Answers to the Review Questions for the Final Exam: 
Vartanian: Data II

1. **Cell 1:** Father’s education has a more negative effect on the likelihood of growing up poor relative to growing up middle class. That is, the higher the level of education, the less likely you are to be in category 1 relative to the excluded category. 
   --Mother’s education has a more negative effect on growing up poor relative to growing up middle class. The higher the level of education of the mother, the less likely you are to be in category 1 relative to the excluded category. 
   --EDDADWF has more negative effect on growing up poor relative to growing up middle class. The higher the education of the wife’s dad, the less likely you are to be in category 1 relative to the excluded category. 
   --EDMOMWF has no effect.

Cell 5: The higher the level of education of the dad, the more likely the person will have grown up rich relative to growing up middle class. The higher the value of EDDAD, the more likely you are to be in category 5 relative to the excluded group. 
   - The higher the level of education of the mom, the more likely the person will have grown up rich relative to growing up middle class. The higher the value of EDMOM, the more likely you are to be in category 5 relative to the excluded group. 
   - The higher the level of education of the father of the wife, the less likely the person grew up rich relative to growing up middle class. The higher EDDADWF, the less likely you are to be in category 5 relative to the excluded category. 
   - EDMOMWF has no effect on the outcome.

From the second set of regressions, we can determine the differences between category 5 (Rich) and category 1 (Poor).

Cell 5: The higher the level of education, the more likely you are to be in the rich group relative to the poor group. 
   - The higher the level of mother’s education, the more likely you are to be in category 5 relative to category 1. 
   - The other variables are not related to differences in category 5 and category 1.

B. If the Hausman test showed significance, this would indicate that we reject the null of iia, and we would therefore need
to run another type of test.

#2.

A. If we have a relatively large number of observations with 0 hours of work, our estimates for hours of work may be biased. We run the selection model, examining factors that may affect whether or not someone works, and the correct for our estimates by this Heckman process. In other words, the Heckman selection model allows us to use information from non-working women to improve the estimates of the parameters in the regression model. The Heckman selection model provides consistent, asymptotically efficient estimates for all parameters in the model.

b. There seem to be slight differences in the models. The Heckman model seems to show that the effects of being married are less than what the OLS model indicates. It also shows that the effects of owning your own home is more negative than in the OLS model. The biggest difference appears to be in the effects of health of the head. The heckman models shows that the effects of having excellent health are quite a bit lower than the OLS model – an additional 92 hours relative to those who don’t have excellent health, versus 104 in the OLS model for those with excellent health relative to those with less than excellent health.

C. This tests to determine if the error terms in the two models are related to each other. If they are not, then the OLS regression estimates are unbiased. If they are related to one another then the OLS regression estimates are biased. The null hypothesis is that Rho =0. Here, we reject the null hypothesis and state that Rho ≠0, in all likelihood, and we have better estimates with our selection model.

3.

A. 

\[
p = \frac{e^{0.05+1.5+0.0003*10,000+(0.950)^2}}{1+e^{0.05+1.5+0.0003*10,000+(0.950)^2}} = \frac{0.44}{1.44} = 0.31
\]
In other words, there is a 31% chance that someone will buy a car.

B.

\[ p = \frac{e^{0.05 + 1.5 \times 1.0 + 0.0003 \times 20,000 + (-0.950) \times 1.0}}{1 + e^{0.05 + 1.5 \times 1.0 + 0.0003 \times 20,000 + (-0.950) \times 1.0}} = \frac{e^{1.2}}{1 + e^{1.2}} = \frac{3.32}{4.32} = 0.7685 \]

Thus, with these characteristics, the probability of buying a car increases from 31% to 77%.

C.

\[ e^{3 \times -0.950} = e^{-2.85} = 0.0578 \]
\[ e^{0 \times -0.950} = e^0 = 1 \]

In other words, people with 3 kids are .0578/1 as likely to buy a car as those people with 0 kids. Or, people with 0 kids are 1/(0.0578) = 17.30 as likely to buy a car as those people with 3 kids.

D. Females are 4.48 times as likely to buy a car as males, controlling for the other variables within the model.

\[ e^{1.50} = 4.48 \]

5.

A. Logistic Regression is your best choice.

B. DV=likelihood of drug use
   IV=Peer pressure, either measured at the interval or nominal level.
   Control=Family income, district income and whether or not the person smokes.
C. The null hypothesis is that peer pressure will have no effect on the likelihood of drug use.

D. You'll accept the null hypothesis in the first model where peer pressure is measured at the nominal level. In the second model, where peer pressure is measured at the interval level, you'll reject the null hypothesis at the .05 level.

E. For people who experience peer pressure, have $0 in family and district income, and do not smoke, the probability of drug use is

\[ P(\text{drug use}) = \frac{e^{1 + (0.5)(0) + 0 + 0}}{1 + e^{1 + (0.5)(0) + 0 + 0}} \]

\[ = \frac{4.48}{5.48} = 0.8178 = 81.78\% \]

The probability for those people who do not experience peer pressure but who smoke is

\[ P(\text{drug use}) = \frac{e^{1 + (0)(0) + 0 + 2 + 1}}{1 + e^{1 + (0)(0) + 0 + 2 + 1}} \]

\[ = \frac{20.09}{21.09} = 0.9524 = 95.24\% \]

5.

A and B.

Children with parents who drink a lot of caffeine are found to be more likely to throw rocks than children with parents who drink little or no caffeine. You will thus accept your null hypothesis that there is no relationship between caffeine drinking parents and rock throwing since the relationship is in the opposite direction of your research hypothesis. Those children who ride motor scooters to school are shown to be less likely to throw rocks at birds. In fact, this is
the best predictor we have of whether or not a child will throw rocks at birds. Children who ride motor scooters are only 1.8% as likely to throw rocks as children who do not ride motor scooters. This means that children who do not ride motor scooters are 54.598 times as likely to throw rocks as children who ride motor scooters (1/0.0183). As predicted, Libras are more likely to throw rocks, but unexpectedly, Scorpios are less likely to throw rocks than non-Scorpios.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Null Hypothesis</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffeine:</td>
<td>Accept</td>
<td>49% more likely for those who drink 1 ounce versus those who drink 0 ounces of caffeine.</td>
</tr>
<tr>
<td>Scorpio:</td>
<td>Accept</td>
<td>13.53% as likely</td>
</tr>
<tr>
<td>Libra:</td>
<td>Reject</td>
<td>7.389 times as likely</td>
</tr>
<tr>
<td>Scooter:</td>
<td>Reject</td>
<td>1.83% as likely</td>
</tr>
</tbody>
</table>

C. Yes. You've made a type II error. This is when you accept the null hypothesis when there truly is a relationship between the variables.

D. The likelihood of rock throwing for kids who ride a scooter, are Scorpios, and have parents who do not drink caffeine:

\[
P(\text{stone throwing}) = \frac{e^{1+(-2) \times 1+0+2+(-4) \times 1+(40) \times 0}}{1+e^{1+(-2) \times 1+0+2+(-4) \times 1+(40) \times 0}} = \frac{0.0067}{1.0067} = 0.006693 = 0.6693\%
\]

The most highly significant variable is riding a scooter. So please, parents, for the sake of our birds, give your children scooters.

6.

A. You will not use those observations that are left censored. Since we do not know the beginning of their spell, we cannot know how long they've been unemployed. You can use the right censored observations since we know when they started and we can then use this information to determine the likelihood of
getting off of the state of unemployment in month 1 or month 2 (if they're censored in month 3). You'll use a hazard rate regression model to determine the likelihoods of exiting an unemployment state, given that you're still in the state.

B. You can use a set of dummy variables that will indicate the time you've spent on welfare to date. These dummy variables could be name something like month1-month6+, where month6+ stands for months 6 and above. You may need to do this if too few observations are unemployed for 7 or more months. Remember that there must be variation in the variable in order to determine parameter estimates. Also remember that like any set of dummy variables you must omit one of the variables and the parameter estimates for the set of included variables are relative to the excluded category. In other words, these parameter estimates for month2-month6+ will tell you the likelihood of exiting unemployment relative to the first month. They may thus give you information on duration dependence -- the longer you're unemployed, the lower is your likelihood of exiting the state. This will be true when your coefficient estimates for time unemployed are negative and significant the longer you're unemployed. However, even if these time coefficients are negative and significant, it does not necessarily mean that there is duration dependence. It may mean that there is unobserved heterogeneity within the population that you're examining. That is there may be unobserved differences between people that you are unable to measure that are instead showing up in the coefficient estimates of the time variables. If every individual has the same probability of exit regardless of time they've been unemployed, the exit probabilities for a group of heterogeneous individuals will decline with time on the program. This is because the exit probability will be based on all individuals in the early months, including those with a high exit probability and those with a low exit probability. In later months the exit probabilities are based on only those individuals who stay unemployed. Those who remain will have lower exit probabilities than those who have left. Therefore the exit probability will be lower than in the first month.

C. The significant variables (at the .05 level) are kids, age, at the beginning of the spell, living in the South and West relative to the East, and months 3 and above relative to month 1.
D. The hazard rate is

$$Hazard = e^{(-.5d + 1.1 + 0.05*25)} = 19.2\%$$

Given that you're in the sample, there is a 19.2% chance of leaving the state of unemployment in the first month.

7. No question 7.

8. 

A. You'll probably choose either 2 or 3 principal components since the first three eigenvalues explain 92.5% of the variation in the 8 variables.

The first principal component appears to be measuring the income level of the neighborhood. The higher the level of this principal component, the lower is the income level.

The second principal component appears to be measuring the employment circumstances of the neighborhood. A high value for the principal component indicates that employment levels for both males and females is relatively high.

The third principal component appears to be measuring the types of jobs that are worked in the neighborhood. A relatively high value for this component means more people are working in professional and managerial jobs, while a lower number or negative number indicates that people in the neighborhood are working manufacturing and service sector jobs.

B. You could use these principal components within regression models or any other type of statistical model that uses interval scale variables.

C. The b coefficient for Prin1 indicates that as income levels in the neighborhood decrease, the length of spell also decreases. The b coefficient for Prin2 indicates that as employment rates increase in the neighborhood, employment lengths also increase. The coefficient for Prin3 indicates that the greater the percentage of manufacturing and service sector jobs, and the lower the level of professional and executive jobs, the shorter will be the employment spell. However, this is not a significant relationship so we will conclude that there is no difference among areas with different types of jobs.
9.

A. For those who are married, they are significantly less likely to have low income relative to those who are not married. The odds ratio indicates that they are around 25.9% as likely to be at the next lowest level of income relative to those who are married. Hours of work of the head has a significant negative effect on the likelihood of being poorer. Number of kids has a negative effect on the likelihood of having lower income. For each additional child in the household, there is an 85% increase in the likelihood of being at the next lowest income level.

B. The chi-square value for the proportional odds assumption is significant, so no, this is not an appropriate model for these independent and dependent variables. The independent variables do not have proportionate effects on the different levels of the dependent variable.

C. Yes, the -2 Log of the Likelihood indicates significance at the p=.0001 level.

D. In this model, the lowest level of the variable is excluded, or level 1 is excluded. We can determine the likelihood of being in level 1 by subtracting the overall likelihood for being in response value 2 from 1. Or 1-.831 = .169. Or there is a 16.9% chance of being in the highest income category. You could check this by examining the response profile, examining the totals in each of the categories and the relative to the total sample. In this case, it would be 981.4912/5805 = 16.9%. For those families with 3 children, this likelihood is 1-.946 = .054, or 5.4%.

To examine the likelihood of being lowest income category, you could examine response value=5. This indicates that 5.059% are in this lowest category. For those with 3 children, this likelihood is 16.09%.

For the second highest category of income, this likelihood is .38 -.05 = .33, or a 33% likelihood of being in the second highest income category. For those with 3 kids, this likelihood is .65-.16 = .49 or 49%.

For the third highest category, the value is .588-.383 = .205, or 20.5%. For those with 3 kids, the values are .82-.65 = .17, or 17%. For the category 3 from the response profile, 1197.056/5805= 20.6.
Etc.

E. Given above.

10. The interaction indicates that the interactive effects of kids and marriage are: Kids have a more negative effect on lowering income for married families than on non-married families. This is the interpretation because higher values of the dependent variable are associated with lower levels of income. In other words, kids have a more positive effect on income in married families than in non-married families.

11. 

A. The time variables (work2-work10) indicate that there is a positive effect of time on the likelihood of leaving a state of employment. These effects do not get larger or more significant the longer someone has been employed.

B. Job training weeks has a positive effect on the likelihood of leaving a state of employment. That is, the longer you’ve been trained, the more likely you’ll leave employment.

C. The hazard for leaving employment in month 1 is .0078. In month 2 the hazard is .04588, and in month 3 the hazard is .057. In month 4 this is .0933. These are given by the Pr_XB probabilities.

D. The overall likelihood of survival in month 1 = 1 -.0078 = .9922. The likelihood of survival in month 2 = .9922 - .0459*.9922 = .9467. In month 3 the overall likelihood of survival = .9467 - .057*.9467 = .8927. In month 4 the overall likelihood of survival = .8927 - .093*.8927 = .8097. Thus, the overall likelihood of survival through 4 months is around 80.97 percent.

E. The hazard for those with health insurance in month 1 and month 2 are: .003 and .019, respectively. For those without HI, these hazards are .009 and .054. The survival likelihoods for those with health insurance are:

<table>
<thead>
<tr>
<th>Period</th>
<th>For those with HI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>1 - .003 = .997</td>
</tr>
<tr>
<td>2:</td>
<td>.997 - .019*.997 = .978</td>
</tr>
<tr>
<td>3:</td>
<td>.978 - .024*.978 = .955</td>
</tr>
</tbody>
</table>
Period: For those without HI
1: $1 - .009 = .991$
2: $.991 - .054*.991 = .937$
3: $.937 - .068*.937 = .873$

So, for those with health insurance, the likelihood of survival in employment is around 95.5 after 3 months, while for those without health insurance, this likelihood is 87.3%.

12.

A. The Hausman test indicates that it is not appropriate to run a random effects model.

B. The random effects model has more statistical power because we do not lose all those degrees of freedom by using dummy variables for each family.

C. We will use the fixed effects model. Birth order has a negative effect on the log of income – for each additional child before this child, income decreases by around 3.5%. If we allow a .10 significance level, the coefficient for move indicates that income decreases by around 7.4 if the person moves during adolescence. (Use the eb formula for log dependent variables for determining the exact meaning of these coefficient estimates.)

D. Within family differences explain around 1.17% of the variance of the dependent variable.

E. 1121.