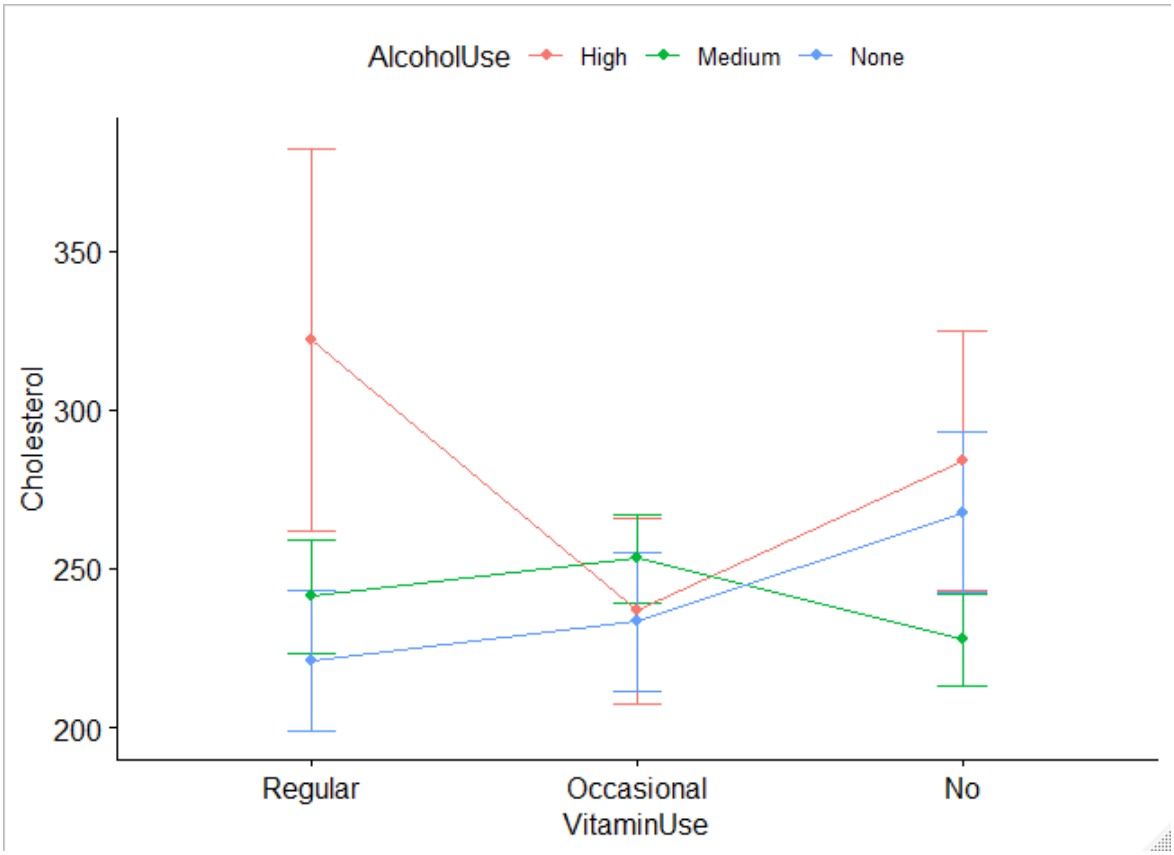
   Use these categories to create a set of indicator variables for ALCOHOL that use effect coding. Save these to your dataset.

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. For the Reduced model, fit an OLS multiple regression model using only the effect coded variables for VITAMIN and ALCOHOL to predict CHOLESTEROL. Conduct a nested model F-test using the

Full and Reduced Models described here. Be sure to state the null and alternative hypothesis, make a decision regarding the test, and interpret the result. Obtain a means plot to illustrate any interaction, or lack thereof, to help explain the result.

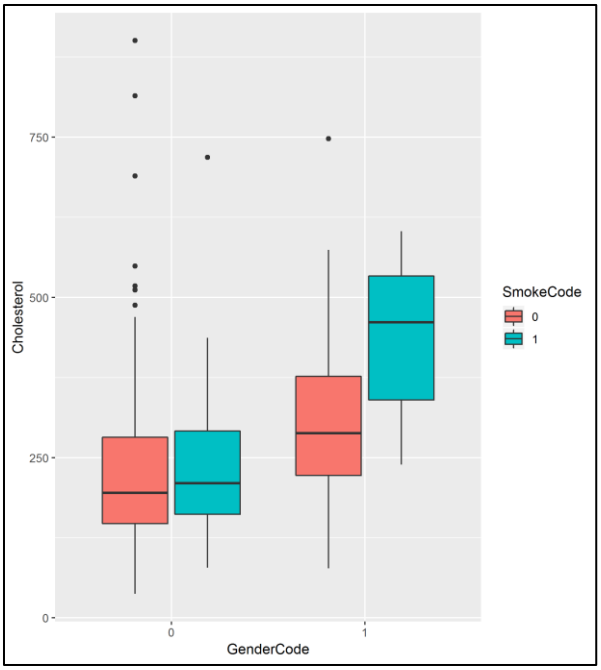
```
Model 1: Cholesterol ~ effect_regular + effect_occasional + alcohol_high +
    alcohol_medium
Model 2: Cholesterol ~ effect_regular + effect_occasional + alcohol_high +
    alcohol_medium + reg_high + reg_med + occ_high + occ_med
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1    310 5426297
2    306 5342216  4     84081 1.204 0.3091
```

- H_0: $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$
- H_A: At least one interaction beta does not equal zero
- The F value of the nested model is 1.204. We do not reject the null hypothesis, which means we accept that all the interaction variables are not significantly different from zero.



- A couple of the slopes in the means plot appear to be different, which would imply the interactions are not zero. However there are parallel lines between certain interactions as the Vitamin variable changes. That combined with the nested F test would lead me to believe that there is no significant interaction.

7) There are 2 other categorical variables in this dataset, namely GENDER and SMOKE. Do these variables interact amongst themselves or with VITAMIN or ALCOHOL when it comes to modeling CHOLESTEROL? Obtain means plots to see if there is interaction. Conduct nested model F-tests to rule out randomness as the explanation for observed patterns. Report your findings.

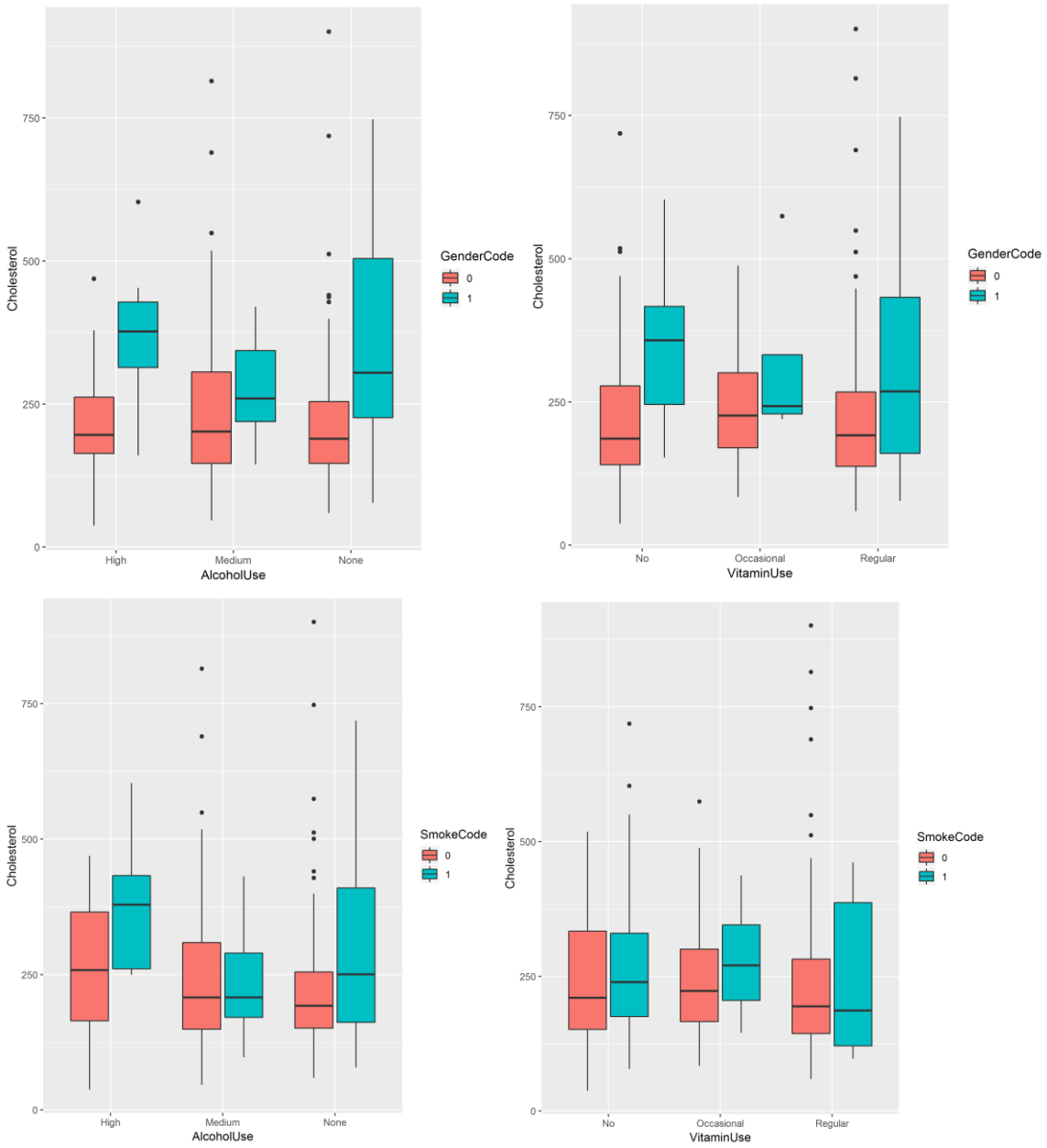


There does seem to be some interaction between gender and smoking. Both smokers and non-smokers have similar Cholesterol levels among females, however male smokers have much higher Cholesterol than non-smokers.

```
Analysis of Variance Table

Model 1: Cholesterol ~ GenderCode + SmokeCode + GenderCode * SmokeCode
Model 2: Cholesterol ~ GenderCode + SmokeCode
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1    311 5011965
2    312 5078043 -1    -66078 4.1002 0.04373 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

After running the nested model F-test to isolate the interaction of gender and smoking, we see that we reject the null hypothesis (H_0: $\beta_{interaction} = 0$, H_A: $\beta_{interaction}$ does not equal 0) at the 95% confidence level. It seems like the interaction between gender and smoking is a significant predictor of Cholesterol.



Based on these graphs, there does not seem to be any other significant interaction between gender, alcohol, vitamin use, and smoking among any of the remaining pairs of variables.

8) Please write a reflection on your experiences from this assignment.

In this assignment I've learned how to create models using dummy and effect coded variables. The most interesting thing to me is that the predictions, goodness of fit tests, and residuals after using both dummy and effect coded variables are the same. Using one over the other will not help boost the model's r-squared value. The only true difference to them is the intercepts and beta estimates. Any given observation will have the same final prediction regardless of what method is used. So what that has taught me is that the most efficient thing would be to learn to be comfortable with using both dummy and effect coding so that each can be used in the most logical place. In models where it is helpful to know the grand mean, we should use effect coding. But for data where there is a logical baseline that we want to interpret changes from, like when we used the people who don't take any vitamins and could see if there were any effects from starting to use vitamins, it is most logical to use dummy variables.