JSC270 Assignment 2 Report - *Zihan Guo*

Chosen data: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

Background: The *UCI Adult dataset*, also known as the *Census Income dataset*, was extracted from the 1994 U.S. Census database. It contains 48,842 instances and includes various demographic and economic attributes such as *age, workclass, education level, marital status, occupation, race, sex, hours worked per week,* and *income group*. The dataset was originally created for predictive modeling tasks, particularly to determine whether an individual earns more than $50K per year. It is widely used in machine learning, statistical analysis, and labor market research to study income prediction, employment trends, and socioeconomic disparities.

Research Question: How does marital status influence weekly work hours?

Research Details: The goal is to determine whether being married affects the number of hours an individual works per week and whether additional factors such as *education and income* contribute to this relationship. To answer this question, *a linear regression model* was fitted using *hours-per-week* as the dependent variable. The initial model included *marital status* as the sole predictor, where individuals were categorized as either *alone* (if they were divorced, widowed, or never married) or *married* (the rest of non-null status). To refine the model, *education level (education-num)* and *income (income_binary)* were added to control for their potential influence on work hours.

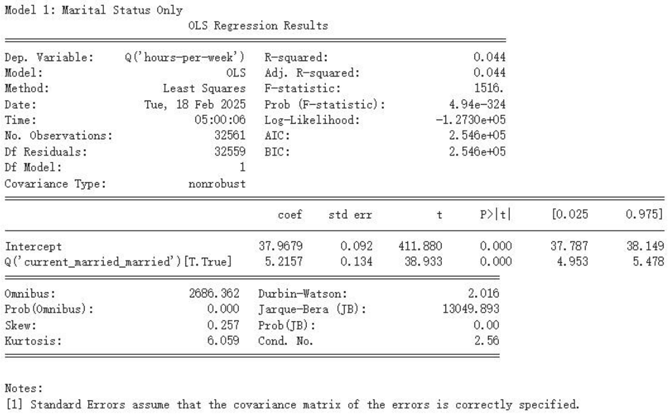
Further improvement was made by incorporating *workclass, occupation, race, relationship status, and native country* to capture broader socioeconomic effects.

Research Process: Using four models to figure out how marital status influence the working time. Each model builds upon the previous one, improving explanatory power and reducing prediction error, while also revealing potential model limitations. In Model 1(Appendix 1&5), *marital status* is the only predictor, showing that married individuals work *5.22 more hours* per week than those who are not married. This relationship is statistically significant (p-value = 0.000), but the low R-squared (0.044) suggests that marital status alone explains only a small portion of the variance in work hours. The high RMSE (12.07) indicates significant prediction errors, implying missing influential factors. Model 2(Appendix 2&5) adds *education (education-num)*, reducing the marital status effect to 4.96 hours, meaning education partly explains work hour differences. Each additional year of education *increases work hours by 0.64*, and the model fit improves (R-squared = 0.062, RMSE = 11.96), though errors remain high. Model 3(Appendix 3&5) incorporates income (income_binary), further reducing the marital status effect, confirming that income explains part of the relationship. The model's predictive power improves (R-squared = 0.075, RMSE = 11.87), but residual errors remain large. Model 4[improving model](Appendix 4&5) expands with additional predictors (*workclass, occupation, race, relationship status, and native country*), yielding the best fit (*R-squared increases, RMSE drops to 11.07*). However, multicollinearity issues suggest that some predictors are highly correlated, potentially distorting coefficient estimates.

Despite improvements, RMSE remains high, likely due to high variability in work hours, missing key predictors (job type, industry, household responsibilities), multicollinearity, and potential non-linearity. To reduce errors, we should remove multicollinear variables (using VIF), apply log transformations to skewed variables, introduce interaction terms, and explore non-linear models like decision trees or random forests to improve predictive accuracy.

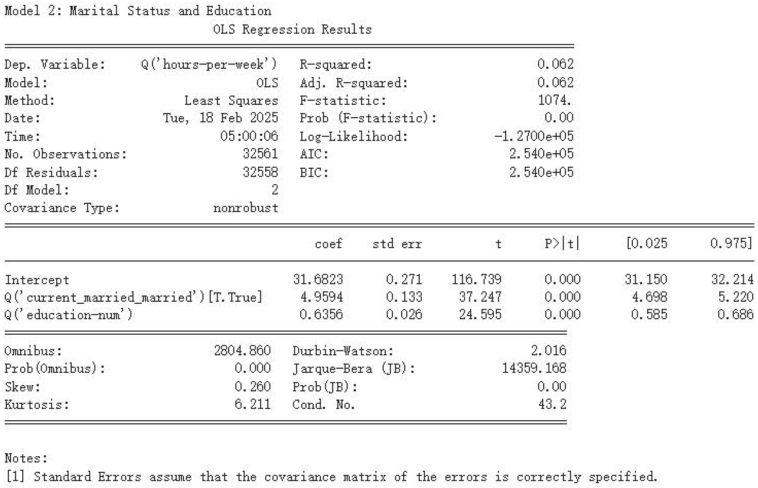
Appendix:

Appendix 1:



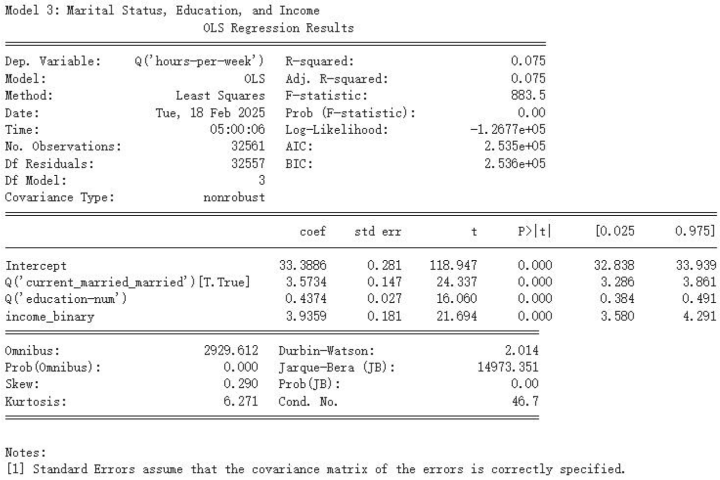
```
Model 1: Marital Status Only
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q(’hours-per-week’)   R-squared:                       0.044
Model:                             OLS   Adj. R-squared:                  0.044
Method:                  Least Squares   F-statistic:                     1516.
Date:                 Tue, 18 Feb 2025   Prob (F-statistic):          4.94e-324
Time:                         05:00:06   Log-Likelihood:            -1.2730e+05
No. Observations:                32561   AIC:                         2.546e+05
Df Residuals:                    32559   BIC:                         2.546e+05
Df Model:                            1
Covariance Type:             nonrobust
==================================================================================================
                                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                             37.9679      0.092    411.880      0.000      37.787      38.149
Q(’current_married_married’)[T.True]   5.2157      0.134     38.933      0.000       4.953       5.478
==============================================================================
Omnibus:                     2686.362   Durbin-Watson:                   2.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13049.893
Skew:                           0.257   Prob(JB):                         0.00
Kurtosis:                       6.059   Cond. No.                         2.56
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Appendix 2:



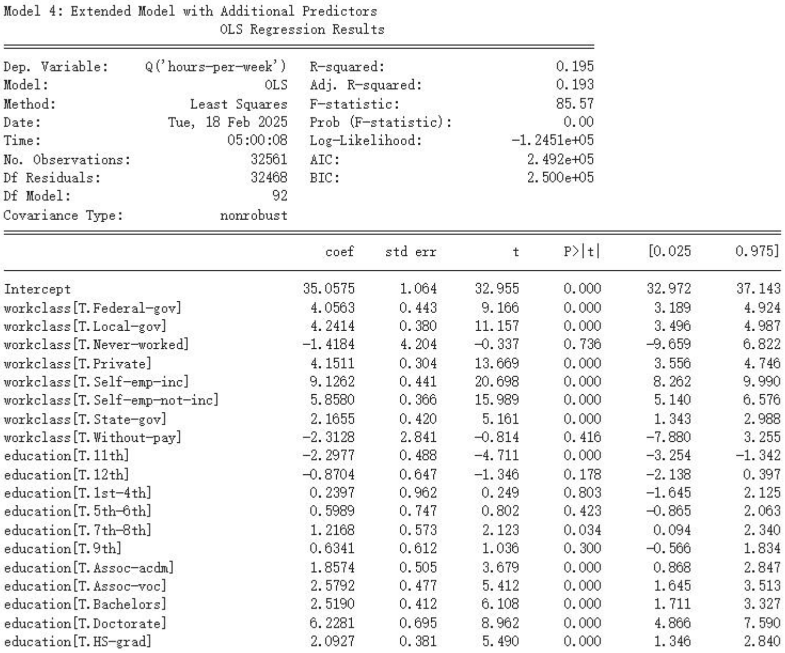
```
Model 2: Marital Status and Education
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q(’hours-per-week’)   R-squared:                       0.062
Model:                             OLS   Adj. R-squared:                  0.062
Method:                  Least Squares   F-statistic:                     1074.
Date:                 Tue, 18 Feb 2025   Prob (F-statistic):               0.00
Time:                         05:00:06   Log-Likelihood:            -1.2700e+05
No. Observations:                32561   AIC:                         2.540e+05
Df Residuals:                    32558   BIC:                         2.540e+05
Df Model:                            2
Covariance Type:             nonrobust
==================================================================================================
                                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                             31.6823      0.271    116.739      0.000      31.150      32.214
Q(’current_married_married’)[T.True]   4.9594      0.133     37.247      0.000       4.698       5.220
Q(’education-num’)                     0.6356      0.026     24.595      0.000       0.585       0.686
==============================================================================
Omnibus:                     2804.860   Durbin-Watson:                   2.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14359.168
Skew:                           0.260   Prob(JB):                         0.00
Kurtosis:                       6.211   Cond. No.                         43.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Appendix 3:



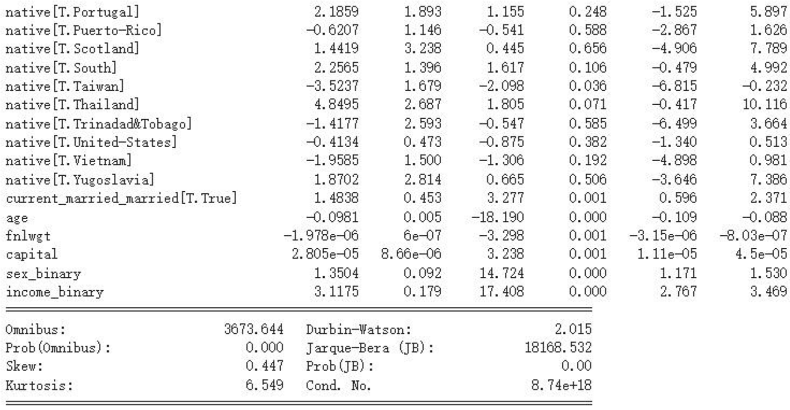
```
Model 3: Marital Status, Education, and Income
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q(’hours-per-week’)   R-squared:                       0.075
Model:                             OLS   Adj. R-squared:                  0.075
Method:                  Least Squares   F-statistic:                     883.5
Date:                 Tue, 18 Feb 2025   Prob (F-statistic):               0.00
Time:                         05:00:06   Log-Likelihood:            -1.2677e+05
No. Observations:                32561   AIC:                         2.535e+05
Df Residuals:                    32557   BIC:                         2.536e+05
Df Model:                            3
Covariance Type:             nonrobust
==================================================================================================
                                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                             33.3886      0.281    118.947      0.000      32.838      33.939
Q(’current_married_married’)[T.True]   3.5734      0.147     24.337      0.000       3.286       3.861
Q(’education-num’)                     0.4374      0.027     16.060      0.000       0.384       0.491
income_binary                          3.9359      0.181     21.694      0.000       3.580       4.291
==============================================================================
Omnibus:                     2929.612   Durbin-Watson:                   2.014
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14973.351
Skew:                           0.290   Prob(JB):                         0.00
Kurtosis:                       6.271   Cond. No.                         46.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Appendix 4:



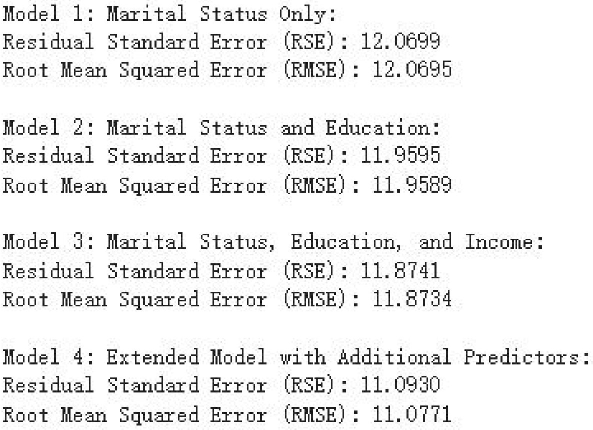
```
Model 4: Extended Model with Additional Predictors
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q('hours-per-week')   R-squared:                       0.195
Model:                             OLS   Adj. R-squared:                  0.193
Method:                  Least Squares   F-statistic:                     85.57
Date:                 Tue, 18 Feb 2025   Prob (F-statistic):               0.00
Time:                         05:00:08   Log-Likelihood:            -1.2451e+05
No. Observations:                32561   AIC:                         2.492e+05
Df Residuals:                    32468   BIC:                         2.500e+05
Df Model:                           92
Covariance Type:             nonrobust
==============================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                         35.0575      1.064     32.955      0.000      32.972      37.143
workclass[T.Federal-gov]           4.0563      0.443      9.166      0.000       3.189       4.924
workclass[T.Local-gov]             4.2414      0.380     11.157      0.000       3.496       4.987
workclass[T.Never-worked]         -1.4184      4.204     -0.337      0.736      -9.659       6.822
workclass[T.Private]               4.1511      0.304     13.669      0.000       3.556       4.746
workclass[T.Self-emp-inc]          9.1262      0.441     20.698      0.000       8.262       9.990
workclass[T.Self-emp-not-inc]      5.8580      0.366     15.989      0.000       5.140       6.576
workclass[T.State-gov]             2.1655      0.420      5.161      0.000       1.343       2.988
workclass[T.Without-pay]          -2.3128      2.841     -0.814      0.416      -7.880       3.255
education[T.11th]                 -2.2977      0.488     -4.711      0.000      -3.254      -1.342
education[T.12th]                 -0.8704      0.647     -1.346      0.178      -2.138       0.397
education[T.1st-4th]               0.2397      0.962      0.249      0.803      -1.645       2.125
education[T.5th-6th]               0.5989      0.747      0.802      0.423      -0.865       2.063
education[T.7th-8th]               1.2168      0.573      2.123      0.034       0.094       2.340
education[T.9th]                   0.6341      0.612      1.036      0.300      -0.566       1.834
education[T.Assoc-acdm]            1.8574      0.505      3.679      0.000       0.868       2.847
education[T.Assoc-voc]             2.5792      0.477      5.412      0.000       1.645       3.513
education[T.Bachelors]             2.5190      0.412      6.108      0.000       1.711       3.327
education[T.Doctorate]             6.2281      0.695      8.962      0.000       4.866       7.590
education[T.HS-grad]               2.0927      0.381      5.490      0.000       1.346       2.840
```

...



```
native[T.Portugal]                 2.1859      1.893      1.155      0.248      -1.525       5.897
native[T.Puerto-Rico]             -0.6207      1.146     -0.541      0.588      -2.867       1.626
native[T.Scotland]                 1.4419      3.238      0.445      0.656      -4.906       7.789
native[T.South]                    2.2565      1.396      1.617      0.106      -0.479       4.992
native[T.Taiwan]                  -3.5237      1.679     -2.098      0.036      -6.815      -0.232
native[T.Thailand]                 4.8495      2.687      1.805      0.071      -0.417      10.116
native[T.Trinadad&Tobago]         -1.4177      2.593     -0.547      0.585      -6.499       3.664
native[T.United-States]           -0.4134      0.473     -0.875      0.382      -1.340       0.513
native[T.Vietnam]                 -1.9585      1.500     -1.306      0.192      -4.898       0.981
native[T.Yugoslavia]               1.8702      2.814      0.665      0.506      -3.646       7.386
current_married_married[T.True]    1.4838      0.453      3.277      0.001       0.596       2.371
age                               -0.0981      0.005    -18.190      0.000      -0.109      -0.088
fnlwgt                         -1.978e-06      6e-07     -3.298      0.001   -3.15e-06   -8.03e-07
capital                         2.805e-05   8.66e-06      3.238      0.001    1.11e-05     4.5e-05
sex_binary                         1.3504      0.092     14.724      0.000       1.171       1.530
income_binary                      3.1175      0.179     17.408      0.000       2.767       3.469
==============================================================================
Omnibus:                     3673.644   Durbin-Watson:                   2.015
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            18168.532
Skew:                           0.447   Prob(JB):                         0.00
Kurtosis:                       6.549   Cond. No.                     8.74e+18
==============================================================================
```

Appendix 5:



```
Model 1: Marital Status Only:
Residual Standard Error (RSE): 12.0699
Root Mean Squared Error (RMSE): 12.0695

Model 2: Marital Status and Education:
Residual Standard Error (RSE): 11.9595
Root Mean Squared Error (RMSE): 11.9589

Model 3: Marital Status, Education, and Income:
Residual Standard Error (RSE): 11.8741
Root Mean Squared Error (RMSE): 11.8734

Model 4: Extended Model with Additional Predictors:
Residual Standard Error (RSE): 11.0930
Root Mean Squared Error (RMSE): 11.0771
```