Hamzah Ahmed
Econ 613
Assignment 2

***Note to Grader(s): Sometimes I would rerun my code a handful of times and, for the parts where randomness is used (such as for optimization), the outputs in my markdown may be different than those on this document. In hindsight, I could have set a seed but I was having issues initially with the optimization and seeding every time would not have allowed to me to work through my issues.**

***Also, dropped people under 18 from the data for all problems (professor said to in Slack) (in markdownNoKids).**

Exercise 1

I filtered out people with missing wage and age information as well as people with 0 wage.

a. Correlation is defined as $\rho = \frac{cov(X,Y)}{sd(X) \cdot sd(Y)}$. X was defined as age and Y was defined as wage and those were the values I used. Output was:
[1] 0.1398806

b. Used the given formula and use %*% for the matrix multiplication (normal * would be entry-wise which is not what we want). Added a column of 1's to X data to indicate intercept is included. Output was:
intercept: 14362.5594
Age: 226.3447

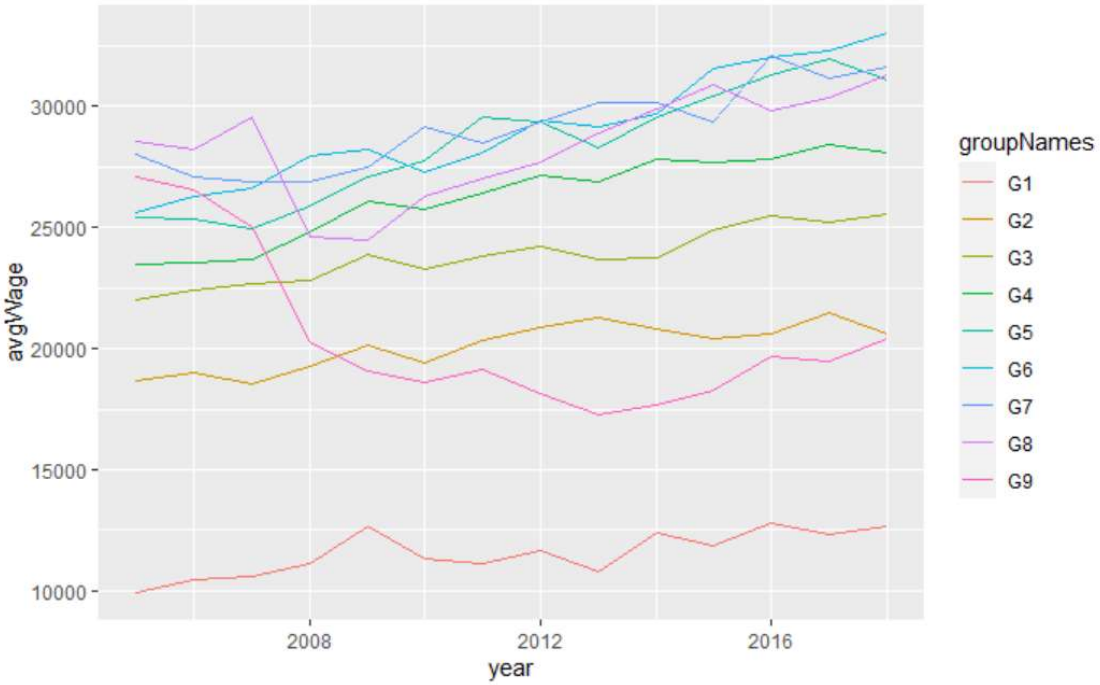
c. OLS SE was calculated by the following process: $\widehat{\sigma^2} = \frac{\sum(\hat{y}-y)^2}{n-k+1}$ where $\hat{y}$ was calculated by multiplying the $\hat{\beta}$ by the age values (I assumed homoscedasticity but I think that it may be heteroskedastic). Then got the variance covariance matrix (by doing $inv(X'X)$ times what I got. With bootstrapping, I did it twice, once with 49 replications and once with 499 replications. Both situations were quite similar (just the latter was repeated 450 more times). My sample size was the size of the dataset but I sampled with replacement so there were possibilities were some people were included multiple times and others were omitted and I found the $\hat{\beta}$ of this "resample". I collected all of them and then found the square root of the variance and that was my SE estimate through bootstrapping. Outputs (OLS, 49 Replications, 499 replications) were:
[1] 650.70041 (intercept) and 14.98449 (Age)
[1] 595.7059 (intercept) and 15.41032 (Age)
[1] 593.7494 (intercept) and 15.78028 (Age)

I actually ran this code multiple times and the bootstrapped values varied a bit. They were generally both quite close to each other but the intercept SE was slightly lower than OLS and the coefficient SE was slightly higher. Bootstrapping is kind of a resampling process in which the weights of each observation can vary where some observations have higher weights than others and some observations have no weights (as opposed to with OLS which gave equal weight to each observation). From class, we learned that bootstrapping randomly changes weights to potentially highlight influential estimates. We can take the variance of our $\beta's$ to get an informative estimate of our variance of β using all the observations equally. This process can be beneficial in cases where our number of observations is really big. Between the two bootstrapping techniques, one just had more "sample data" as it was repeated more times

(both had sample sizes equal to the amount of data entries). Difference between bootstraps were small and due to randomness in selection. Difference between bootstrap and OLS was due to random subsampling with various weights versus using total data with equal weights. Since bootstrapped values were similar, may indicate that there aren't any super influential estimates that could lead to outliers. OLS requires that we have a closed form solution to the estimator, in the case that we don't, bootstrapping gives us a way of estimating variance and it can potentially more convenient with decent approximations.

Hamzah Ahmed
Econ 613
Assignment 2
Exercise 2

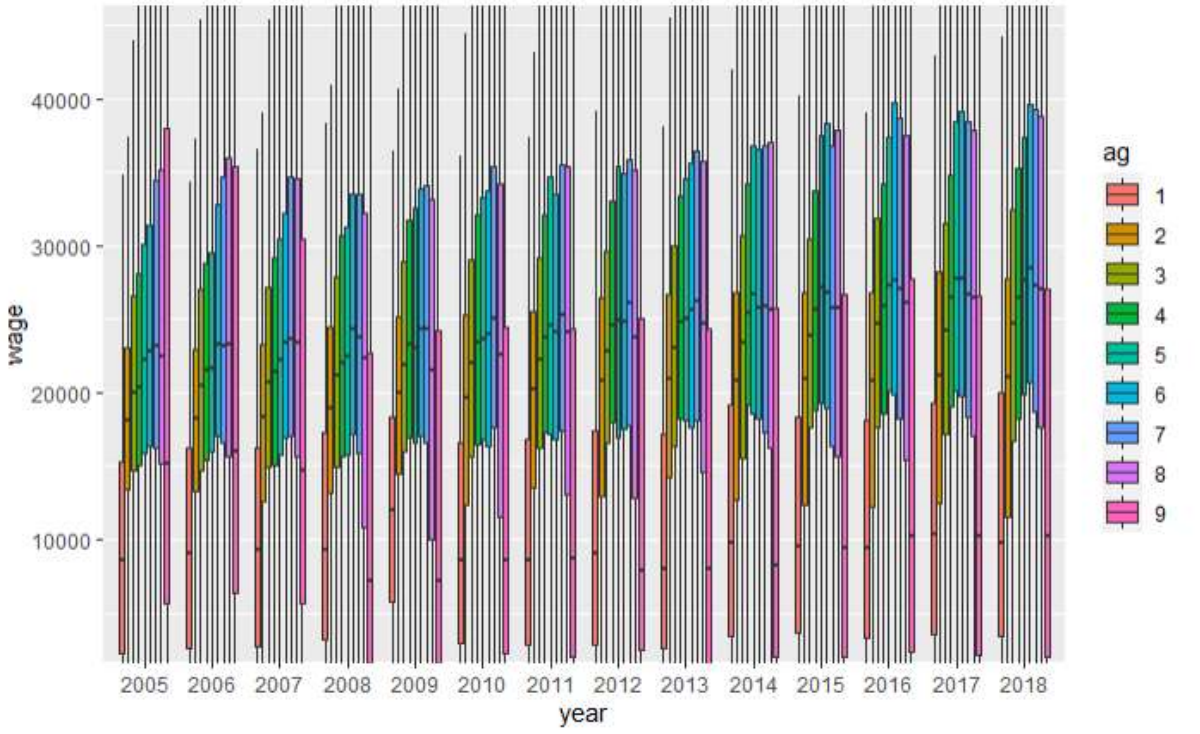
a. Dropped people who had 0 wage. Made the ag categorical variables and labeled them 1-9 in order by the age bins given (under 18 were dropped). I then found the average wage in each of these age groups in each year and then plotted them using ggplot. I had to append all the data together to better do multiple lines and added all this data to wageAndAgePerYear. The plot is below:



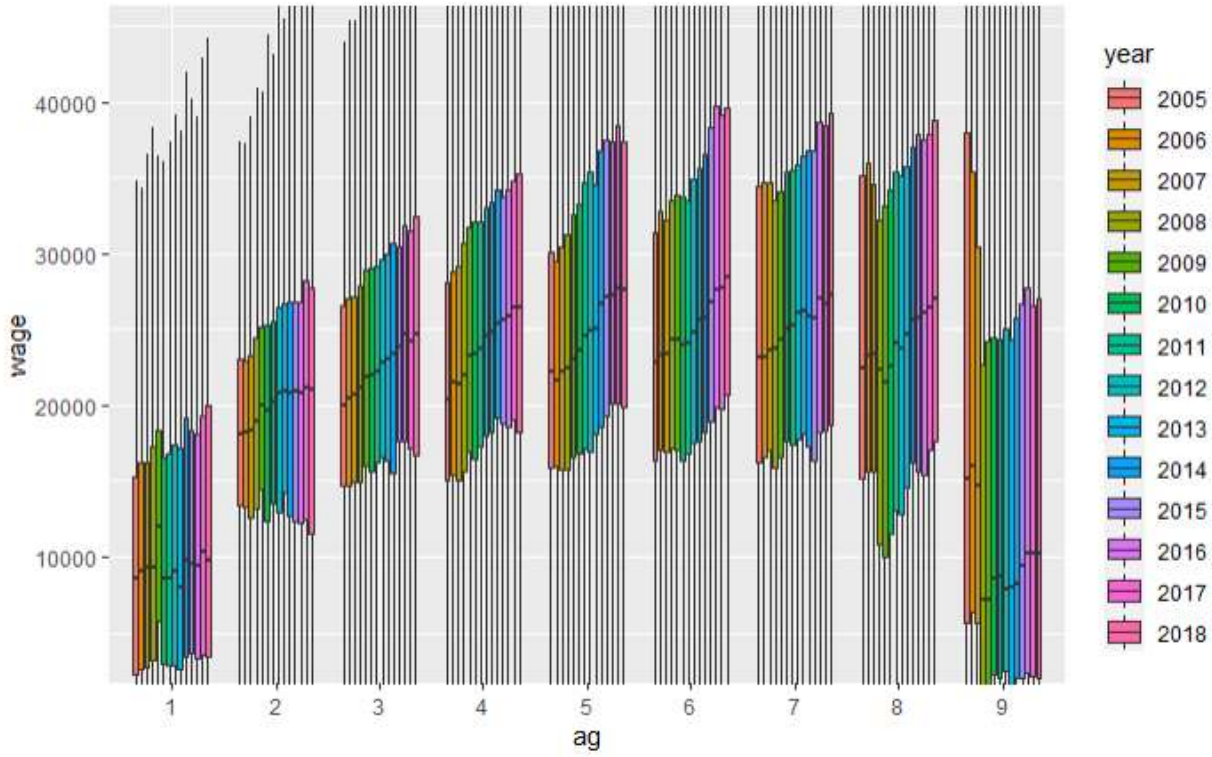
There were definitely differences between each age group by it seemed as though there were overall growth in average wages (not taking into account inflation) over time. I am a bit colorblind so I can't fully discern the different groups but there was one group that started just under 30,000 and ended around 20,000 and it seemed to be the only group with a negative trend over time. Overall, positive trends in most groups. Potential instance where using time fixed effects could be useful.

Later, the professor said to do boxplots and I did many of them (all should be available in my markdown). I did aggregate boxplots that did each year by age group as well as each age group by year. The question asked to plot wage of each group across years. To better see the data, I cropped the y-axis so that we see the data between the 10^{th} and 90^{th} deciles. If we included the outliers, you could not see any trends as the y scaling made all the boxplots look identical. I included graphs for individual groups in my markdown (both with and without outliers) as well as the aggregate graphs without outliers. The ones shown below seem the most relevant to the problem:

Hamzah Ahmed
Econ 613
Assignment 2



It seemed that, for the most part, the median wages across years tended to rise for all groups. It seems that, at each year, the groups in the middle tend to have the highest median wages and those between-year trends seem to stay (although overall levels generally rise) across years.



This data seems much more clear to see trends (which is why I included it). It seems as though each age group tends to see a general increase in median wages as time progresses. Many of them do see a dip around 2008 (most largely for the retired group) but then things start to pick back up. It seems as though the retired group was the only one that ended in a (much) worse place than they began (and perhaps the younger groups as well).

b. I did this using factor and manually with the use of as.numeric(year == X) and seemed to get the same results either way (with and without using lm function). After incorporating time fixed effects by considering $Y_{it} = \beta X_{it} + \gamma_t + e_{it}$, the new coefficient I got for $\hat{\beta}$ was 10625.074

(intercept) and 297.817 (age). Each year (base was 2005) also now have resulting coefficients and they were all positive (except for 2008) to incorporate how intercept can change with year (ie incorporating time fixed effects relative to base year). Now, the estimate for age increased in magnitude (while the intercept seemed to decrease). This could be because the previous model did not incorporate the fixed effects associated with a year (perhaps certain years where better/worse financially for most people than others). I can think of the 2008 financial crisis could have greatly affected older/retired people so if G9 was this group with the downward trend, that impact could have explained why it suddenly dropped and maybe, generally, the overall economy was doing better in the years following that (good growth after recovery). Now, after including time fixed effects, age seems to contribute more to the wage level. Results from fully manually regression with time fixed effects (intercept and then all coefficients): Intercept, Age, and dummies for 2006,…, 2018.

[,1]

[1,] 10625.07352 (intercept)

[2,] 297.81653 (age)

Dummy Variables (for Years)

[3,] 114.58388

[4,] 135.20977

[5,] -90.79323

[6,] 758.23827

[7,] 568.06612

[8,] 1106.64241

[9,] 1550.65094

[10,] 1424.06266

[11,] 2029.96350

[12,] 2305.74479

[13,] 2982.29116

[14,] 2940.21715

[15,] 3044.37546

If we think year FE should be included, then our initial estimates were probably quite biased. There could also be issues with not including a quadratic term but, for simplicity, I stuck with a linear model as the TA recommended.

Hamzah Ahmed
Econ 613
Assignment 2
Exercise 3

a. Filtered out empstat that were “Inactive” and “Retired” and those under 18.
b. Followed directions. Created XB with an intercept and then calculated CDF and then found loglikelihood.

c. I ran the optimization 500 times and stored the intercept and coefficient values using the BFGS method. After each optimization, I took note of what the log likelihood was. After the optimization was completed, I found the index of the minimum log likelihood (as BFGS looks for minimization) and found the associated optimized intercept and coefficient. Not fully converging (perhaps not identified so not always converging). Results were (generally around):

   [1] 1.058024974 0.006621276

   In general, when using probit (or logit) regression, the actual magnitudes of the coefficients don’t give us any useful information. The signs are more important. A positive coefficient means that an increase in the associated independent variable leads to an increase in the predicted probability of the variable being measured (in this case, of being employed). In this model, an increase in age leads to an increase in probability of being employed.

   However, economically, there can be a case made that there is some threshold where age may also decrease probability so I added age^2 to see what happens. Economically, I would expect the squared coefficient to be negative as there should be some threshold in which the probability that people are employed decreases as you get older. The results were:
   [1] -1.951058215 0.158064195 -0.001683646 but when I ran it a few times, sometimes the signs would flip (with BFGS) so perhaps optimization doesn’t properly converge so I did not want to comment on it as I am hesitant about the results.

d. First thing to point out is that we are modeling choice and if we based decisions on wages that people receive, we can only observe wages for those who are employed. We don’t have (or should not have) any information on wages for those who are not employed (ie wages they turned down if they were to be employed) so we can’t have data that would tell us the impact of an increase in wage would have on being employed since we only have wage data on those who are employed.

   More generally, I do not believe using wages as a determinant of labor market participation would be logical as they both, in theory, indicate the same thing. If somebody is employed, they are most likely earning a wage; if somebody is earning a wage, they are most likely employed. I think we could replace empstat with wage and check if it is above 0 and get similar results but I do not think using it as a determinant of participation is smart or would be useful given how wage data is collected and its implications on the individual it corresponds to.

Hamzah Ahmed
Econ 613
Assignment 2

Exercise 4

For the sake of simplicity, I will be using linear models from now on:

a. Combined the data and filtered out inactive and retired people and those under 18.
b. Basically did similar process (including my implementation of fixed effects using as.numeric when writing out XB for each year). Then, depending on the type of model, I wrote out the loglikelihood function. For linear, it was a product of densities (from standard normal); for probit, it was the cdf; for logit, it was $\frac{e^x}{(1+e^x)}$. I individually ran optimizations for each type using similar methods as earlier: did 200 optimizations and found the one that had the highest likelihood (as that should be what is used but may not always be achieved due to inconsistent convergences). Because of convergence issues, I did 200 iterations which was going to take a long time so I went to the grocery store while optim() did its thing. 10 iterations gave me quasi consistent results but I thought 200 would be better and hopefully have more consistence convergence. After having a good idea of what, I believe, the correct parameters are, I ran the code again and got the Hessians (I did while loops to ensure that the Hessian was non-singular before continuing) and then went about finding the standard errors by finding the Fisher Information Matrix and taking the square roots of the diagonals (which correspond to each parameter in order). If the parameters weren't similar to what my original optimization (with 200 runs) was, I reran the code for hessians because you can't trust a computer.
c. Outputs consist of optimized parameters, SE from Hessian and Significance:

	V1	linearParam	linearSE	linearSig	probitParam	probitSE	probitSig	logitParam	logitSE	logitSig
1	Intercept	0.798793359201427	0.0138042751669704	57.8656502814943	0.748620944804245	0.0226850766907143	33.0005913143189	1.12080886604055	0.0438446581764077	25.5631794763004
2	Age	0.00231700972460191	0.000244528161880015	9.475430996528	0.0123335867054422	0.000405031940114672	30.4508990129281	0.0252977065667717	0.00080921619265312	31.2619875831079
3	2006	0.00385903720025647	0.0134790150056829	0.286299644197254	0.0172089934153462	0.0227756803991257	0.755586358509264	0.0318999558216975	0.0440151992236924	0.724748641022315
4	2007	0.0143447676770697	0.0133591985274764	1.07377457169802	0.0807097733727018	0.0229338516687436	3.51924197201906	0.157473146058562	0.0447322271488098	3.52035112257431
5	2008	0.0161575455618408	0.0133837444677359	1.20725149832259	0.109889954079866	0.0231535443513943	4.74613961526148	0.212894224877227	0.0453188279559616	4.69769926715903
6	2009	0.00599661912735817	0.0133881886273709	0.447903692893798	0.026611270385853	0.0226857473148352	1.17303917814736	0.0456261062292807	0.0438479431804292	1.0405529409107
7	2010	0.00447980351045161	0.0132771300217708	0.337407519780703	0.0219427427698306	0.0224849962134685	0.975883765401167	0.037236390206836	0.0434628424429601	0.856740795443934
8	2011	0.00802982584612543	0.0131976105664361	0.608430276503742	0.0553183939930824	0.0225435149688838	2.45384954695117	0.100769635355354	0.0437656513963256	2.30248224670123
9	2012	0.000339936890505236	0.0130269314237141	0.0260949320640791	0.0103905502980317	0.0220232881831435	0.471798316928195	0.0119756518095457	0.0425049567519666	0.28174718255634
10	2013	-0.00954078750169117	0.0133119572263628	-0.716708094794412	-0.0397326779314001	0.0222649586103716	-1.78453859388229	-0.0850423304656633	0.0427181862042226	-1.99077578011159
11	2014	-0.00727226366902789	0.0132689487817455	-0.548066300401472	-0.0330533513906345	0.0222519615188955	-1.48541293146525	-0.0716890543230198	0.0427684665593527	-1.6762128757535
12	2015	-0.0123353887168351	0.0133118003001574	-0.926650673740146	-0.0531465190527419	0.0222388858166599	-2.38980133676158	-0.111141780710895	0.0426487178121723	-2.60598176011693

When taking into account year fixed effects (which simply shift the intercept based on the year of the individual at the time of survey so it doesn't affect the age coefficient), it looks as though all the coefficients for age tell the same/similar story: that an increase in age leads to an increase in the predicted probability of being employed (holding all else constant). The year fixed effects indicate the probability also increases (relative to 2005) for all years except the last 3 (2013-2015). For probit and logit, the magnitude of the coefficients do not tell us much.

The linear probability model goes a little bit more specific and says that the probability of being employed increases by roughly .23% when age increases by 1 year (holding all else constant). However, linear probability models are not great as they can predict probabilities over 1 which don't make sense. Probit and Logit should always give values between 0 and 1.The parameters 3-12 tell us if the intercept moves up or down relative to the base year which is 2005. Significance is an interesting term here. Since most of these coefficients only discuss correlated changes in probability, the actual value of the probit and logit coefficients are not that important (although their corresponding test statistics do indicate being the true parameter is non-zero with a high degree of confidence. What is more interesting (and reassuring) is that all the sign of all the coefficients seem to match: parameters 1-9 are positive and 10-12 are negative. This is good as each of these binary probability models should be highlighting the same things regarding changes in variables and how they affect the likelihood of employment (all else being constant).

It looks as though the intercept and Age coefficient are statistically significant (given asymptotic normality of MLE and OLS) but some of the year fixed effects may not be. It looks as though the interpretations of coefficients for age are consistent and robust across the three models.

Hamzah Ahmed
Econ 613
Assignment 2

Exercise 5

a. Did the Marginal Effects at the Mean where I found the average of each of the values (I included the dummy variables for all years but 2005 which is base year) and multiplied by corresponding coefficients for age given by the binary model optimization. Since I am doing MEM, I took the first derivative of the probability function that we were given for either probit (pdf of standard normal) or logit (first derivative of $\left(e^{(X\beta)}\right)/\left(1+e^{(X\beta)}\right)$: Results for MEM of age are:

[1] "Probit"
0.002179040
[1] "Logit"
0.002278034

Magnitudes, which now tell us how much we expect probability of employment to increase from a unit increase in age of these are slightly different but signs are all the same. Now, we expect a 1 unit increase in age to increase probability of being employed by .21 or .22%, respectively.

b. Basic methodology: Got random resample of data, got Y, X1, X2 and tried to find optimal parameter (using similar approach as earlier where I run optim 100 times and extract the parameters with lowest likelihood). Then found averages for variables (age and then proportions of people in each year). After bootstrapping to take resamples (with replacement), I found the MEM of the resampled data and kept it in a vector. I did bootstrapping techniques 49 times with sample sizes of the number of entries in the dataset. I then took column variances to get variances of each MEM (I included the math for fixed effects but they are not relevant) and then took square roots to get SE. Due to convergence issues and realistic computational and time limits, there can be instances of some of my coefficients still being incorrect and that is simply due to my optim functions finding local (as opposed to global) extremas (you can check logitParams and probitParams to see which coefficients were used for each MEM). There could be, and very likely is, variations in parameters due to the resampled data used and issues with optim() not finding the global extrema (in 100 tries) as well. That being said, these MEM were still relatively consistent and so the SE's were not that large. Results below:

[1] "Probit BS SE"
0.0002820159
[1] "Logit BS SE"
0.00003301368