Part I [20 pts]

1. [5 pts] Between 1995 and 2005, the percentage of Americans without health insurance rose from 20 percent to 49 percent. Explain the change between 1995 and 2005 in percent and in percentage point.

Percentage point change= 49-20= 29 (Do Not put %)

% change=100 * (49-20)/ 20 = 145%

2. [5 pts] Interpret the coefficient of log (income)

$$\widehat{\text{test score}} = 557.8 + 16.42 \log\,(Income)$$

When income goes up by 1 percent, we expect, on average, test scores to increase by 0.1642 points.

3. [5 pts] Interpret the coefficient of log(income)

$$\log\,(\widehat{\text{test score}}) = 6.336 + 0.0124 \log\,(Income)$$

When income goes up by 1 %, test scores go up by 0.0124%.

4. [5 pts] When would model in part 3 (log-log model) be more useful than model 2 (linear-log model)?

When we want to understand the effect of percentage changes in income on percentage changes in test scores.

Part II

1. Suppose you are trying to estimate the below model:

$$Earnings_i = \beta_0 + \beta_1 male_i + \beta_2 female_i + u_i$$

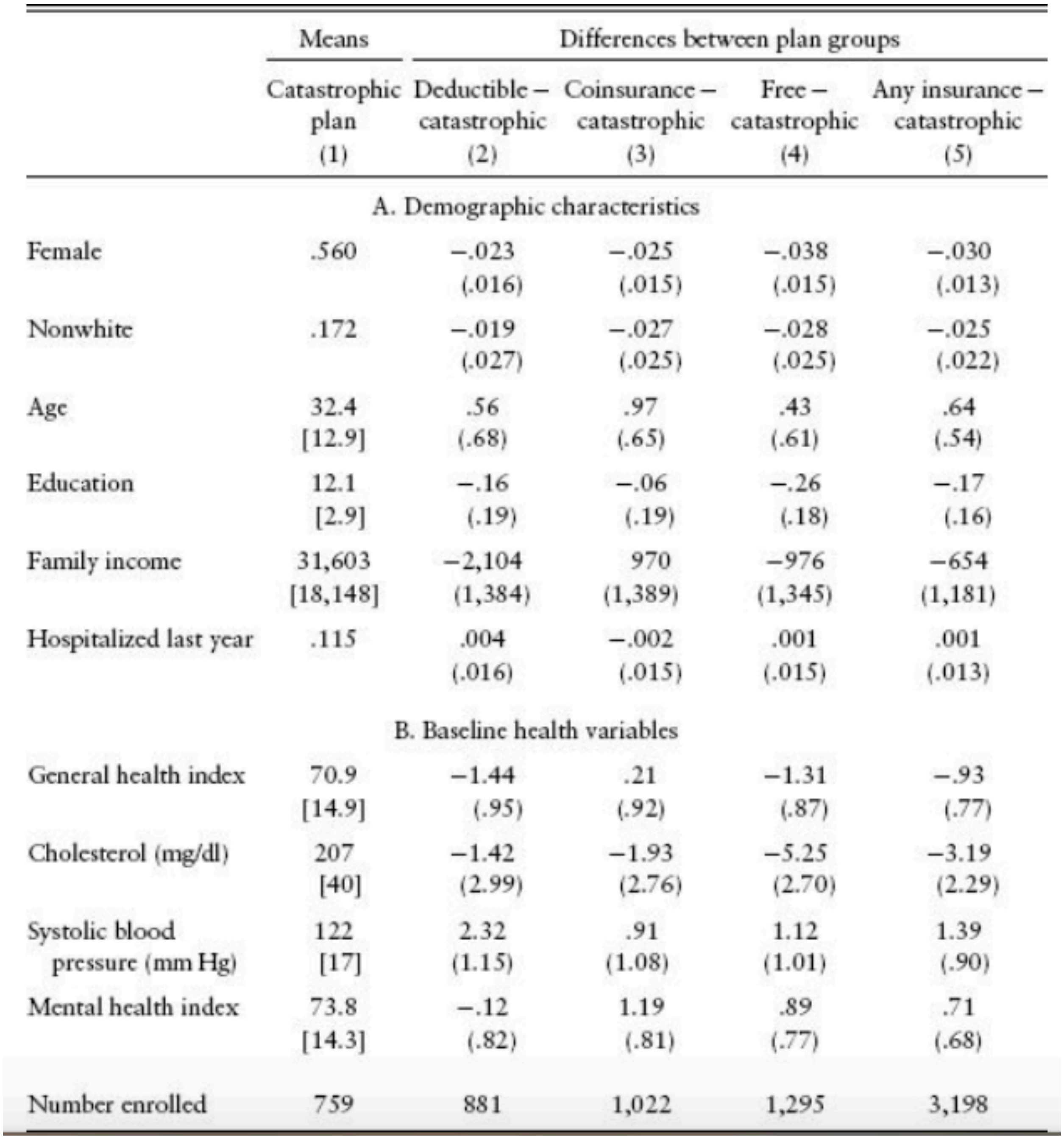
a) Is there any problem with this regression? If so, explain the problem.

Dummy variable trap/ Multicollinearity. Not able to estimate β_0 since it is the intercept coefficient which means both male and female variables need to be equal to zero for us to be able to estimate it. However, both male and female dummy cannot be zero at the same time. Hence, we have too many dummy variables in this regression implying that it cannot be estimated.

b) Also, explain how you can fix the problem.

Drop constant term or drop male or female dummy variable. The regression model can be run now.

Part III



	Means	Differences between plan groups			
	Catastrophic plan	Deductible – catastrophic	Coinsurance – catastrophic	Free – catastrophic	Any insurance – catastrophic
	(1)	(2)	(3)	(4)	(5)
		A. Demographic characteristics			
Female	.560	−.023	−.025	−.038	−.030
		(.016)	(.015)	(.015)	(.013)
Nonwhite	.172	−.019	−.027	−.028	−.025
		(.027)	(.025)	(.025)	(.022)
Age	32.4	.56	.97	.43	.64
	[12.9]	(.68)	(.65)	(.61)	(.54)
Education	12.1	−.16	−.06	−.26	−.17
	[2.9]	(.19)	(.19)	(.18)	(.16)
Family income	31,603	−2,104	970	−976	−654
	[18,148]	(1,384)	(1,389)	(1,345)	(1,181)
Hospitalized last year	.115	.004	−.002	.001	.001
		(.016)	(.015)	(.015)	(.013)
		B. Baseline health variables			
General health index	70.9	−1.44	.21	−1.31	−.93
	[14.9]	(.95)	(.92)	(.87)	(.77)
Cholesterol (mg/dl)	207	−1.42	−1.93	−5.25	−3.19
	[40]	(2.99)	(2.76)	(2.70)	(2.29)
Systolic blood pressure (mm Hg)	122	2.32	.91	1.12	1.39
	[17]	(1.15)	(1.08)	(1.01)	(.90)
Mental health index	73.8	−.12	1.19	.89	.71
	[14.3]	(.82)	(.81)	(.77)	(.68)
Number enrolled	759	881	1,022	1,295	3,198

For the following questions, focusing on the three outcome variables (Cholesterol, Blood pressure, and Mental health index).

1. [15 pts] Compare the three treatment effect estimates in Column 5 (any treatment – catastrophic) for Cholesterol (mg/dl) , Systolic blood pressure, and Mental health index (higher value means better mental health) with the control (Column 1) means. Explain each estimate **separately and if possible say something about the size of the effect**.

Cholesterol: The cholesterol level for the treatment group, on average, is 3.19 mg/dl smaller than that of the control (catastrophic) group. Given that the control group mean is 207 mg/dl this represents a difference of (100 * 3.19/207)= 1.54% which seems to be a small difference between the control and treatment group.

Systolic Blood pressure: The systolic blood pressure for the treatment group is on average 1.39 mm Hg higher than that of the control (catastrophic) group. Given that the control group mean is 122 mm Hg this represents a difference of (1.39/122) * 100= 1.14% which seems to be a small difference between the control and treatment group.

Mental Health Index: The mental health index for treatment group is 0.71 points higher than that of the control group. This represents the size of effect of (0.71/73.8)* 100= 0.96% difference between control group and any treatment group. This difference also seems small.

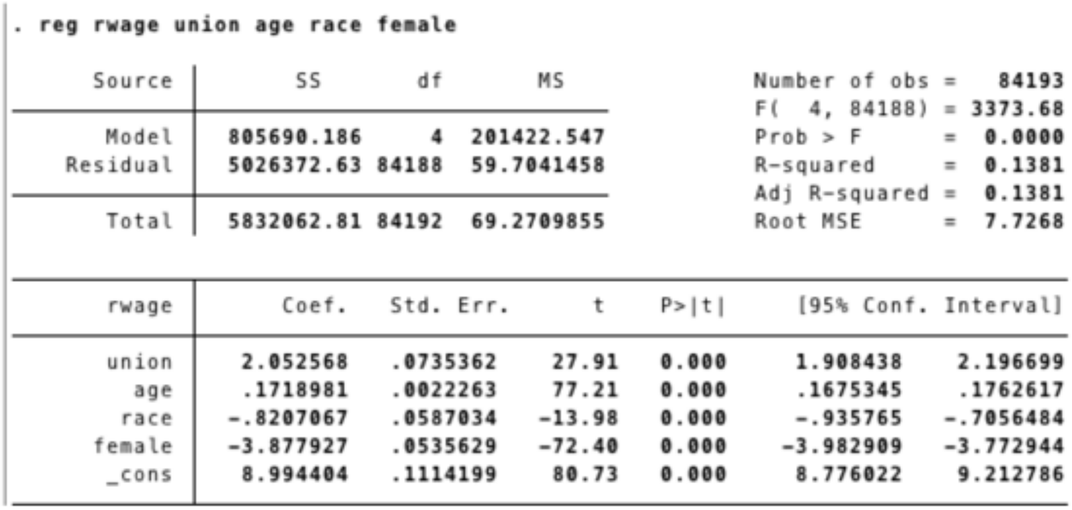
2. [15 pts] Perform statistical inference on the difference between the treatment (Cholesterol, Blood pressure, and Mental health index) and control to see if **each of these three** these estimates are statistically significant at α=0.05. (Hint: use rule of thumb: compare t-stat with t-critical value of 2).

Cholesterol: t-stat= coefficient/ standard error = =3.19/2.29= -1.39
Absolute value of t-stat is smaller than t- critical value of 2 => We fail to reject the null which states that the two means are not statistically different.

Systolic Blood pressure:1.39/ 0.90= 1.54 <2 => Fail to reject null => Not statistically significant

Mental Health Index= 0.71/0.68 = 1.04 => Fail to reject null => Not statistically significant

Part IV



```
. reg rwage union age race female

      Source |       SS       df       MS              Number of obs =   84193
-------------+------------------------------           F(  4, 84188) = 3373.68
       Model |  805690.186     4  201422.547           Prob > F      =  0.0000
    Residual |  5026372.63 84188  59.7041458           R-squared     =  0.1381
-------------+------------------------------           Adj R-squared =  0.1381
       Total |  5832062.81 84192  69.2709855           Root MSE      =  7.7268

------------------------------------------------------------------------------
       rwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   2.052568   .0735362    27.91   0.000     1.908438    2.196699
         age |   .1718981   .0022263    77.21   0.000     .1675345    .1762617
        race |  -.8207067   .0587034   -13.98   0.000     -.935765   -.7056484
      female |  -3.877927   .0535629   -72.40   0.000    -3.982909   -3.772944
       _cons |   8.994404   .1114199    80.73   0.000     8.776022    9.212786
------------------------------------------------------------------------------
```

1. What is the explained sum of square (ESS)? What is the residual sum of squares (RSS)? What do they mean in words?

ESS= Model sum of squares. It measures the part of variation in the wage rate that can be explained by the regression model that we are running.

RSS= Part of variation in the y-variable (wage rate in this context) which cannot be explained by the model.

2. What is the R-square? What does it mean in this example?

R-square is the part of variation in y-variable that can be explained by the regression model.

R-square= ESS/ TSS= ESS/ (ESS + RSS)= 805690.186/ 5832062.81 = 0.1381

⇨ **This model explains around 13.81% variation in wage rates.**

3. What is the F-stat shown on the output? Specify the null hypothesis for this F-stat. Based on its p-value (Prob>F), do you reject this null hypothesis at a=0.05?

Ho: β_{union}= β_{age}= β_{race}= β_{female}=0

Ha: atleast one $\beta_i \neq 0$ where i= union, age, race, female.

Part V (See problem Set 3 Solution)

1. Suppose you're interested in estimating the effect of family income on health status. Your original model is given as follows: health index $= \beta_0 + \beta_1 \text{family income} + u$

   (a) How would you change the model if you think that the effect of family income on health can be different for immigrant and non-immigrant?

(b) In the modified model in (a), how would you test whether there is a differential effect of family income on health status depending on immigrant status? State the null hypothesis and alternative hypothesis.

(c) How would you change the model if you think that the effect of family income on health can be different for people with different ages?

(d) In the modified model in (c), how would you test whether there is a differential effect of family income on health status depending on ages? State the null hypothesis and alternative hypothesis.

Part VI: Short Questions 4/5

Potential topics

1. F-test hypothesis **(joint test of all coefficients on regressors =0)**
2. Internal and External Validity (**Internal Validity is about whether we have reasons to believe that the coefficient might be biased because of any source of bias whereas external validity is about whether the results of the study will hold in a different context. Example we do a study of class size on test scores in IL. Internal Validity will judge if we have omitted variable bias, wrong functional form, measurement error, reverse causality/ simultaneous causality, sample selection bias. External Validity is about whether results from IL will generalize to other states in the US and possibly to other countries.**
3. Critique of Multiple regression (**5 types of biases in multiple regression)**
4. Linear Probability model (dependent binary variable) and its critique (**easy to estimate but can generate probabilities below zero or above 1!)**
5. R-square and adjusted R-square **(Adjusted r-square penalizes you for adding more variables to a regression. When adding a variable to a regression, R-square goes up for sure whereas adjusted r-square need not go up. It values how much new information are the new regressors adding to the regression once we account for the fact that we are adding more variables).**
6. Why naïve comparison is typically misleading and when it is not? **(A naive comparison is typically misleading since for example when we compare people with and without health insurance, they are likely to be different sets of people so the comparison is not apple to apple. However, if we were to randomly assign people to control and treatment groups as in the case of a Randomized control trial, a simple mean comparison tends to yield results that are not misleading).**

7. Balance test in RCT, results table in RCT **(A balance test allows us to test whether randomization was done successfully or not. It basically compares the demographics of people in control and treatment groups *before* the experiment has started. If the group of people in treatment and control groups are of similar demographics (e.g. similar age, similar fraction of females/ race) then we are confident that the randomization was done well. The results table in the RCT**

looks at the mean comparison of outcomes for control and treatment group once the experiment has been done).

8. What is the attrition problem in RCT. **(The attrition problem can arise in a RCT if for some reason more people drop out in either the control or treatment group and this dropping out is not random but is in fact correlated with the outcome variable of interest. See slides for examples).**
9. Dummy variable trap and how to fix it. **(Drop either the constant term, or one of the categories of variables).**
10. When is log-log model more useful? **When we want to understand the effect of percentage changes in income on percentage changes in test scores.**