Problem 1 (40 Points)

Consider the following simple specification that tests for regional differences in hours worked:

\[ \text{hours} = \beta_0 + \beta_{urban} + u \]

*hours* is average hours worked per week, and *urban* is a dummy variable that takes on a value of 1 if the respondent lives in a metropolitan area, and 0 otherwise. The results from estimating this equation are below:

| Estimate  | Std. Error | t value | Pr(>|t|)         |
|-----------|------------|---------|-----------------|
| (Intercept) | 43.7386 | 0.4448  | XXXXXXXXXXXXXXXX |
| urban      | 0.2658   | 0.5251  | XXXXXXXXXXXXXXXX |

---

Multiple R-squared: 0.0002747, Adjusted R-squared: -0.0007969  
F-statistic: 0.2563 on 1 and 933 DF, SSR=48731.95

a.) Please construct and interpret a 95% confidence interval for the intercept. (10 Points)

\[ t_{crit} = 1.96 \]

\[ 43.7 - 0.44 \cdot 1.96 < \beta_0 < 43.7 + 0.44 \cdot 1.96 \]

\[ 42.84 < \beta_0 < 44.56 \]

With 95% confidence, a person that lives in a rural (non-urban) location works between 42.84 and 44.56 hours per week.
b.) I claim that urban residents work a number of hours that is significantly different than rural residents. What is the probability that I'm wrong? (10 Points)

\[
P_r(\frac{z}{0.5251} > \frac{0.2658}{0.5251}) = 2 \left( 1 - P_r(z < \frac{0.2658}{0.5251}) \right)
\]

\[
+9
\]

\[
= 2 \left( 1 - P_r(z < 0.506) \right)
\]

\[
= 2 \left( 1 - 0.6956 \right)
\]

\[
= 0.61
\]

-6

reasonable

work is ok


c.) Suppose that instead of the regression in 'a', I run the following regression:

\[\text{hours} = \beta_0 + \beta_{\text{urban}} + \beta_{\text{educ}} + \epsilon\]

where \text{educ} is the years of education of the respondent. The results from estimating this equation are below:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept)  | 39.8146    | 1.4889  | XXXXXXXXXXXX |
| urban       | 0.1613     | 0.5246  | XXXXXXXXXXXX |
| educ        | 0.2969     | 0.1076  | XXXXXXXXXXXX |

---
Multiple R-squared: 0.008383, Adjusted R-squared: 0.006255
F-statistic: 3.94 on 2 and 932 DF, SSR=48336.7

In comparing the regression in ‘a’ and the regression in ‘c’, what is the correlation between \text{educ} and \text{urban}? Why? (10 Points)

The correlation is positive. In ‘a’, \text{educ} is an omitted variable. Given the positive relationship between \text{educ} and \text{hours}, and given \text{urban} goes down, there must be a positive correlation between \text{educ} and \text{urban}.
d.) Does the model in ‘c’ tell us anything about hours worked? If a hypothesis test is warranted, test this hypothesis at the 95% level, stating your null and alternative hypotheses. If not, provide other evidence for your answer. (10 Points)

\[ F_{\text{stat}} = 3.911 \]

\[ q = 2 \Rightarrow F_{\text{crit}} = 3.00 \]

\[ F_{\text{stat}} > F_{\text{crit}} \Rightarrow \text{Reject null in favor of alternative.} \]

\[ H_0: \beta_1 = \beta_2 = 0 \]

\[ H_a: \text{not true} \]

\[ + \]

\[ \beta \]

e.) Suppose that I modify the regression in ‘c’ to include age and age2, which are the age and age squared of the respondent.

\[ \text{hours} = \beta_0 + \beta_{\text{urban}} + \beta_{\text{educ}} + \beta_{\text{age}} + \beta_{\text{age}^2} + u \]

The results from this regression are below:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 42.078061  | 31.524587 | XXXXXXXXXX |
| urban      | 0.166803   | 0.525423  | XXXXXXXXXX |
| educ       | 0.299204   | 0.108045  | XXXXXXXXXX |
| age        | -0.200219  | 1.914023  | XXXXXXXXXX |
| I(age^2)   | 0.003918   | 0.028749  | XXXXXXXXXX |

---

Multiple R-squared: 0.009078, Adjusted R-squared: 0.004816
F-statistic: 2.13 on 4 and 930 DF, SSR=48302.81

At what age is average hours worked minimized? Show your work!! (10 Points)

\[ \frac{\text{d}(\hat{\text{hours}})}{\text{d} \text{age}} = \hat{\beta}_3 + 2\hat{\beta}_4 \text{age} = 0 \]

\[ \text{age} = -\frac{\hat{\beta}_3}{2\hat{\beta}_4} \]

\[ = \frac{0.200}{2 \cdot 0.0039} \]

\[ = 25.64 \]
f.) Is the model in ‘e’ preferred to the model in ‘c’? If a hypothesis test is warranted, test this hypothesis at the 95% level, stating your null and alternative hypotheses. If not, provide other evidence for your answer. (10 Points)

\[ F_{\text{stat}} = \frac{48,336.7 - 48,302.8}{48,302.8} = 0.326 \]

\[ F_{\text{crit}} = 3.60 \quad F_{\text{stat}} < F_{\text{crit}} \]

Fail to reject \( H_0 \).
Problem 2 (40 Points)

a.) For this problem, we wish to study the impact of health insurance on the smoking behavior of pregnant mothers. While difficult to assess, we will leverage a family’s eligibility for prenatal care via Medicaid to determine the effects of health insurance on behavior. To do so, we run the following regression:

\[ smoke = \beta_0 + \beta_1 \text{faminc} + \beta_2 \text{medicaid} + \beta_3 \text{faminc} \cdot \text{medicaid} + u \]

Here, \( smoke \) takes on a value of 1 if a mother smoked during pregnancy, and zero otherwise. Further, \( \text{faminc} \) is yearly family income (in thousands) and \( \text{medicaid} \) is a dummy variable taking a value of 1 if \( \text{faminc} \) is below 22 (which is $22,000) and zero otherwise. What kind of regression technique is this? (10 Points)

b.) The results from estimating the regression in ‘a’ are below:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 0.2053321  | 0.0329005 | XXXXXXXXXX |
| faminc    | -0.0024584 | 0.0007648 | XXXXXXXXXX |
| medicaid  | 0.0810480  | 0.0601873 | XXXXXXXXXX |
| I(\text{medicaid} \cdot \text{faminc}) | -0.0037862 | 0.0036022 | XXXXXXXXXX |

Multiple R-squared: 0.02766, Adjusted R-squared: 0.0252
F-statistic: 11.26 on 3 and 1187 DF, SSR= 135.3847

Please use a t-test to test whether Medicaid eligibility (at the eligibility threshold) affects smoking behavior. Please state your null and alternative hypotheses, and test the null against the alternative at the 99% level. (10 Points)

\[ H_0: \beta_2 = 0 \]
\[ H_A: \beta_2 \neq 0 \]

\[ t_{stat} = \frac{0.081 - 0}{0.0601} = 1.35 \]

\[ t_{crit} = 2.575 \]

\[ |t_{stat}| < t_{crit} \rightarrow \text{Fail to reject } H_0 \]

Eligibility does not have an effect on smoking behavior at the threshold.
c.) Does the relationship between family income and maternal smoking behavior depend on whether the family is eligible for Medicaid? Test this hypothesis at the 98% level using a two-sided test. State your null and alternative, and show your work! (10 Points)

\[
\begin{align*}
H_0: \beta_2 &= 0 \\
H_1: \beta_2 &\neq 0
\end{align*}
\]

\[
E_{stat} = \frac{-0.00378}{0.0036} = -1.058
\]

\[
E_{crit} = 2.325
\]

\[
|E_{stat}| < E_{crit} \Rightarrow \text{Fail to reject null}
\]

There is no significant relationship between income and smoking that is conditional on medicaid.

d.) Suppose that instead of the above model, we estimate the following model:

\[
\text{smoke} = \beta_0 + \beta_{\text{mothers}} + \beta_{\text{medicaid}} + u
\]

where \text{mothers} is the mother’s education level in years. The results are below:

\[
\begin{array}{cccc}
\text{Estimate} & \text{Std. Error} & \text{t value} & \text{Pr(>|t|)} \\
\hline
\text{(Intercept)} & 0.520542 & 0.059349 & XXXXXXXXXXXXXXX \\
\text{mothers} & -0.030276 & 0.004257 & XXXXXXXXXXXXXXX \\
\text{medicaid} & 0.040296 & 0.022489 & XXXXXXXXXXXXXXX \\
\hline
\end{array}
\]

Multiple R-squared: 0.05677, Adjusted R-squared: 0.05519
F-statistic: 35.75 on 2 and 1188 DF, SSR=131.3309

Please interpret the coefficient on \text{medicaid}, and test whether this coefficient is significantly different from zero. Please state your null and alternative hypotheses, and test the null against the alternative at the 90% level. (10 Points)

\[
\begin{align*}
H_0: \beta_2 &= 0 \\
H_1: \beta_2 &\neq 0
\end{align*}
\]

\[
E_{stat} = \frac{0.040296 - 0}{0.022489} = 1.79 \\
E_{crit} = 1.645
\]

\[
|E_{stat}| > E_{crit} \Rightarrow \text{Reject null and favor alternative}
\]

Mothers eligible for medicaid have a 0.0102 larger probability of smoking.
e.) Which regression is preferred, the regression in ‘2b’ or the regression in ‘2d’? If a hypothesis test is warranted, test this hypothesis at the 95% level, stating your null and alternative hypotheses. If not, provide other evidence for your answer. (10 Points)

\[ \text{Non-Nested} \Rightarrow \text{Adjusted } R^2 \]

\[ \text{Adj } R^2_b = 0.0252 + 4 \]

\[ \text{Adj } R^2_d = 0.0552 + 1 \]

\[ \Rightarrow \text{Model in "d" is preferred.} \]

\[ +3 \text{ (or F-test)} \]
f.) Using the previous regression equation in ‘d’, we wish to predict the probability of smoking for a mother with 20 years of education that is eligible for Medicaid. Please derive a regression equation that allows us to generate this prediction with standard error, and write the R commands that would estimate this particular equation. Show your work!! (10 Points)

\[ \Theta = b_0 + b_1 \cdot 20 + b_2 \cdot 1 \]

\[ \Rightarrow b_0 = \Theta - b_1 \cdot 20 - b_2 \cdot 1 \]

\[ = \frac{1}{y} \text{smoke} = (\Theta - b_1 \cdot 20 - b_2 \cdot 1) + b_1 \text{motheredu} + b_2 \text{medicaid} + u \]

\[ \text{smoke} = \Theta + b_1 \text{(motheredu - 20)} + b_2 \text{(medicaid - 1)} + u \]

\[ x_{\text{M20}} = x_{\text{motheredu}} - 20 \]

\[ x_{\text{Med1}} = x_{\text{medicaid}} - 1 \]

\[ \text{lm(smoke} = x_{\text{M20}} + x_{\text{Med1}}, x) \]

or

\[ \text{lm(smoke} \sim (x_{\text{motheredu}} - 20) + x_{\text{medicaid}} - 1), x) \]

Have a nice holiday!!!