1. (3pts) You learn from your professor that, historically, 33% of seniors who take Stat2 earn an A in the course, while only 20% of sophomores who take Stat2 earn an A. Calculate (but do not interpret) the odds ratio of earning an A for seniors versus sophomores.

Odds for SRs = 0.33/0.67 = 0.5

Odds for Sophs = 0.20/0.80 = 0.25 🡪 OR = 0.5/0.25 = 2

1. (3pts) The empirical logit plot for one of the predictor variables in a multiple logistic regression model is as below. What action would you suggest?

Chart, scatter chart

Description automatically generated

1. No action: this looks fine
2. You can’t use this *xvar* in the model because this plot is not linear.
3. **Try a transformation of *xvar* to achieve linearity.**
4. Try a transformation of y to achieve linearity.
5. Trick Question! This is not even the right plot to be looking at.

**For the rest of the test, we will be using the US\_states data set that we worked with at the beginning of the semester. You have been given the codebook, along with 6 pages of R output.**

**Your response variable is *Elect2020*.**

1. (2pts each) First, we investigate several variables to see if they look like good predictors of *Elect2020*. Consider the EDA on the first page of the output. Using this EDA, indicate (circle) whether each of the variables below looks like it would be: **useful** in the model to predict *Elect2020*, **not useful**, or **can’t tell** from the plot/table given.
   1. *Region* useful not useful **can’t tell**
   2. *EnoughFruit* **useful** not useful can’t tell
   3. *HouseholdIncome* useful not useful **can’t tell**
   4. *EighthGradeMath* useful not useful **can’t tell**
   5. *Obese* useful not useful **can’t tell**
   6. *HeavyDrinkers* useful **not useful** can’t tell
   7. *College* **useful** not useful can’t tell
2. (3pts) Choose one of the variables that you called “useful” above. Then write a one-sentence summary of the relationship between that variable and *Elect2020*, based on the EDA.

States that vote D have a higher overall % of college grads (median = 39% vs. 30% for states that vote R).

States that vote D are much more likely to have high fruit consumption (17/25) than states that vote R (2/25).

**You decide to take a “backwards stepwise” approach to model-building, and first fit a model (*Model1*) with many variables.**

1. (3pts) In these models, R defines “success” to be *Elect2020*=R. In a sentence or two, tell me how you would know this, or how you could verify this if you had access to this data and R code.

I know this because R chooses alphabetically. So D = 0 and R = 1, thus R is “success”.

Alternatively, you could look at the Model1 output and see that “LowFruit” *increases* a state’s odds of success, while higher rates of “College” *decrease* a state’s odds of success. Matching this to the conclusions in #4, we must conclude that “success” is Elect2020=R.

1. (5pts) Which variables in *Model1* are statistically significant (at a 10% level of significance)? (circle all that apply)
   1. *EnoughFruit*
   2. *HouseholdIncome*
   3. ***College***
   4. ***EighthGradeMath***
   5. ***Obese***
2. (4pts) Interpret the coefficient on EnoughFruit in context.

exp(1.027) = 2.79

Holding all other variables constant, LowFruit states (those with less than 65% of residents eating the daily recommended amount) have 2.79x the odds of voting R compared to HighFruit states (those with more than 65% of residents eating the daily recommended amount).

1. (3pts) How would we assess the condition of linearity for the variable *EnoughFruit*? (circle one)
   1. If the p-value on *EnoughFruit* is significant, the relationship is linear.
   2. If the coefficient on *EnoughFruit* is positive, the relationship is linear.
   3. **The relationship is guaranteed to be linear because the explanatory variable is binary.**
   4. If the plot of the empirical log odds of *Elect2020* against *EnoughFruit* is linear, the linearity condition is met.
   5. If the plot of the empirical probability (proportion) of *Elect2020* against *EnoughFruit* is linear, the linearity condition is met.

***Model2* is a reduced model.**

1. (4pts) State a conclusion in context based on the **nested** drop-in-deviance test (nested likelihood ratio test) on page 3 of the R output. Be specific about the p-value that you are basing your conclusion on.

Since the p-value is large (0.5466), there is no evidence that *EnoughFruit* or *HouseholdIncome* add significantly to the model containing *College, EighthGradeMath*, and *Obese*.

1. (4pts) According to *Model2,* is *College* a useful predictor of *Elect2020*? Report the p-value and make a conclusion in context.

Assuming the model contains *EighthGradeMath* and *Obese*, *College* is a significant predictor of *Elect2020*, with a p-value of 0.00629.

1. (3pts) In *Model2*, how would we assess the condition of linearity for the variable *Obese*? (circle one)
   1. If the p-value on *Obese* is significant, the relationship is linear.
   2. If the coefficient on *Obese* is positive, the relationship is linear.
   3. The relationship is guaranteed to be linear because the explanatory variable is binary.
   4. **If the plot of the empirical log odds of *Elect2020* against *Obese* is linear, the linearity condition is met.**
   5. If the plot of the empirical probability (proportion) of *Elect2020* against *Obese* is linear, the linearity condition is met.

***Model3* deletes the least significant variable: EighthGradeMath**

1. (4pts) Why would I delete a significant variable?

This was the least significant of the variables in Model2, and I should try a simpler model to see if this variable is really worth it.

1. (5pts) Interpret a 95% confidence interval for the effect of *Obese* in *Model3* in context.

exp(0.1374) = 1.15 We are 95% confident that a 1% increase in the % of residents

exp(0.7118) = 2.04 classified as obese will increase the state’s odds of voting

Republican by between 15% and 104%, on average, when holding rates of college graduation constant.

1. (3pts) Consider a hypothetical state where 30% of residents age 25-34 have college diplomas and 24% of residents are classified as obese. Using *Model3*, calculate the probability of this state voting Republican in 2020.

fitted value for this state = -5.856 – 0.1977(30) + 0.3915(24) = -2.391

Prob = exp(fitted value) = 0.084

1+exp(fitted value)

1. (3pts) According to *Model3*, does graduating from college raise or lower a person’s chances of voting Republican? (circle one)
   1. Raises
   2. Lowers
   3. **This model does not tell me whether graduating from college raise or lower a person’s chances of voting Republican.**

**You know (or a friend tells you) that in the US, states tend to vote very similarly based on region. For example, the northeast tends to vote Democratic while the south is extremely Republican. (This is mainly for cultural and historical reasons, in case you’re wondering.) So you decide to fit a model with only one variable: *Region*. This is *Model4*.**

1. (4pts) Is Region a good predictor of *Elect2020*? Be specific about what leads you to conclude that it is or is not.

Yes, because the drop-in-deviance test shows the variable is highly significant (p-value = 1.615x10-5).

1. (3pts) One of the p-values is 0.122. Interpret this p-value – be specific about what, exactly, it’s telling you!

The odds of a western state voting Republican is not significantly different than the odds of a midwestern state (the reference group) voting Republican.

1. (3pts) The intercept of the model is 0.811. Interpret this quantity in context.

exp(0.811) = 2.25

The odds of a midwestern state voting Republican is 2.25 = 9/4. So for every 9 R states in the Midwest, we expect 4 D states.

1. (3pts) Consider the states in the west compared to the southern states. Using *Model4*, calculate (but do not interpret) the odds ratio of voting Republican between these two states.

Log(odds of R for W) = 0.8109 – 1.2809 = -0.47 🡪 odds = exp(-0.47) = 0.625

Log(odds of R for S) = 0.8109 + 0.8938 = 1.705 🡪 odds = exp(1.705) = 5.5

Odds ratio = 0.625/5.5 = 0.114 or Odds ratio = 5.5/0.625 = 8.8

1. (3pts) In *Model4*, how would we assess the condition of linearity for the variable *Region*? (circle one)
   1. If the p-value on *Region* is significant, the relationship is linear.
   2. If the coefficient on *Region* is positive, the relationship is linear.
   3. **The relationship is guaranteed to be linear because the explanatory variable is categorical.**
   4. If the plot of the empirical log odds of *Elect2020* against *Region* is linear, the linearity condition is met.
   5. If the plot of the empirical probability (proportion) of *Elect2020* against *Region* is linear, the linearity condition is met.
2. (5pts) Of the 4 models presented, which is your “favorite” model, and why? Use the misclassification tables on the final page along with the summaries from each model to make your choice. If referencing particular numbers (misclassification rates, p-values, deviances), make sure you are specific! You may assume that all conditions are met for all 4 models.

Summary:

Model1: misclassification = 16%; 5 variables (3 sig); resid deviance = 31.14

Model2: misclassification = 18%; 3 variables (3 sig); resid deviance = 32.35

Model3: misclassification = 18%; 2 variables (2 sig); resid deviance = 41.78

Model4: misclassification = 22%; 1 variable with 3 categories (none sig, although variable is sig); resid deviance = 44.53

Weighing misclassification, complexity, and deviance, Models 2 or 3 seem like the best option. I would probably prefer Model2, since the deviance is much lower.

1. (3pts) You learn from your professor that, historically, 40% of seniors who take Stat2 earn an A in the course, while only 10% of sophomores who take Stat2 earn an A. Calculate (but do not interpret) the odds ratio of earning an A for seniors versus sophomores.

Odds for SRs = 0.4/0.6 = 0.6667

Odds for Sophs = 0.1/0.9 = 0.1111 🡪 OR = 0.6667/0.1111 = 6

1. (3pts) The empirical probability plot for one of the predictor variables in a multiple logistic regression model is as below. What action would you suggest?

Chart, scatter chart

Description automatically generated

* 1. No action: this looks fine
  2. You can’t use this *xvar* in the model because this plot is not linear.
  3. Try a transformation of *xvar* to achieve linearity.
  4. Try a transformation of y to achieve linearity.
  5. **Trick Question! This is not even the right plot to be looking at.**

**For the rest of the test, we will be using the US\_states data set that we worked with at the beginning of the semester. You have been given the codebook, along with 6 pages of R output.**

**Your response variable is *majority\_vax*.**

1. (2pts each) First, we investigate several variables to see if they look like good predictors of *majority\_vax*. Consider the EDA on the first page of the output. Using this EDA, indicate (circle) whether each of the variables below looks like it would be: **useful** in the model to predict *majority\_vax*, **not useful**, or **can’t tell** from the plot/table given.
   1. *Region* **useful** not useful can’t tell
   2. *EnoughFruit* useful not useful **can’t tell**
   3. *HouseholdIncome* **useful** not useful can’t tell
   4. *EighthGradeMath* useful not useful **can’t tell**
   5. *Obese* **useful** not useful can’t tell
   6. *HeavyDrinkers* useful not useful **can’t tell**
   7. *College* useful not useful **can’t tell**
2. (3pts) Choose one of the variables that you called “useful” above. Then write a one-sentence summary of the relationship between that variable and *majority\_vax*, based on the EDA.

States that are majority vaxxed have a higher average household income (median = 65K vs. 52K for states not majority vaxxed).

States that are majority vaxxed have a lower overall obesity rate (median = 30% vs. 34% for states not majority vaxxed).

States in the northeast are most likely to be majority vaxxed (11/11), then states in the west (6/13), then states in the Midwest and south (3/13 and 1/13).

**You decide to take a “backwards stepwise” approach to model-building, and first fit a model (*Model1*) with many variables.**

1. (3pts) In these models, R defines “success” to be *majority\_vax*=Yes. In a sentence or two, tell me how you would know this, or how you could verify this if you had access to this data and R code.

I know this because R chooses alphabetically. So No = 0 and Yes = 1, thus Yes is “success”.

Alternatively, you could look at the Model1 output and see that higher Household Incomes *increases* a state’s odds of success, while higher rates of “Obese” *decrease* a state’s odds of success. Matching this to the conclusions in #4, we must conclude that “success” is majority\_vax=Yes.

1. (5pts) Which variables in *Model1* are statistically significant (at a 10% level of significance)? (circle all that apply)
   1. *EnoughFruit*
   2. *HouseholdIncome*
   3. ***College***
   4. ***EighthGradeMath***
   5. ***Obese***
2. (4pts) Interpret the coefficient on EnoughFruit in context.

exp(-1.616) = 0.1986

Holding all other variables constant, LowFruit states (those with less than 65% of residents eating the daily recommended amount) have 20% the odds of being majority vaxxed compared to HighFruit states (those with more than 65% of residents eating the daily recommended amount).

1. (3pts) How would we assess the condition of linearity for the variable *EnoughFruit*? (circle one)
   1. If the p-value on *EnoughFruit* is significant, the relationship is linear.
   2. If the coefficient on *EnoughFruit* is positive, the relationship is linear.
   3. **The relationship is guaranteed to be linear because the explanatory variable is binary.**
   4. If the plot of the empirical log odds of *majority\_vax* against *EnoughFruit* is linear, the linearity condition is met.
   5. If the plot of the empirical probability (proportion) of *majority\_vax* against *EnoughFruit* is linear, the linearity condition is met.

***Model2* is a reduced model.**

1. (4pts) State a conclusion in context based on the **nested** drop-in-deviance test (nested likelihood ratio test) on page 3 of the R output. Be specific about the p-value that you are basing your conclusion on.

Since the p-value is large (0.4014), there is no evidence that *EnoughFruit* or *HouseholdIncome* add significantly to the model containing *College, EighthGradeMath*, and *Obese*.

1. (4pts) According to *Model2,* is *College* a useful predictor of *majority\_vax*? Report the p-value and make a conclusion in context.

Assuming the model contains *EighthGradeMath* and *Obese*, *College* is a significant predictor of *majority\_vax*, with a p-value of 0.00403.

1. (3pts) In *Model2*, how would we assess the condition of linearity for the variable *Obese*? (circle one)
   1. If the p-value on *Obese* is significant, the relationship is linear.
   2. If the coefficient on *Obese* is positive, the relationship is linear.
   3. The relationship is guaranteed to be linear because the explanatory variable is binary.
   4. **If the plot of the empirical log odds of *majority\_vax* against *Obese* is linear, the linearity condition is met.**
   5. If the plot of the empirical probability (proportion) of *majority\_vax* against *Obese* is linear, the linearity condition is met.

***Model3* deletes the least significant variable: EighthGradeMath**

1. (4pts) Why would I delete a significant variable?

This was the least significant of the variables in Model2, and I should try a simpler model to see if this variable is really worth it.

1. (5pts) Interpret a 95% confidence interval for the effect of *College* in *Model3* in context.

exp(0.1254) = 1.134 We are 95% confident that a 1% increase in the % of residents

exp(0.5461) = 1.726 with college degrees will increase the state’s odds of being

majority vaxxed by between 13% and 73%, on average, when holding obesity rates constant.

1. (3pts) Consider a hypothetical state where 30% of residents age 25-34 have college diplomas and 24% of residents are classified as obese. Using *Model3*, calculate the probability of this state having at least 50% of their residents vaccinated.

fitted value for this state = -0.6411 + 0.3002(30) - 0.3121(24) = 0.8745

Prob = exp(fitted value) = 0.7056

1+exp(fitted value)

1. (3pts) According to *Model3*, does graduating from college raise or lower a person’s chances of being vaccinated? (circle one)
   1. Raises
   2. Lowers
   3. **This model does not tell me whether graduating from college raise or lower a person’s chances of being vaccinated.**

**In the US, Covid vaccination is highly politicized, so Republicans are less likely to be vaccinated against Covid-19. Region is a strong predictor of voting behavior. For example, the northeast states tends to vote Democratic while the south is extremely Republican. So you decide to fit a model with only one variable: *Region*. This is *Model4*.**

1. (4pts) Is Region a good predictor of *majority\_vax*? Be specific about what leads you to conclude that it is or is not.

Yes, because the drop-in-deviance test shows the variable is highly significant (p-value = 2.252x10-6).

1. (3pts) One of the p-values is 0.298. Interpret this p-value – be specific about what, exactly, it’s telling you!

The odds of a southern state being majority vaxxed is not significantly different than the odds of a midwestern state (the reference group) being majority vaxxed.

1. (3pts) The intercept of the model is -1.204. Interpret this quantity in context.

exp(-1.204) = 0.3

The odds of a midwestern state being majority vaxxed is 0.3 = 3/10. So for every 3 states in the Midwest with a majority of their residents vaccinated, we expect 10 states that are not majority vaxxed.

1. (3pts) Consider the states in the west compared to the southern states. Using *Model4*, calculate (but do not interpret) the odds ratio of *majority\_vax* between these two states.

Log(odds for W) = -1.2040 + 1.0498 = -0.1542 🡪 odds = exp(-0.154) = 0.857

Log(odds for S) = -1.2040 - 1.2809 = -2.4849 🡪 odds = exp(-2.485) = 0.083

Odds ratio = 0.857/0.083 = 10.33 or Odds ratio = 0.083/0.857 = 0.097

1. (3pts) In *Model4*, how would we assess the condition of linearity for the variable *Region*? (circle one)
   1. If the p-value on *Region* is significant, the relationship is linear.
   2. If the coefficient on *Region* is positive, the relationship is linear.
   3. **The relationship is guaranteed to be linear because the explanatory variable is categorical.**
   4. If the plot of the empirical log odds of *majority\_vax* against *Region* is linear, the linearity condition is met.
   5. If the plot of the empirical probability (proportion) of *majority\_vax* against *Region* is linear, the linearity condition is met.
2. (5pts) Of the 4 models presented, which is your “favorite” model, and why? Use the misclassification tables on the final page along with the summaries from each model to make your choice. If referencing particular numbers (misclassification rates, p-values, deviances), make sure you are specific! You may assume that all conditions are met for all 4 models.

Summary:

Model1: misclassification = 10%; 5 variables (3 sig); resid deviance = 30.26

Model2: misclassification = 10%; 3 variables (3 sig); resid deviance = 32.09

Model3: misclassification = 16%; 2 variables (2 sig); resid deviance = 37.03

Model4: misclassification = 20%; 1 variable with 3 categories (none sig, although variable is sig); resid deviance = 39.04

Weighing misclassification, complexity, and deviance, Model2 seems like the best option. It’s more complex than Model3 but has a much lower deviance and misclassification rate.