Introduction to Statistics: Homework 2 (Model Answers)

Dichotomous/Nominal Variables and Non-linear Functional Forms

1. A 1976 dataset provides a way to examine the relationship between a number of individual-level characteristics and monthly wages. The outcome variable is monthly wages. Here are some regression results:

	Coef.	SE	T	P
Black (1=yes)	-241.332	38.472	-6.270	0.000
Lives in the South (1=yes)	-72.764	27.235	-2.670	0.008
Lives in an Urban Area (1=yes)	183.197	27.934	6.560	0.000
Age (in years)	19.314	4.007	4.820	0.000
Constant	243.356	135.558	1.800	0.073

a. The constant means that individuals who are not Black, do not live in the south, do not live in an urban area, and who are (theoretically) zero years old are expected to make $243.36 per month. Because the p-value is greater than .05 we can not be confident that the estimate 243.36 is statistically distinguishable from zero.

b. The coefficient on *Black* indicates that – controlling for the other variables in the model – black respondents are expected to make $241.33/month less than non-black respondents. Because the p-value is less than .05 we can be 95% confident that this difference is different from zero.

The coefficient on *Lives in the South* indicates that – holding other variables in the model constant – individuals living in the South are expected to earn $72.76/month less than those not living in the South. Again, because the p-value is less than .05 we can say that this difference is statistically significant (distinguishable from zero).

The coefficient on Lives in an Urban Area means that – after controlling for the effects of *Black*, *Lives in the South*, and *Age* – individuals living in urban areas are expected to earn $183.20/month more than those not living in an urban area. Because the p-value is below .05 (and, correspondingly, because the absolute value of the t-statistic is greater than 2) we can be confident that this difference is different from zero.

The coefficient on *Age* indicates that – holding other variables in the model constant – for each additional year of age we expect an individual to make $19.34/month more. Again, the p-value tells us that we can be confident that this relationship is different from zero.

c. 243.36 – 241.33 + 0 + 183.20 + 30*19.31 = 764.53

d. The positive relationship between age and wages may reflect a dynamic where employers feel obligated to pay older workers more simply because they are older. I would guess that this estimate of the relationship between age and wags is biased. For example, “years of experience” is likely to be strongly correlated with both age and wages. By not controlling for this confound, we may well be overestimating the independent effect of age on wages.

2. For items in this section use the turnout2008 dataset. The dependent variable we’ll focus on is turnout in the 2008 presidential election. [Note: there are more advanced statistical approached for dealing with dichotomous dependent variables. However, they rarely lead to meaningfully different findings. In this case, think about predicted values as the predicted probability that an individual turns out.]

Variable	Description
Turnout	Voted in 2008 (1=yes)
Anycontact	Report being contacted by Democratic or Republican campaign (1=yes)
strength_pid	Strength of party affiliation (0=pure independent; 1="leaner"; 2=weak identifier; 3=strong identifier)
Education	Education (1=no high school; 2=some HS, no diploma; 3=HS diploma; 4=some college; 5=associate degree; 6=bachelor's degree; 7 advanced degree)

a.

i. The coefficient on the constant is .662. This is the expected probability of turning out to vote among people who did not report being contacted by a political party.

ii. The coefficient on contact is .214. This means that the model predicts that people who reported being contacted by a party were 21.4 percentage points more likely to report turning out to vote than those who reported that they had not been contacted. The fact that the p-value is less than .05 means that we can be 95% confident that this difference is not due to chance – i.e., that the difference is statistically distinguishable from zero.

b.

i. The coefficient on the constant is .233. This means that the model predicts that pure independents who reported not being contacted by a party and who have an education level of zero (which is not even part of the education scale) are expected to have a 23.3% chance of turning out.

ii. The coefficient on *anycontact* is .154 which means that – controlling for other variables in the model – those who reported being contacted by a party are expected to turn out 15.4 percentage points more frequently than those who report not having been contacted. The p-value is less than .05, indicating that this difference is statistically significant.

The coefficient on *strength_pid* means that – holding other variables in the model constant – each one unit increase in strength of party attachment is associated with an expected 11.7 percentage point increase in the likelihood that an individual will turn out. Again, the p-value is below .05, indicating that this relationship is statistically distinguishable from zero.

The coefficient on *education* is .061. This means that, controlling for other variables in the model, each one unit increase in education is associated with a 6.1 percentage point increase in the likelihood that an individual turned out in 2008. Again, the p-value is below .05, indicating that this relationship is statistically significant (distinguishable from zero).

iii. The coefficient on *anycontact* dropped from .214 to .154. There is no clear benchmark for whether this change is substantial. To the extent that the coefficient did change, it is a product of the relationship between the likelihood of being contacted by a party and the two IVs that were added to the model, as well as the relationship between those variables and turning out to vote. For example, individuals who are strong partisans may be more likely to be contacted by their political party than independents. If we fail to account for this relationship, we may overestimate the effects of being contacted because that relationship is confounded by other factors that might affect turnout (e.g., strength of partisan attachments).

iv. The estimate of the effect of turnout is probably still biased. There are probably a variety of other factors that are associated with both turnout and likelihood of being contacted by a political party. For example, parties may be more likely to contact people who have voted before because they think they have a better chance of getting them to turnout than people who have never voted. Because past turnout behavior was also probably associated with the likelihood of turning out in 2008 the estimate of the effects of being contacted may be biased upward. Controlling for past turnout would reduce this bias.

c.

i. The coefficient on the constant is .587. This is the predicted probability of turning out for a pure independent with a bachelor’s degree who reports not having been contacted by a party.

ii. The coefficient on “advanced degree” is .031, indicating that, after controlling for other variables in the model, those with an advanced degree are expected to be 3 percentage points more likely to turn out than an individual with a bachelor’s degree. However the p-value is .419. Therefore, we can not be confident that there is a difference in the likelihood that the likelihood of those with a bachelor's degree turning out is different than the likelihood that of someone with an advanced degree. The coefficient on *some college* is -.058. This means that, controlling for other variables in the model, those with some college are 5.8 percentage points less likely to turn out than those with a bachelor's degree. This difference is statistically significant (p-value is less than .05).

iii. .587 + 2*.117 - .162 + .154 = .813

iv. To test the difference between those with some college and other education groups we want to rerun the model swapping the excluded category. In other words, we want to include the indicator for bachelor's degree and exclude the indicator for some college. The coefficient on *associates degree* in that model is .002. However, the p-value on the coefficient is greater than .05. Therefore we can not reject the null hypothesis that, after controlling for the other variables in the model, these two education groups have the same likelihood of turning out – i.e., we can not be confident that they are different.

v. Using the same model estimated for iv., we find that the coefficient on high school grad is -.245. The p-value is less than .05, so we can be confident that, controlling for the other variables in the model, those with some college were more likely to vote than those with only a HS diploma. Specifically, they were 24.5 percentage points more likely to vote.

3. For these items use the literacy dataset. This dataset is from 1970 and has information about countries around the world. The following variables are included:

Variable	Description
LITERACY	Literacy rate (number of literate residents per 1000)
GDPCA	GDP per capita
GDPCA2	GDP per capita (squared)
DEMOCRAC	Democracy (1=yes)

Estimate a model predicting literacy rate with the other three variables in the dataset.

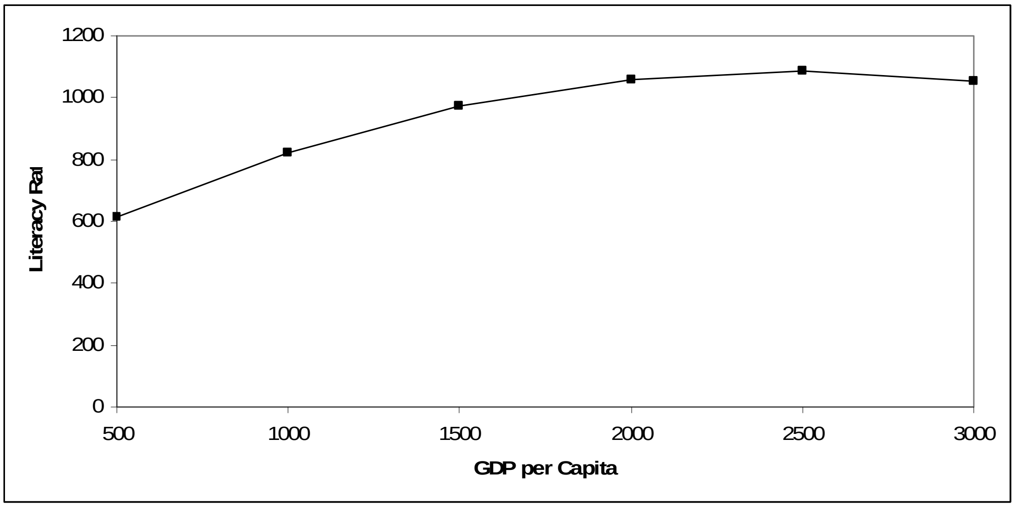
i. The coefficient on GDPCA2 tells us that allowing the relationship between GDP per capita and literacy to "bend" significantly improves the fit of the model. We know because the p-value is less than .05.

ii. No, if the coefficients on GDPCA and GDPCA2 were both statistically insignificant, we should not necessarily conclude that there is no relationship between GDP per capita and literacy rates. Because these variables are likely to be highly correlated with one another (in fact, in this dataset they are correlated at .934), taken together (jointly) they may significantly improve the fit of the model. We could test this using an F-test.

iii.

GDP per capita	Predicted literacy rate
500	238.8699+107.975+.5982508*500-.0001207*(500^2) = **615.7953**
1000	**824.3957**
1500	**972.6461**
2000	**1060.5465**
2500	**1088.0969**
3000	**1055.2973**

iv. The graph shows that increasing levels of GDP per capita is associated with higher literacy rates. However, the relationship flattens out around $2500 per capita GDP and then actually begins to decline.


Literacy Rat
GDP per Capita