[redacted]

[redacted]

Problem 1 from Chapter 7

i. The coefficient for males is 87.75. This indicates that men are estimated to sleep around 1.5 hours more per week than women. Calculating this difference gives 87.75/34.33 = 2.56. This value is near the 1% significance level for a two-sided test which is approximately 2.58. This provides robust evidence of a notable difference in sleep patterns between genders.

ii. For the variable totwrk, we can compute the t-statistic as -0.163/0.018, which equals -9.06. This value highlights a significant statistical difference. In practical terms, every additional hour of work, which is 60 minutes, leads to a decrease in sleep by about 9.8 minutes.

iii. To check the impact of age on sleep while holding other factors constant, a partial F-test is necessary. First, we must calculate the R^2 value from a model without the age and age^2 variables. We test the assumption: $\beta age = \beta age^2 = 0$. If this holds, we use the model without the age variables. If not, we include age and age^2 in the full regression model.

Problem 3 from Chapter 7

i. When evaluating the variable 'hsize^2', its t-statistic is determined as -2.19/0.53, = -4.13. This provides compelling evidence for the inclusion of 'hsize^2' in the model.
To ascertain the most appropriate high school size, we differentiate 'sat' WRT 'hsize,' keeping it constant, and then set it to zero:
19.3 + 2 * 2.19 * 'hsize' = 0
This results in 'hsize' = -4.406. Keeping in mind that 'hsize' is scaled in hundreds, this implies the best graduating class size is approximately 441.

ii. The disparity in SAT scores between non-black females and non-black males is inferred from the coefficient of 'female' (when 'black' = 0). Non-black females tend to score about 45.09 points lower than non-black males. The t-statistic, calculated as -45.09/4.29 = -10.51. Given the expansive sample size, this is statistically very significant.

iii. The coefficient for 'black' suggests that a black student is likely to score around 169.81 points < non-black peer. With a t-statistic > 13, we can decisively reject the null hypothesis of no difference in their scores, confirming a significant difference.

iv. Inserting values 'black' = 1 and 'female' = 1 for black females and 'black' = 0 and 'female' = 1 for non-black females, the computed difference becomes -169.81 + 62.31 = -107.50. This interpretation relies on both coefficients and isn't a straightforward t-test. To perform a comprehensive analysis, one must frame an F-test using linear constraints to ascertain the significance of the combined effect of the 'race' and 'gender' dummy variables.

An alternative methodology to evaluate this difference is by setting up linear constraints:
H0: β'black' + β'female-black' = 0 versus
Ha: β'black' + β'female-black' ≠ 0.

Problem C9 from Chapter 7

i. Out of the total sample, 39% of the families are eligible for participation in a 401(k) plan.

```
# (i) Fraction of families eligible for 401k
eligible_families = data['e401k'].sum()
total_families = len(data)
fraction_eligible = eligible_families / total_families
print(f"(i) Fraction of families eligible for 401K: {fraction_eligible:.2f}")
```

ii. In the Linear Probability Model (LPM) results:
   - The effect of income (inc) on eligibility is positive, suggesting that as income increases, the probability of being eligible for a 401(k) also increases.

   - Age (age) has a positive effect on eligibility, but the quadratic term for age (agesq) is negative. This suggests that as age increases, the probability of being eligible initially increases but at a decreasing rate.

   - The coefficient of the male variable is not statistically significant (given its p-value is 0.770), suggesting that gender might not have a significant effect on eligibility when considering other variables.

```
# (ii) Linear Probability Model (LPM)
data['incsq'] = data['inc'] ** 2
data['agesq'] = data['age'] ** 2

X = data[['inc', 'age', 'incsq', 'agesq', 'male']]
X = sm.add_constant(X)
y = data['e401k']

model = sm.OLS(y, X).fit()
print("\n(ii) LPM Results:")
print(model.summary())
```

iii. As inferred from the results:
   - 401(k) eligibility appears to be dependent on income and age given their statistically significant coefficients.
   - Gender doesn't appear to have a significant effect on 401(k) eligibility.

iv. The range of fitted values from the LPM model:
   - Minimum Fitted Value: 0.0299
   - Maximum Fitted Value: 0.6972
   - From the estimated values, none of the 9275 predicted values fall outside the range [0, 1]. Interestingly, no anomalies were observed with the LPM in this context.

```
# (iv) Range of fitted values
fitted_vals = model.predict(X)
print(f"\n(iv) Min Fitted Value: {fitted_vals.min()}")
print(f"Max Fitted Value: {fitted_vals.max()}")
```

v. Using the defined criteria, 2460 families out of 9,275 are predicted to be eligible for a 401(k) plan.

```
# (v) Predicted eligible families
predicted_eligible = np.where(fitted_vals >= 0.5, 1, 0)
print(f"\n(v) Predicted Eligible Families: {predicted_eligible.sum()}")
```

vi. Among the families:
   - 81.71% of the 5,638 families not eligible for a 401(k) are predicted not to have a 401(k).
   - 39.29% of the 3,637 families eligible for a 401(k) plan are predicted to have one.

```
# (vi) Percentages for not eligible and eligible families
actual_not_eligible = data[data['e401k'] == 0]
predicted_not_eligible_from_actual = predicted_eligible[actual_not_eligible.index]
percentage_not_eligible = 100 * (1 - predicted_not_eligible_from_actual.mean())
print(f"\n(vi) Percentage of the 5,638 families predicted not to have a 401(k): {percentage_not_eligi

actual_eligible = data[data['e401k'] == 1]
predicted_eligible_from_actual = predicted_eligible[actual_eligible.index]
percentage_eligible = 100 * predicted_eligible_from_actual.mean()
print(f"Percentage of the 3,637 families predicted to have a 401(k): {percentage_eligible:.2f}%")
```

vii. The accuracy rate of the predictions is approximately 64.9%. This is derived from a weighted average of the predictions from part (vi). The model performs commendably when predicting ineligibility. However, it struggles to predict eligibility, being accurate less than 40% of the time.

OUTPUT for C9:

```
hw3')
(i) Fraction of families eligible for 401K: 0.39

(ii) LPM Results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                  e401k   R-squared:                       0.094
Model:                            OLS   Adj. R-squared:                  0.094
Method:                 Least Squares   F-statistic:                     193.0
Date:                Thu, 02 Nov 2023   Prob (F-statistic):          3.41e-196
Time:                        16:46:21   Log-Likelihood:                -6051.5
No. Observations:                9275   AIC:                         1.211e+04
Df Residuals:                    9269   BIC:                         1.216e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.5063      0.081     -6.243      0.000      -0.665      -0.347
inc            0.0124      0.001     20.993      0.000       0.011       0.014
age            0.0265      0.004      6.758      0.000       0.019       0.034
incsq      -6.165e-05   4.73e-06    -13.028      0.000   -7.09e-05   -5.24e-05
agesq         -0.0003    4.5e-05     -6.782      0.000      -0.000      -0.000
male          -0.0035      0.012     -0.292      0.770      -0.027       0.020
==============================================================================
==============================================================================
Omnibus:                    65188.981   Durbin-Watson:                   1.970
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1031.991
Skew:                           0.369   Prob(JB):                    8.05e-225
Kurtosis:                       1.542   Cond. No.                     6.51e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 6.51e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

(iv) Min Fitted Value: 0.02991717362643631
Max Fitted Value: 0.6971898950315519

(v) Predicted Eligible Families: 2460

(vi) Percentage of the 5,638 families predicted not to have a 401(k): 81.71%
Percentage of the 3,637 families predicted to have a 401(k): 39.29%
```

Problem C15 from Chapter 7

i. The min and max values of children observed are 0 and 13, respectively. The mean number of children is approximately 2.27. Clearly, a woman can't have an average of 2.27 children.

```
# (i) Compute smallest, largest, and average values of children
min_children = data['children'].min()
max_children = data['children'].max()
avg_children = data['children'].mean()
```

ii. Among the 4,358 women where data on electricity is available, 611, or approximately 14.02%, have access to electricity in their households.

```
# (ii) Compute percentage of women with electricity in the home
percentage_electric = (data['electric'].sum() / data['electric'].count()) * 100
```

iii. When excluding data for the 3 women without electricity details, the mean number of children for women lacking electricity is approximately 2.33. For those with electricity, the mean is around 1.90. By using a simple regression on children against electric, we see that the difference in the average number of children between these groups is around -429. Factoring in robust standard errors for heteroskedasticity, we obtain a t-statistic of -5.237, indicating this result is statistically significant.

```
# (iii) Compute average of children for those without/with electricity
avg_children_no_electric = data.loc[data['electric'] == 0, 'children'].mean()
avg_children_with_electric = data.loc[data['electric'] == 1, 'children'].mean()
```

iv. Establishing a direct causation between the presence of electricity and the number of children is challenging. External factors, such as a woman's income or education level, can influence these numbers.

v. When performing a regression of children on variables like electric, age, age^2, urban, spirit, protest, and catholic, the coefficient of electric is around -.306 (standard error = .064). Even after considering the added variables, the effect of having electricity on the average number of children remains significant, albeit less than in the previous model. The t-statistic becomes -4.761, reinforcing its statistical significance.

```python
# Regression: children on electric
X = sm.add_constant(data['electric'].dropna())
y = data.loc[data['electric'].notna(), 'children']
model = sm.OLS(y, X).fit(cov_type='HC1') # Using heteroskedasticity robust standard errors

# (v) Regression with multiple variables
X_mult = sm.add_constant(data[['electric', 'age', 'age2', 'urban', 'spirit', 'protest', 'catholic']].dropna())
model_mult = sm.OLS(y, X_mult).fit(cov_type='HC1') # Using heteroskedasticity robust standard errors
```

vi. Introducing an interaction term between electric and educ results in a coefficient of approximately -.022, with its associated t-statistic being -1.174 (with an adjusted p-value of .24). This result indicates that the effect of electric diminishes as education increases. Still, this interaction isn't statistically significant at conventional levels when focusing on a subgroup where educ = 0.

```python
# (vi) Adding interaction term between electric and educ
data['interaction'] = data['electric'] * data['educ']
X_interact = sm.add_constant(data[['electric', 'educ', 'interaction', 'age', 'age2', 'urban', 'spirit', 'protest', 'catholic']].dropna())
model_interact = sm.OLS(y, X_interact).fit(cov_type='HC1') # Using heteroskedasticity robust standard errors
```

Output for C15:

```
(i) Min children: 0, Max children: 13, Avg children: 2.27
(ii) Percentage of women with electricity: 14.02%
(iii) Average children with no electricity: 2.33, with electricity: 1.90
(iii) Regression results:
                            OLS Regression Results
==============================================================================
Dep. Variable:               children   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     27.49
Date:                Thu, 02 Nov 2023   Prob (F-statistic):           1.65e-07
Time:                        18:29:03   Log-Likelihood:                -9652.7
No. Observations:                4358   AIC:                         1.931e+04
Df Residuals:                    4356   BIC:                         1.932e+04
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.3277      0.037     62.558      0.000       2.255       2.401
electric      -0.4292      0.082     -5.243      0.000      -0.590      -0.269
==============================================================================
Omnibus:                      614.395   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              902.845
Skew:                           1.055   Prob(JB):                    8.91e-197
Kurtosis:                       3.719   Cond. No.                         2.94
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
(v) Regression with multiple variables results:
                            OLS Regression Results
==============================================================================
Dep. Variable:               children   R-squared:                       0.560
Model:                            OLS   Adj. R-squared:                  0.560
Method:                 Least Squares   F-statistic:                     797.1
Date:                Thu, 02 Nov 2023   Prob (F-statistic):               0.00
Time:                        18:29:03   Log-Likelihood:                -7871.6
No. Observations:                4358   AIC:                         1.576e+04
Df Residuals:                    4350   BIC:                         1.581e+04
Df Model:                           7
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.9100      0.250    -19.617      0.000      -5.401      -4.419
electric      -0.5316      0.065     -8.193      0.000      -0.659      -0.404
age            0.3514      0.020     17.966      0.000       0.313       0.390
age2          -0.0027      0.000     -7.673      0.000      -0.003      -0.002
urban         -0.2708      0.046     -5.857      0.000      -0.361      -0.180
spirit         0.1008      0.057      1.763      0.078      -0.011       0.213
protest       -0.0675      0.066     -1.016      0.310      -0.198       0.063
catholic      -0.0453      0.080     -0.566      0.572      -0.202       0.112
==============================================================================
Omnibus:                      197.573   Durbin-Watson:                   1.869
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              621.997
Skew:                           0.129   Prob(JB):                    8.61e-136
Kurtosis:                       4.833   Cond. No.                     1.07e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 1.07e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
(vi) Regression with interaction results:
                            OLS Regression Results
==============================================================================
Dep. Variable:               children   R-squared:                       0.574
Model:                            OLS   Adj. R-squared:                  0.573
Method:                 Least Squares   F-statistic:                     648.1
Date:                Thu, 02 Nov 2023   Prob (F-statistic):               0.00
Time:                        18:29:03   Log-Likelihood:                -7802.1
No. Observations:                4358   AIC:                         1.562e+04
Df Residuals:                    4348   BIC:                         1.569e+04
Df Model:                           9
Covariance Type:                  HC1
==============================================================================
```

```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.3599      0.250    -17.452      0.000      -4.850      -3.870
electric      -0.1291      0.183     -0.704      0.482      -0.489       0.230
educ          -0.0721      0.007    -10.004      0.000      -0.086      -0.058
interaction   -0.0216      0.018     -1.183      0.237      -0.057       0.014
age            0.3434      0.019     17.897      0.000       0.306       0.381
age2          -0.0028      0.000     -7.905      0.000      -0.003      -0.002
urban         -0.2097      0.046     -4.564      0.000      -0.300      -0.120
spirit         0.1353      0.056      2.396      0.017       0.025       0.246
protest        0.0709      0.066      1.073      0.283      -0.059       0.200
catholic       0.1158      0.079      1.466      0.143      -0.039       0.271
==============================================================================
Omnibus:                      203.361   Durbin-Watson:                   1.890
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              716.487
Skew:                           0.013   Prob(JB):                    2.61e-156
Kurtosis:                       4.986   Cond. No.                     1.10e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 1.1e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```