BIOSTAT 201B Homework 3

Lillian Chen (c.lillian@ucla.edu)

February 9, 2022

Warm-up Problems

1. Probit Regression Basics:

(a) Explain what the distributions, link functions and systematic components are for the probit model.

$Y \sim Bernoulli$, link function is the probit $g(p) = \Phi^{-1}(p) = X\beta$, systematic component is the linear combination of the predictors $\beta_0 + \beta_1 X_1 + ... + \beta_m X_m$. The probit function takes a Z-score and returns the probability of a normal distribution being less than that value.

(b) Give examples of circumstances where you might want to use this model.

In dose-response or toxicology studies where lethal dose is often tracked for, probit models are more common.

2. Multinomial and Ordinal Logistic Regression Basics:

(a) Explain what the three basic GLM components are for a multinomial logistic regression.

$Y \sim multinomial$ s.t. Y has k levels whose probabilities $p_1 + ...p_k = 1$ for the $k+1$ outcome categories, link function is the logit $g(p) = \log \frac{p}{1-p} = X\beta$, systematic component is the linear combination of the predictors $\beta_0 + \beta_1 X_1 + ... + \beta_m X_m$ represented by a separate equation for each level (thus, k sets of coefficients comparing a given outcome level to the reference level.

(b) Explain how the ideas of odds and odds ratios are extended for multinomial logistic regression.

In logistic regression, we simply defined the odds as $p/(1-p)$ with p being the probability given a set of predictor values. In multinomial logistic regression, we choose one outcome level as the reference group, and we get the odds relative to that reference group. For a given category j with probability p_j, the odds would be defined as p_j/p_0 with p_0 being the probability for the reference level. The odds ratio corresponding to a particular variable is still obtained by exponentiating, but for interpretation we have to specify 1) which variable we are computing the odds ratio for and 2) that the odds are relative to a reference category that also needs to be specified.

(c) Explain how multinomial logistic regression can be adapted if there is a natural ordering to the outcome categories and what the proportional odds assumption means.

If there is a natural ordering to the outcome categories, we would be able to adapt this to an ordinal logistic regression model under the proportional odds assumption and get the model

$$log \frac{P(Y \leq j}{Y > j} = \beta_{j,0} + \beta_{j,1} X_1 + ... + \beta_{j,m} X_m \, j = 0, 1, ..., k-1$$

We get the odds of lower values of Y relative to higher values of Y, with the assumption that the split can be made at any point. What this additionally means is that our odds ratios are directional, and the effect of the regression coefficients on X variables is the same regardless of where the cut point is for higher and lower values.

The proportional odds assumption means that if we look on the log odds scale the relationship between the predictors and the different outcome levels is the same–i.e. we get parallel lines. This is quite a strong

assumption. We can relax the assumption by allowing different coefficients at each cut point (generalized ordinal logistic) or with other groupings of the categories.

3. Gasping For Breath Some More:

This problem continues the Asthma warm-up problem from assignments 1 and 2. Recall that Professor Urtha Green is studying risk factors for childhood asthma. She has participated in a study that followed 1000 children from birth to age 10, recording whether or not they developed asthma during that period (Y = 1 for yes and Y = 0 for no). The study also collected information on potential risk factors and protective effects from the child's first year of life including X1, an indicator for whether the child's family lived in an urban setting (Yes = 1, No = 0), X2, the average annual pollution level in thousands of particles per cm3 for the county in which the child lived, X3, an index of socio-economic status for the child's family (higher is better), X4, the number of months for which the child was breast-fed, X5, an indicator for whether there was a family history of asthma (1 = Yes, 0 = No), and X6, sex (1 = Female, 0 = Male). The data are given in the accompanying file.

(a) On HWs 1 and 2 you fit a standard logistic model using all the predictors. Suppose Professor Green informs you that the study actually oversampled children living in urban environments and/or having a family history to make sure that there were enough asthma cases to get good estimates of the effects of the various predictors. She says the true rate of childhood asthma cases in the population is closer to 5%. Explain what you would have to do to adjust the model from Assignment 3 to get good predicted probabilities under this assumption.

We have oversampled the number of cases/rare-outcomes in the population since the true rate of asthma cases is lower than expected. In this situation, odds ratios and odds interpretations are still valid, but the intercept is not and the current intercept will not give us accurate predicted probabilities. To do this, we need to recalibrate the intercept using the following (controls in numerator n_0, cases in denominator n_1:

$$\hat{\beta_0} = \beta_0 + log\frac{Pn_0}{(1-P)n_1} = -0.47 - 0.980 = -1.027$$

We get a lower β_0 which means our predicted probability p will be lower, which reflects the adjustment that the proportion of cases in our sample was higher than the true proportion in the population.

(b) Suppose Professor Green had told you that the data in her sample came from 500 sibling pairs. Explain how this would affect your analyses. (You do not need to actually run a model, just explain how you would do so, what variables you would include, and why it would be important to take account of the pairing.)

We would want to rerun the model to take advantage of the matched pairs for higher power in our model. In this case, we would rerun the model as a conditional logistic regression. If we were doing this manually we would select all these pairs, take the differences in the predictor variables between the cases and non-cases and fit a logistic regression with the outcome set to 1 for all pairs and the intercept forced to 0.

For parts (c)-(d) go back to assuming that the data are a representative sample of children from this region.
(c) Fit a probit regression model to these data and provide an interpretation of the coefficients for the family history and pollution variables on the probit scale. Is there an easy interpretation on the probability scale?

(d) Find the predicted probability of developing asthma for a boy born in the city of Los Seraphim where the pollution level is 35 thousand cm3 to parents who had asthma and an SES index score of 50 and who was breastfed for 6 months based on your probit model. Compare your answer to what you found in Problem 5k on HW1.

(e) Compare the log likelihoods for your probit model and the standard logistic model from HWs 1 and 2. Does one appear to be better than the other? Do you think the difference is substantial?

4. The Accidental Statistician:

The problem continues with the data set from the 2011 201a final on drinking and traffic accidents which we saw as Warm-up Problem 1 on HW2. Recall that there were three types of accidents: single car (solo) accidents, two-car accidents and a three-or-more (multi) car accidents. We now want to understand what factors differentiate among the different accident types. I have recoded the data so that our outcome variable is "type" which has three levels: 0 = a two-car accident, 1 = a solo accident and 2 = a multi-car accident. As predictors we consider the following subset of the original variables: whether or not the accident involved a fatality (fatal = 1 for yes and 0 for no), whether the accident took place at night (dark = 1 for yes and 0 for no), the sex of the driver (1=Female and 0=Male) and the speed at which the cars were travelling (in miles per hour.) The data are given in the accompanying file.

(a) Fit a multinomial logistic regression for these data using the two-car accident as the reference category. Is this model overall significant? Explain briefly.

(b) Give a careful interpretation of the coefficients for the coefficients of "fatal" and "speed" for each of the two model components (solo versus two-car and multi versus two-car).

(c) Which factors appear to be important for differentiating among the three accident types? You should explain both in terms of the p-values in the individual model components and by testing for the overall significance of the variables. Make sure you are able to write down the null and alternative hypotheses you are testing, both mathematically and in words..

(d) For the variables that are significant, does it appear that their coefficients differ across the two components of the model? Check by performing an appropriate set of tests. Again, be sure you know how to write out the formal hypotheses.

(e Based on your answers to (a)-(d) describe as carefully as you can what sort of conditions are more likely to lead to what sort of accident.

(f) Find the predicted probability of each type of accident given you know that the at fault car was driven by a man at 90 miles per hour after dark and that there were no fatalities involved.

For the remainder of the problem, suppose that you consider the outcome categories to be ordered in terms of seriousness (two-car crash is the least serious, then a solo crash, then a multi-car crash).

(g) Fit an appropriate model for this situation using the same set of predictors as in parts (a)-(e). Is the model overall significant?

(h) Give careful interpretations of the coefficients of fatal, dark and speed and their corresponding odds ratios. Which of these variables appear to be significant predictors?

(i) Find the predicted probability of each type of accident based on this model given you know that the at fault car was driven by a man at 90 miles per hour after dark and that there were no fatalities involved.

(j) Compare and contrast your results with those from the multinomial logistic model.

Problems to Turn In

5. Cancer Conundrums:

At a chemical plant workers are exposed to compounds which may cause skin cancer. In order to evaluate the effect of exposure on risk of cancer, 50 plant workers who developed skin cancer are identified and are matched on age and sex with workers who were employed during the same time period but did not develop skin cancer. The variables for this problem are cancer (1 = yes, 0 = no), exposure (a continuous score with higher being more exposure) and pairid which identifies how the employees were matched up. Our goal is to evaluate the impact of exposure.

(a) Fit a conditional logistic regression to these data. Does there appear to be a significant effect of exposure? Give a careful interpretation of the odds ratio for the exposure variable.

Based on the output for the conditional logistic regression, there does appear to be a significant effect of exposure ($p < .001$). The odds ratio for exposure is 1.394, with 95% confidence limits of [1.154, 1.683]. We can say that the odds of developing cancer are 1.394 higher for each additional 1-unit increase in exposure score, adjusting for the matched variables.

Figure 1: Conditional logistic regression of cancer on exposure in matched pairs of chemical plant workers

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	25.3819	1	<.0001
Score	19.4059	1	<.0001
Wald	11.9165	1	0.0006

Analysis of Conditional Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
exposure	1	0.3319	0.0962	11.9165	0.0006

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
exposure	1.394	1.154	1.683

(b) Now rerun the model ignoring the pairing–i.e supposing that we were simply performing a case control study–and obtain the odds ratio for exposure. Does the effect of exposure seem stronger or weaker than in the conditional logistic model? Is the variable more or less significant? Does this fit with your expectations of how these models work? Why or why not?

The odds ratio for exposure in this model, ignoring the pairing, is 1.105, with 95% confidence limits of [1.034, 1.181]. The effect of exposure does seem weaker in this model as compared to that in the conditional logistic model, as the odds ratio estimate is closer to 1. We also see that the exposure variable is less significant ($p = 0.003$) in this model as compared to that in the conditional logistic model ($p < .001$). This does fit with my expectations of how these two models work, since matching does power the model more since we save degrees of freedom by matching on other nuisance variables and not needing to include the matched variables into our model. By excluding the matched variables in the rerun model which didn't account for the matching, we also lose information related to age and sex for the cases and the controls.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	10.0094	1	0.0016
Score	9.5012	1	0.0021
Wald	8.5817	1	0.0034

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4276	0.5257	7.3736	0.0066
exposure	1	0.0997	0.0340	8.5817	0.0034

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
exposure	1.105	1.034	1.181

Figure 2: Logistic regression of cancer on exposure in chemical plant workers

(c) Suppose we really had been doing a case-control rather than a matched pair study but were told that the actual rate of skin cancer in the general population was only 5%. How would we adjust the model from part (b) so that it would give appropriate estimates of the predicted probabilities? How much would this change the predicted probability for a person with an exposure score of 20?

The scenario here means that we oversampled the number of cases and that our estimates of the predicted probabilities will be too high, so we need to recalibrate the intercept term to get more appropriate estimates of the predicted probabilities by adjusting for the degree of imbalance in the sample.

$$\hat{\beta_0}^* = \hat{\beta}_0 + \ln\left(\frac{\hat{P}n_0}{(1-\hat{P})n_1}\right) = -1.4276 + \ln\frac{0.05*50}{0.95*50} = -4.372$$

We get a new intercept of -4.372, so the model would now simply have the new intercept in place of the old one, and the model would be

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta_0}^* + \hat{\beta}_1 X_1 = -4.372 + 0.0997 X_1$$

For a person with an exposure score of 20, the predicted probability using the unadjusted intercept is

$$p = \frac{\exp(-1.4276 + 0.0997 * 20)}{1 + \exp(-1.4276 + 0.0997 * 20)} = 0.638$$

Using the adjusted intercept, the recalibrated predicted probability is

$$p^* = \frac{\exp(-4.372 + 0.0997 * 20)}{1 + \exp(-4.372 + 0.0997 * 20)} = 0.0526$$

For a person with an exposure score of 20, the use of the recalibrated intercept in the model would decrease that person's predicted probability by 0.553.

6. I Await Your Response:

Researchers are pioneering a new medication and are interested in knowing how high the dose needs to be before patients will respond. There is also some concern that too high a dose may actually start to be toxic and decrease the chance of a response. They have tried the medication at different doses on a sample of 500 patients and recorded whether or not the patients had a positive response to treatment. The data are given in the accompanying file.

(a) Fit a probit regression model with response as the outcome and dose as the predictor and say whether there is a significant dose-response relationship. Explain why a probit model may be a natural choice for these data.

When we fit the probit regression model, we see that the regression coefficient for dose on the probit scale is 0.0227 ($p < .001$), and based on the p-value being much lower than the significance level of $\alpha = 0.05$ we conclude that there is a significant dose-response relationship. A probit model may be a natural choice for these data because the resulting model yields a more extreme sigmoidal curve which is often observed in biological or assay data, and there is no need to employ the logit model for easy interpretations of independent predictors since there is essentially only one predictor in this model.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8094	0.1166	48.1786	<.0001
dose	1	0.0227	0.00262	75.3777	<.0001

Figure 3: Probit regression of response on dose for treating patients on a new medication

(b) Give an interpretation of the intercept in this model. (Note–you may find this easier to do if you transform back to the probability scale. Recall that you can use the display normal(z) command to get the probability that a standard normal random variable is less than a particular value, z.)

The intercept on the probit scale is -0.8094, which corresponds to 0.209 on the probability scale. We can say that the probability that a patient will respond when they are administered a dose of 0 is 0.209, or 20.9%.

(c) Find the dose at which the probability of a response is 50%. Explain your reasoning.

The dose at which the probability of a response is 50% is at 35.67 dose units. We want to find dosage such that the systematic component equals to the inverse normal CDF associated with $p = 0.50$, which happens to be when $Z = 0$. Using this information, we obtain $X_1 = 35.67$ as the dose that will have a 50% probability of a response.

$$\begin{aligned} \Phi^{-1}(0.50) = \beta_0 + \beta_1 X_1 &= -0.8094 + 0.0227 X_1 \\ 0 &= -0.8094 + 0.0227 X_1 \\ X_1 &= 35.67 \end{aligned}$$

(d) Perform a Hosmer-Lemeshow goodness of fit test for this model. Does the model appear to be well-calibrated? Explain briefly.

The test statistic for the Hosmer-Lemeshow goodness of fit test is $\chi^2 = 67.28$ with $p < .001$, indicating that this model does not appear to be overall well-calibrated, The high test-statistic value signals that there is a large difference in the observed and expected response rates (the numerator of the test statistic) relative to the expected proportions for each group. We reject the null hypothesis that the observed and expected response proportions are the same across all doses and conclude that the model is not overall well calibrated.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
67.2750	8	<.0001

Figure 4: Hosmer-Lemeshow goodness of fit test for probit regression of response on dose

(e) Now rerun the model including both dose and dose squared as predictors. (The variable dosesq has been included for your convenience.) Is the curvilinear model in dose an improvement? Explain briefly. Explain what the signs of the dose and dose-squared terms tell you about the dose-response relationship. Do they confirm the researchers' theories? Is the calibration of this model adequate?

The curvilinear model is a significant improvement, as we see that the test statistic for the goodness of fit test is now $\chi^2 = 4.11$ with $p = .847$, indicating that this model is overall well-calibrated. The positive linear term for dose and the negative dose-squared term indicates that the dose-response relationship appears to be a parabola, with the probability of a response increasing with dose up until a certain point, after which the probability of a response decreases. This confirms the researchers' theories that too high of a dose may decrease the chance of a response. The calibration of this model appears to be adequate as the goodness of fit test shows significant improvements in the lowered test statistic and associated p value.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3142	0.2501	85.5990	<.0001
dose	1	0.1210	0.0131	85.8041	<.0001
dosesq	1	-0.00116	0.000145	63.5756	<.0001

Figure 5: Probit regression of response on dose and dosesq for treating patients on a new medication

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
4.1141	8	0.8467

Figure 6: Hosmer-Lemeshow goodness of fit test for probit regression of response on dose and dosesq

(f) Based on the fitted model from part (e) identify the optimal dose of this medication and the response rate at that optimal dose. (Note: You may find it useful to recall that the peak of the parabola $ax^2 + bx + c$ occurs at $x = -b/2a$.)

Since we know that the model is a parabola opening downwards, the peak of the parabola is where the medication is at the optimal dose and yields the maximum response rate. The optimal dose is

$$X_{optimal} = \frac{-(0.1210)}{2(-0.00116)} = 52.16 \text{ dose units}$$

and the response rate at that optimal dose is

$$Y_{max} = \Phi(-2.3142 + 0.1210 * 52.16 - 0.00116 * (52.16)^2) = \Phi(0.841) = 0.800, \text{ or } 80.0\%.$$

(g) Obtain the fitted probabilities for your model from part (e). Then rerun the model using a logit link and obtain the corresponding fitted probabilities. Do the two models seem very different? Discuss this by

(i) Comparing the predicted probabilities of response when the person receives no medication. (ii) Comparing the optimal doses. (iii) Comparing the log likelihoods for the two models (make sure you understand why this is a reasonable comparison!) Say which model (if either) appears better by this criterion. (iv) Calculating the correlation between the predicted probabilities for the two models. (v) Obtaining a scatterplot showing the relationship between the two sets of predicted probabilities.

(i) The predicted probability of response when the person receives no medication is $\Phi(-2.3142) = 0.0103$ for the probit model and $\exp(-3.8366) = 0.0216$ in the logit model. The predicted probability of response for a person receiving no medication in the logit model is higher than that from the probit model by around 1%, but the magnitudes of the predicted probability are similar.
(ii) The optimal dose for the probit model is 52.16 dose units as calculated above, and the optimal dose for the logit model is

$$X_{optimal} = \frac{-(0.2005)}{2(-0.00192)} = 52.21 \text{ dose units.} \tag{1}$$

The two optimal doses are remarkably close to each other.
(iii) The log likelihood for the probit model is $\frac{536.2}{-2} = -268.1$, as compared to $\frac{537.7}{-2} = -268.8$ for the logit model. Though we cannot actually determine what constitutes a big difference in the log likelihoods since the two models have different fits, the observed difference appears to be small (0.7 difference) and the probit model appears to be just a tiny bit better (more positive/less negative) than the logit model, but the difference is small and both models are comparable. This is a reasonable comparison because both models maximize the same likelihood function, just with different link functions to produce different predicted probabilities.
(iv) The calculated correlation between the probit and logit predicted probabilities is $\rho = 0.9999$ with $p < .001$, indicating an almost perfect correlation between the two. Very little would be lost in using the logit predicted values instead of the probit predicted values.
(v) The scatter plot of the two sets of probabilities shows that they are correlated in a nearly perfect straight line, confirming our conclusions from the correlation for the two models.
Overall, the two models do not seem very different.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.8366	0.4539	71.4569	<.0001
dose	1	0.2005	0.0232	74.4497	<.0001
dosesq	1	-0.00192	0.000253	57.5559	<.0001

Figure 7: Logistic regression of response on dose and dosesq for treating patients on a new medication

Pearson Correlation Coefficients, N = 500
Prob > |r| under H0: Rho=0

	probit_p	logit_p
probit_p	1.00000	0.99989
Estimated Probability		<.0001
logit_p	0.99989	1.00000
Estimated Probability	<.0001	

Figure 8: Correlation between logit and probit models

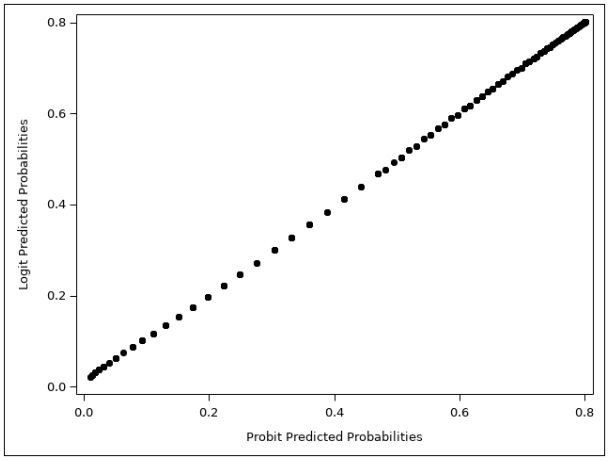

Figure 9: Scatter plot of logit vs. probit predicted probabilities

7. Healthy and Happy? (ACM 12.22):

This problem uses the depression data set which was featured in HW2, Problems 4 and 5. However, we are now going to reverse our set-up and evaluate health status as a function of age, income and depression.

(a) Fit a multinomial logistic regression for these data using the "excellent health" category as the reference. Is this model overall significant? Explain briefly.

The model does appear to be overall significant, as the likelihood ratio chi-squared test for the overall model has a test statistic of $\chi^2 = 48.24$ ($p < .001$), indicating that we should reject the null hypothesis that all the slope coefficients are 0 (that none of age, income, or depression have any effect on any levels of health), and we conclude that at least one of age, income, and/or depression is important for at least one category of health.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	48.2433	9	<.0001
Score	46.1054	9	<.0001
Wald	40.0857	9	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	3	24.9095	<.0001
income	3	6.7891	0.0789
depressed	3	9.9403	0.0191

Analysis of Maximum Likelihood Estimates

Parameter	health	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-0.9916	0.4480	4.8990	0.0269
Intercept	2	1	-2.5675	0.7216	12.6602	0.0004
Intercept	3	1	-6.3676	1.3702	21.5960	<.0001
age	1	1	0.0282	0.00802	12.3388	0.0004
age	2	1	0.0381	0.0115	10.9216	0.0010
age	3	1	0.0753	0.0190	15.6634	<.0001
income	1	1	-0.0177	0.00906	3.8209	0.0506
income	2	1	-0.0340	0.0167	4.1141	0.0425
income	3	1	0.00267	0.0213	0.0157	0.9004
depressed	1	1	0.2772	0.3821	0.5262	0.4682
depressed	2	1	1.0834	0.4869	4.9505	0.0261
depressed	3	1	1.7937	0.6756	7.0492	0.0079

Figure 10: Multinomial logistic regression of health on age, income, and depression, referent: **health** = Excellent

(b) Give a careful interpretation of the coefficients for the age, income and depression variables for each of the three model components (good, fair or poor vs excellent health). What would be the effect of an increase in age of 10 years on the odds of poor health? What would be the effect of an increase in income of $5000 on the odds of fair health? Show your work.

Good (1) vs. Excellent (0) Health

- Age: For every additional year in age, the log odds of having good health relative to excellent health increase by 0.0282 units, all else constant. For every additional year in age, the odds of having good health relative to excellent health increase by 2.9%, all else constant.
- Income: For every additional $1000 in income, the log odds of having good health relative to excellent health decrease by 0.0177 units, all else constant. For every additional $1000 in income, the odds of having good health relative to excellent health decrease by 1.8%, all else constant.
- Depression: The log odds of having good health relative to excellent health are 0.2772 units higher when depression is present than when no depression is present, all else constant. The odds of having good health

relative to excellent health are 1.3% higher when depression is present than when no depression is present, all else constant.

Fair (2) vs. Excellent (0) Health

- Age: For every additional year in age, the log odds of having fair health relative to excellent health increase by 0.0381 units, all else constant. For every additional year in age, the odds of having fair health relative to excellent health increase by 3.9%, all else constant.
- Income: For every additional $1000 in income, the log odds of having fair health relative to excellent health decrease by 0.0340 units, all else constant. For every additional $1000 in income, the odds of having fair health relative to excellent health decrease by 3.3%, all else constant.
- Depression: The log odds of having fair health relative to excellent health are 1.0834 units higher when depression is present than when no depression is present, all else constant. The odds of having fair health relative to excellent health are 195% higher when depression is present than when no depression is present, all else constant.

Poor (3) vs. Excellent (0) Health

- Age: For every additional year in age, the log odds of having poor health relative to excellent health increase by 0.0753 units, all else constant. For every additional year in age, the odds of having poor health relative to excellent health increase by 7.8%, all else constant.
- Income: For every additional $1000 in income, the log odds of having poor health relative to excellent health increase by 0.00267 units, all else constant. For every additional $1000 in income, the odds of having poor health relative to excellent health increase by 0.3%, all else constant.
- Depression: The log odds of having poor health relative to excellent health are 1.7937 units higher when depression is present than when no depression is present, all else constant. The odds of having poor health relative to excellent health are 501% higher when depression is present than when no depression is present, all else constant.

The effect of an increase in age of 10 years on the odds of poor health is as follows:

$$OR_3 = (e^{0}.0753)^{1}0 = 2.12$$

The effect of an increase in income of $5000 on the odds of fair health is as follows:

$$OR_2 = (e^{-}0.0340)^{5} = 0.844$$

An increase in age of 10 years is associated with a 112% increase in the odds of poor health relative to excellent health, all else constant. An increase in income of $5000 is associated with a 15.6% decrease in the odds of fair health relative to excellent health, all else constant.

(c) Which factors appear to be important for differentiating among the levels of health status? You should explain both in terms of the p-values in the individual model components and by testing for the overall significance of the variables. For one of the latter tests write the hypotheses mathematically and in words and give the test statistic. For the other tests you may simply give the p-values and your conclusions.

For testing individual variables at each level of the model, we conduct a Wald test and the test statistic is a Wald Z-statistic. For differentiating between good health from excellent health, age ($p < .001$) seems to be the only significant predictor. For differentiating between fair health from excellent health, the significant predictors include age ($p = .001$), income ($p = .043$), and whether or not one is depressed ($p = .026$). For differentiating between poor health from excellent health, the significant predictors include age ($p < .001$) and whether or not one is depressed ($p = .008$).

For testing overall significance of the variables, we can conduct a likelihood chi-squared test by fitting the model with and without the variable, or we can conduct a Wald type test (included in the Type 3 Analysis of Effects from the multinomial model printout from SAS). An example set of hypotheses for the age variable would be as follows:

- H_0 : $\beta_{1,1} = \beta_{2,1} = \beta_{3,1} = 0$; there is no overall relationship between age and health category after adjusting for income and depression status.
- H_A : at least one of $\beta_{1,1}$, $\beta_{2,1}$, $\beta_{3,1} \neq 0$; the relative likelihoods of the different health categories depend on age, even after adjusting for income and depression status.
- The Wald test statistic for age is $\chi^2 = 24.91$ ($p < .001$); we reject the null and conclude that age tells us something additional about health status even after adjusting for income and depression status.

Conducting a Wald style test again for the remaining two predictors, we get $\chi^2 = 6.79$ ($p = .079$) when testing for income, and $\chi^2 = 9.94$ ($p = .019$) when testing for depression status. We see that income does not appear to be significant and that depression appears to be significant with this Wald test. We can say that depression status provides unique information to distinguish between health categories, controlling for other predictors. Income does not appear to have an overall relationship with health after controlling for other predictors.

(d) For the variables that are significant, does it appear that their coefficients differ across the three components of the model? Check by performing an appropriate set of tests. For one of these tests write out the details of the hypotheses and test statistic. For the others you may simply give your p-values and conclusions.

From the previous part, we conclude that age and depression are overall significant in the multinomial model. An example set of hypotheses for the age variable would be as follows:

- H_0 : $\beta_{1,1} = \beta_{2,1} = \beta_{3,1}$; age has the same effect on the odds of good, fair and poor health after adjusting for income and depression status
- H_A : not all of the $\beta_{1,1}$, $\beta_{2,1}$, $\beta_{3,1}$ are equal; age has a different effect on at least one of the good, fair, and poor levels of the model, after adjusting for income and depression status
- The test statistic for the effect of age is $\chi^2_2 = 6.67$ ($p = 0.036$). We conclude that the effect of age differs across the model levels after adjusting for income and depression status.

Repeating the test for depression status, we get $\chi^2_2 = 6.73$ ($p = 0.035$). We conclude that the effect of depression status differs across the model levels after adjusting for age and income.

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
agecoeffeq	6.6732	2	0.0356
depcoeffeq	6.7332	2	0.0345

Figure 11: Test of equal effects of overall significant variables (age, depressed) across all levels of the model, referent: **health** = Excellent

(e Based on your answers to (a)-(d) describe as carefully as you can the relationships among the variables in this model.

Age is significant at the good health level; income is close to significant at this level as well. All three predictors are significant at the fair health level. Age and depression status are significant at the poor health level. Overall, age and depression status are significant, and we also see that there is evidence of different effects across levels of the model for both age and depression. It seems that older age is associated with increased odds of good, fair, and poor health, with the effect being stronger for fair health and strongest for poor health. It also seems that there are increased odds of good, fair, and poor health in the presence of depression vs no depression, with the effect being stronger for fair health and strongest for poor health.

(f) Find the predicted probability of each level of health status for a 50 year old who makes $50,000 per year and is depressed.

$$
\begin{aligned}
X\beta_1 &= -0.9916 + 0.0282 * 50 - 0.0177 * 50 + 0.2772 * 1 = -0.1894 \\
X\beta_2 &= -2.5675 + 0.0381 * 50 - 0.0340 * 50 + 1.0834 * 1 = -1.2791 \\
X\beta_3 &= -6.3676 + 0.0753 * 50 + 0.00267 * 50 + 1.7937 * 1 = -0.6754 \\
P(Y=0) &= \frac{1}{1 + e^{X\beta_1} + e^{X\beta_2} + e^{X\beta_3}} = \frac{1}{1 + e^{-0.1894} + e^{-1.2791} + e^{-0.6754}} = 0.382 \\
P(Y=1) &= \frac{e^{X\beta_1}}{1 + e^{X\beta_1} + e^{X\beta_2} + e^{X\beta_3}} = \frac{e^{-0.1894}}{1 + e^{-0.1894} + e^{-1.2791} + e^{-0.6754}} = 0.316 \\
P(Y=2) &= \frac{e^{X\beta_2}}{1 + e^{X\beta_1} + e^{X\beta_2} + e^{X\beta_3}} = \frac{e^{-1.2791}}{1 + e^{-0.1894} + e^{-1.2791} + e^{-0.6754}} = 0.106 \\
P(Y=3) &= \frac{e^{X\beta_3}}{1 + e^{X\beta_1} + e^{X\beta_2} + e^{X\beta_3}} = \frac{e^{-0.6754}}{1 + e^{-0.1894} + e^{-1.2791} + e^{-0.6754}} = 0.195
\end{aligned}
$$

For a 50 year old who makes \$50,000 per year and is depressed, the chance of having excellent health is 38.2%, the chance of having good health is 31.6%, the chance of having fair health is 10.6%, and the chance of having poor health is 19.5%.

(g) The values for the health status outcome have a natural ordering which our multinomial logit model ignored. Fit an appropriate model taking this into account using the same set of predictors as in parts (a)-(f). Is the model overall significant?

The model appears to be overall significant, as the likelihood ratio test for the overall model yields a test statistic of $X^2 = 40.81$ ($p < .001$), indicating that our model fits significantly better than the null model.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	40.8082	3	<.0001
Score	39.6573	3	<.0001
Wald	37.2491	3	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0	1	0.8925	0.3878	5.2952	0.0214
Intercept	1	1	2.9398	0.4277	47.2382	<.0001
Intercept	2	1	4.3920	0.4925	79.5154	<.0001
age		1	-0.0313	0.00657	22.6984	<.0001
income		1	0.0175	0.00799	4.8146	0.0282
depressed		1	-0.8129	0.3004	7.3201	0.0068

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	0.969	0.957	0.982
income	1.018	1.002	1.034
depressed	0.444	0.246	0.799

Figure 12: Ordinal logistic regression of health on age, income, and depression, referent: **health** = Excellent

(h) Give careful interpretations of the coefficients of age, income and depression and their corresponding odds ratios for the model in part (g). Which of these variables appear to be significant predictors?

- Age: The log odds of better health status decrease by 0.0313 units for each additional year in age, all else equal. The odds of better health status are 3.1% lower for each additional year in age, all else equal.
- Income: The log odds of better health status increase by 0.0175 units for each additional $1000 in income, all else equal. The odds of better health status are 1.8% higher for each additional $1000 in income, all else equal.
- Depression: The log odds of better health status decrease by 0.8129 units if the person is depressed, all else equal. The odds of better health status are 55.6% lower if the person is depressed, all else equal.
- All three variables appear to be significant predictors ($p < .001, p = .028, p = .007$ for age, income, and depressed, respectively).

(i) Find the predicted probability of each level of health status for a 50 year old who makes $50,000 per year and is depressed based on the model from part (g).

$$
\begin{aligned}
X\beta_0 &= 0.8925 - 0.0313 * 50 + 0.0175 * 50 - 0.8129 * 1 = -0.6104 \\
X\beta_1 &= 2.9398 - 0.0313 * 50 + 0.0175 * 50 - 0.8129 * 1 = 1.4369 \\
X\beta_2 &= 4.3920 - 0.0313 * 50 + 0.0175 * 50 - 0.8129 * 1 = 2.8891 \\
P(Y=0) &= \frac{e^{X\beta_0}}{1+e^{X\beta_0}} = \frac{e^{-0.6104}}{1+e^{-0.6104}} = 0.352 \\
P(Y=0,1) &= \frac{e^{X\beta_1}}{1+e^{X\beta_1}} = \frac{e^{1.4369}}{1+e^{1.4369}} = 0.808 \\
P(Y=0,1,2) &= \frac{e^{X\beta_2}}{1+e^{X\beta_2}} = \frac{e^{2.8891}}{1+e^{2.8891}} = 0.947 \\
P(Y=0) &= 0.352 \\
P(Y=1) &= P(Y=0,1) - P(Y=0) = 0.456 \\
P(Y=2) &= P(Y=0,1,2) - P(Y=0,1) = 0.139 \\
P(Y=3) &= 1 - P(Y=0,1,2) = 0.053
\end{aligned}
$$

For a 50 year old who makes $50,000 per year and is depressed, the chance of having excellent health is 35.2%, the chance of having good health is 45.6%, the chance of having fair health is 13.9%, and the chance of having poor health is 5.3%.

(j) Compare and contrast your results with those from the multinomial logistic model. Do you think one of the models fits better than the other? Explain.

The results show that the predicted probabilities of excellent health and fair health are quite similar to those from the multinomial logistic model, but the predicted probabilities are slightly higher for good health in the ordinal logistic model than in the multinomial logistic model, and slightly lower for poor health in the ordinal logistic model than in the multinomial logistic model. In the multinomial logistic model, only age and depression status were overall significant variables, whereas in the ordinal logistic model, all three predictors were significant. The log likelihood of the multinomial logistic model is $\ell_m = \frac{614.040}{-2} = -307.02$, and the log likelihood of the ordinal logistic model is $\ell_o = \frac{621.475}{-2} = -310.74$. The log likelihoods are quite close (with the multinomial logistic model being less negative). The standardized difference of -2 times the difference in log likelihoods is $-2(-310.74 - (-307.02)) = 621.475 - 614.040 = 7.435$, and we can use this as a chi-square value since the standardized difference is close to a chi-square distribution. There are 6 extra parameters in the multinomial logistic model compared to the ordinal logistic model. When we calculate the probability (R command: pchisq(7.435,6, lower.tail = F)), we get $p = .282$. The p-value appears to be large and not significant, so the multinomial logistic model is not a significant improvement over the ordinal logistic model, so either model is suitable and there is not a clear difference in which model fits better.

(k) Do you think the proportional odds assumption has been met? Justify your answer conceptually and with an appropriate test.

The test statistic for the score test for the proportional odds assumption is $X_6^2 = 5.45$ ($p = .487$). We see that the test is not significant, indicating that the ordinal model likely meets the proportional odds assumption and is acceptable to use.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
5.4520	6	0.4873

Figure 13: Score test for the proportional odds assumption