Homework 4 Solutions

Question 2. (**50 points)** The Study of Assets and Health Dynamics among the Oldest Old (AHEAD) is a national longitudinal study with initial sample of 7444 respondents aged 70 years and older, and their spouses (if married). Objectives of the study include: (i) to monitor transitions in physical, functional, and cognitive health; (ii) to examine the relationship of late-life changes in physical and cognitive health to patterns of saving and income flows; (iii) to relate changes in health to economic resources and intergenerational transfers; (iv) to examine how the mix and distribution of economic, family, and program resources affect key outcomes, including institutionalization, saving, and health declines. In this problem, we look at data from the first four waves of the study, collected in 1993, 1995, 1998 and 2000. A reduced data set ahead.xlsx is on the course website. We explore the relationship of a test of cognitive function to some physical functioning indicators. The hypothesis is that, as physical function declines, movement becomes more difficult, and hence there are fewer stimuli, leading to cognitive decline. Here is a description of the variables contained:

id- subject identification number

year- study wave

sex- 1=male,2=female

age- age in years

immword- immediate word recall

delword- delayed word recall

blks- difficulty walking several blocks without help

strs- difficulty climbing flight of stairs without help

push- difficulty pulling or pushing a living room chair

bag- difficulty lifting a bag of groceries

dime- difficulty picking up a dime

iadlany- perform any instrumental activities of daily living

The focus of the study is to assess the association of difficulty performing instrumental activities of daily living (IADLs) to memory and cognition problems. IADLs include activities such as shopping, managing money, taking medications, using the telephone, preparing meals. They are supposed to represent the types of integrative activities that one needs to be able to perform in order to live independently in the community.

The response variable here will be total word recall (totword = immword + delword) and the main exposure is iadlany, an indicator for whether or not the subject reports difficulty with any (at least one) IADL. The goal of this analysis is to determine if there is a relationship between total word recall and the IADL variable using GEE. Consider the variables age and sex as possible confounders and/or effect modifiers for this analysis.

1. **(5 points) Discuss whether you feel that GEE is the appropriate analysis choice for this project.**

Either answer is acceptable with appropriate justification. It really depends on what you are interested in.

For GEE: GEE is a good choice because IADL is not a modifiable risk factor. Here, we are interested in comparing the population with IADL to the population without IADL. This will give us an idea of how these populations are different.

For GLMM: GLMM is a good choice because we are interested in the increased risks a person will have as they transition someone without IADL to someone with IADL. This will tell us what to expect in terms any increased risk for decreased total word recall.

1. **(45 points) Complete an analysis of this data using a GEE that results in the best possible answer to the scientific aims of the study. This should start with exploratory analysis of the data, explore different statistical models based on model fit and scientific hypotheses, and finish with a clear summarization of your findings including the interpretation of the coefficients of interest. Please include only relevant output to how you found the best model. It is encouraged to give the code to replicate your findings.**

**What is being asked?** Note that this question is not necessarily a “longitudinal” question (like a required for the midterm project). Let’s look at the goal of the study:

*“The goal of this analysis is to determine if there is a relationship between total word recall and the IADL variable using GEE.”*

This can be answer from cross-sectional data since all were interested in is the relationship between IADL and total word recall. It doesn’t necessarily need an interaction term between IADL and time/age. It *can* have an interaction term but that isn’t required. All we need to answer this question is to have IADL in the model.

**Analysis:** Since total word recall is a count variable, we’ll consider the Poisson and Negative Binomial models. Both use a log link. Below, I’ll take a look at the relationship between age and the log of average total word recall by the iadlany variable. The code to reproduce this output can be found in the .sas file.

Here, ‘log\_pred’ is the log of the average produced by a lowess smoother.

Chart, line chart

Description automatically generated

The pattern appears to be quadratic, not clear if iadlany has any impact.

To be complete, I’m also going to check if the pattern changes by sex.

Sex = 1

Chart, line chart

Description automatically generated

Sex = 2

Chart, line chart

Description automatically generated

Here the patterns appear to be the same, there doesn’t appear to be an interaction with sex and time or sex by time by iadlany. In the model below, I’ll include age as quadratic with an interaction with iadlany to see if there’s any change over time.

Overall, the negative binomial model fits much better than the Poisson model. Below, I’ll only give the results from the negative binomial.

There’s no need to put iadlany in the class statement as it is coded as 0/1.

**proc** **gee** data=ahead2;

class id sex (param=REF);

model totalwordrecall = sex age\_vary iadlany age\_vary\*iadlany age\_vary\*age\_vary age\_vary\*age\_vary\* iadlany/link=log dist=negbin;

repeated subject=id/corr=exch corrw;

**run**;

**The GEE Procedure**

| **Model Information** | |
| --- | --- |
| **Data Set** | WORK.AHEAD2 |
| **Distribution** | Negative Binomial |
| **Link Function** | Log |
| **Dependent Variable** | totalwordrecall |

|  |  |
| --- | --- |
| **Number of Observations Read** | 26604 |
| **Number of Observations Used** | 18907 |
| **Number of Missing Values** | 7697 |

| **Class Level Information** | | |
| --- | --- | --- |
| **Class** | **Value** | **Design Variables** |
| **sex** | **1** | 1 |
|  | **2** | 0 |

| **GEE Model Information** | |
| --- | --- |
| **Correlation Structure** | Exchangeable |
| **Subject Effect** | id (6651 levels) |
| **Number of Clusters** | 6651 |
| **Clusters With Missing Values** | 3655 |
| **Correlation Matrix Dimension** | 4 |
| **Maximum Cluster Size** | 4 |
| **Minimum Cluster Size** | 0 |

| **Working Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Obs 1** | **Obs 2** | **Obs 3** | **Obs 4** |
| **Obs 1** | 1.0000 | 0.4451 | 0.4451 | 0.4451 |
| **Obs 2** | 0.4451 | 1.0000 | 0.4451 | 0.4451 |
| **Obs 3** | 0.4451 | 0.4451 | 1.0000 | 0.4451 |
| **Obs 4** | 0.4451 | 0.4451 | 0.4451 | 1.0000 |

| **Exchangeable Working Correlation** | |
| --- | --- |
| **Correlation** | 0.4451 |

| **GEE Fit Criteria** | |
| --- | --- |
| QIC | -306649.9960 |
| QICu | -306654.7803 |

| **Parameter Estimates for Response Model** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **with Empirical Standard Error Estimates** | | | | | | | |
| **Parameter** |  | **Estimate** | **Standard Error** | **95% Confidence Limits** | | **Z** | **Pr > |Z|** |
| **Intercept** |  | -0.6943 | 0.6721 | -2.0116 | 0.6231 | -1.03 | 0.3016 |
| **sex** | 1 | -0.1036 | 0.0098 | -0.1228 | -0.0844 | -10.58 | <.0001 |
| **age\_vary** |  | 0.0926 | 0.0170 | 0.0593 | 0.1259 | 5.45 | <.0001 |
| **iadlany** |  | 1.3535 | 1.2557 | -1.1077 | 3.8146 | 1.08 | 0.2811 |
| **age\_vary\*age\_vary** |  | -0.0007 | 0.0001 | -0.0009 | -0.0005 | -6.76 | <.0001 |
| **age\_vary\*iadlany** |  | -0.0352 | 0.0311 | -0.0962 | 0.0257 | -1.13 | 0.2572 |
| **age\_va\*age\_va\*iadlan** |  | 0.0002 | 0.0002 | -0.0002 | 0.0006 | 1.06 | 0.2895 |

Here, we can see that age has a clear quadratic effect on the outcome. It doesn’t appear that iadlany has any interactive effect.

I’m going to remove the interaction term here. Note that the interaction term isn’t necessary to test the hypothesis of interest “”. As a result, the model below is informative and important to look at.

**proc** **gee** data=ahead2;

class id sex (param=REF);

model totalwordrecall = sex age\_vary iadlany age\_vary\*age\_vary/link=log dist=negbin;

repeated subject=id/corr=exch corrw;

**run**;

**The GEE Procedure**

| **Model Information** | |
| --- | --- |
| **Data Set** | WORK.AHEAD2 |
| **Distribution** | Negative Binomial |
| **Link Function** | Log |
| **Dependent Variable** | totalwordrecall |

|  |  |
| --- | --- |
| **Number of Observations Read** | 26604 |
| **Number of Observations Used** | 18907 |
| **Number of Missing Values** | 7697 |

| **Class Level Information** | | |
| --- | --- | --- |
| **Class** | **Value** | **Design Variables** |
| **sex** | **1** | 1 |
|  | **2** | 0 |

| **GEE Model Information** | |
| --- | --- |
| **Correlation Structure** | Exchangeable |
| **Subject Effect** | id (6651 levels) |
| **Number of Clusters** | 6651 |
| **Clusters With Missing Values** | 3655 |
| **Correlation Matrix Dimension** | 4 |
| **Maximum Cluster Size** | 4 |
| **Minimum Cluster Size** | 0 |

| **Working Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Obs 1** | **Obs 2** | **Obs 3** | **Obs 4** |
| **Obs 1** | 1.0000 | 0.4449 | 0.4449 | 0.4449 |
| **Obs 2** | 0.4449 | 1.0000 | 0.4449 | 0.4449 |
| **Obs 3** | 0.4449 | 0.4449 | 1.0000 | 0.4449 |
| **Obs 4** | 0.4449 | 0.4449 | 0.4449 | 1.0000 |

| **Exchangeable Working Correlation** | |
| --- | --- |
| **Correlation** | 0.4449 |

| **GEE Fit Criteria** | |
| --- | --- |
| QIC | -306789.1010 |
| QICu | -306794.1152 |

| **Parameter Estimates for Response Model** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **with Empirical Standard Error Estimates** | | | | | | | |
| **Parameter** |  | **Estimate** | **Standard Error** | **95% Confidence Limits** | | **Z** | **Pr > |Z|** |
| **Intercept** |  | -0.5051 | 0.5841 | -1.6499 | 0.6397 | -0.86 | 0.3872 |
| **sex** | 1 | -0.1035 | 0.0098 | -0.1227 | -0.0843 | -10.57 | <.0001 |
| **age\_vary** |  | 0.0883 | 0.0147 | 0.0595 | 0.1171 | 6.01 | <.0001 |
| **iadlany** |  | -0.1576 | 0.0090 | -0.1751 | -0.1400 | -17.56 | <.0001 |
| **age\_vary\*age\_vary** |  | -0.0007 | 0.0001 | -0.0009 | -0.0005 | -7.59 | <.0001 |

Here, we can see that iadlany clearly has an impact. 1-exp(-0.1576)= 0.146 or about a 15% decrease. As a result, we can interpret the coefficient as follows:

*We estimate that the average rate of total word recall is 15% lower for those with iadlany=1 versus 0, after controlling for age and sex.*

The following analysis only uses a linear effect of age and an interaction between linear age and iadlany. I think the above is better (I’ll discuss more why below)

**The GEE Procedure**

| **Model Information** | |
| --- | --- |
| **Data Set** | WORK.AHEAD2 |
| **Distribution** | Negative Binomial |
| **Link Function** | Log |
| **Dependent Variable** | totalwordrecall |

|  |  |
| --- | --- |
| **Number of Observations Read** | 26604 |
| **Number of Observations Used** | 18907 |
| **Number of Missing Values** | 7697 |

| **Class Level Information** | | |
| --- | --- | --- |
| **Class** | **Value** | **Design Variables** |
| **sex** | **1** | 1 |
|  | **2** | 0 |

| **GEE Model Information** | |
| --- | --- |
| **Correlation Structure** | Exchangeable |
| **Subject Effect** | id (6651 levels) |
| **Number of Clusters** | 6651 |
| **Clusters With Missing Values** | 3655 |
| **Correlation Matrix Dimension** | 4 |
| **Maximum Cluster Size** | 4 |
| **Minimum Cluster Size** | 0 |

| **Working Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Obs 1** | **Obs 2** | **Obs 3** | **Obs 4** |
| **Obs 1** | 1.0000 | 0.4447 | 0.4447 | 0.4447 |
| **Obs 2** | 0.4447 | 1.0000 | 0.4447 | 0.4447 |
| **Obs 3** | 0.4447 | 0.4447 | 1.0000 | 0.4447 |
| **Obs 4** | 0.4447 | 0.4447 | 0.4447 | 1.0000 |

| **Exchangeable Working Correlation** | |
| --- | --- |
| **Correlation** | 0.4447 |

| **GEE Fit Criteria** | |
| --- | --- |
| QIC | -306848.6686 |
| QICu | -306853.2183 |

| **Parameter Estimates for Response Model** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **with Empirical Standard Error Estimates** | | | | | | | |
| **Parameter** |  | **Estimate** | **Standard Error** | **95% Confidence Limits** | | **Z** | **Pr > |Z|** |
| **Intercept** |  | 3.8478 | 0.0601 | 3.7299 | 3.9657 | 63.98 | <.0001 |
| **sex** | 1 | -0.1034 | 0.0098 | -0.1226 | -0.0842 | -10.57 | <.0001 |
| **age\_vary** |  | -0.0223 | 0.0008 | -0.0238 | -0.0208 | -29.17 | <.0001 |
| **iadlany** |  | 0.2064 | 0.1131 | -0.0153 | 0.4282 | 1.82 | 0.0680 |
| **age\_vary\*iadlany** |  | -0.0046 | 0.0014 | -0.0074 | -0.0018 | -3.26 | 0.0011 |

Here we see that there is an interaction between age and iadlany. The interaction term is negative (main effect is positive) which indicates the coefficient for iadlany decreases with age.

Here, are some interpretations from this model:

Effect of iadlany at age 70:

exp(0.2064-70\*0.0046) = 0.8908315 or a 10.9% decrease.

For 70 year old’s, the population with at least one difficulty with an IADL has 10.9% lower total word recall than the population that doesn’t have any difficulty with an IADL.

Effect of iadlany at age 80:

exp(0.2064-80\*0.0046) = 0.8507814or a 14.9% decrease.

For 80 year old’s, the population with at least one difficulty with an IADL has 14.9% lower total word recall than the population that doesn’t have any difficulty with an IADL.

What worries me about this model, is that it goes against what we saw in our graphical analyses. Mainly in the following graph:

Chart, line chart

Description automatically generated

The differences appear to be getting smaller with age. As a result, I think the quadratic approach (without the interaction term) is probably the best way to go.