Logo

Description automatically generated

**ALY 6015: INTERMEDIATE ANALYTICS**

**WEEK 6: FINAL PROJECT REPORT**

**National Longitudinal Survey of Youth 1997**

**Prof. Roy Wada**

**Team Members: Kiran Kasu| Abhinav Jain| Kelvina Pethani| Tanusha Erpula**

**Date: 05/21/2022**

**Introduction**

**Dataset:** Youth National Longitudinal Survey, 1997. In the 1997 National Longitudinal Survey of Youth, 8,984 men and women born between 1980 and 1984 who lived in the United States in 1997 were included. On December 31, 1996, I was between the ages of 12 and 16. From 1997 through 2011, interviews were held annually, then every two years. So far, the current cohort has been surveyed 18 times. From Round 1 (1997-98) until Round 19, data is provided (2019-20).

The NLSY97 collects a lot of information about people's work habits and educational experiences. The poll also asks about the children's homes and neighborhoods, which will help researchers figure out how schooling and other variables affect young newcomers.

**The goal of the project:**

The goal of the NLSY97 is to evaluate youth interviews done in 1979 in the United States.

**Dataset Description:**

The raw dataset has 1,52711 rows and 32 fields. After cleaning the data set, there were 1,41,392 rows and 18 areas. The year of birth is a numeric data type representing a child's birth year. All text data types, including country of birth, sample race, and area, are text data types. The number of Children is a numeric data type that refers to the number of people in the family, whereas Family Size is a numeric data type that refers to the size of the family.

**Methods used:**

Linear regression will be used to create prediction models in this project. The minor square function represents the relationship between one or more independent variables and dependent variables in a linear regression study. A linear combination of one or more model parameters makes up this function (regression coefficients).

We frequently added a regular term to the goal function to keep the parameters as "simple" as possible while ensuring the best fit error and generalizing the model. We also used the Correlation for the selection of the model, moreover we applied stepwise selection model for the selection of the variable with their significance level of variables in the dataset.

**Correlation Model:** We also know a correlation matrix as the Correlogram. It highlights the correlated variable in the data we choose, where in this plot, we're here in this matrix correlation. Coefficients are colour-coded according to the value of the dataset.

The positive coefficient indicates that the value of the independent variables has increased in comparison to the dependent variables. When the dependent variable drops, the coefficient value represents the mean of the dependent variables, which changes when the independent variables move by one unit while the other variables in the model remain constant. This characteristic also holds the other variables constant, which is difficult to do because it permits accessing the effect of different variables on other variables.

**Linear Regression Model:** The motive for choosing this linear regression model is to analyze a group of statistical modelling to display and determine the relationship between the dataset and the variables which includes the wages as a dependent variable whereas independent variables after correlation and stepwise selection method year, marital status, poverty ratio, partner id. The regression analysis was used to define the linear relationship between the independent and dependent variables in the dataset. Inferential statistics are used in this linear regression analysis. The p-values are used to assess whether a relationship exists in the sample and in the larger population. For the null hypothesis that the variables do not correlate with the dependent variable, the p-value for each variable is determined by the independent variables. There is no correlation between the changes in the independent variables and the shifts in the dependent variable if there must be none.

Linear regression will be used to create prediction models in this project. The minor square function represents the relationship between one or more independent variables and dependent variables in a linear regression study. A linear combination of one or more model parameters makes up this function (regression coefficients).

We frequently added a regular term to the goal function to keep the parameters as "simple" as possible while ensuring the best fit error and generalizing the model.

**Research Questions:**

1. Which ‘Race’ is getting the highest wages among all the other races?

2. Which Gender is getting paid the highest irrespective of their profession?

3. Which group in “marital status” gets the highest wages irrespective of the degree?

4. Define the correlation between gender and the wages they earn in their lifetime?

**EXPLORATORY DATA ANALYSIS (EDA)**

**Summary Statistics Tables:**

Descriptive Analysis: Table 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **mean** | **sd** | **median** | **min** | **max** |
| **YEAR** | 2006.20496535797 | 4.40622868298385 | 2007 | 1997 | 2013 |
| **KEY\_SEX\*** | 1.53856205178072 | 0.498518324779857 | 2 | 1 | 2 |
| **KEY\_RACE\*** | 4.36732709371581 | 0.92640663582745 | 5 | 1 | 5 |
| **CV\_SAMPLE\_TYPE\*** | 1.20900692840647 | 0.406605529022681 | 1 | 1 | 2 |
| **KEY\_RACE\_ETHNICITY\*** | 2.96812325270451 | 1.27932002115855 | 4 | 1 | 4 |
| **KEY\_BDATE\_Y** | 1981.83192536769 | 1.37820731784712 | 1982 | 1980 | 1984 |
| **CV\_MARSTAT\_\*** | 5.22085814999392 | 1.1170421820912 | 6 | 1 | 10 |
| **CV\_HH\_POV\_RATIO\_** | 346.189133341437 | 369.50124655369 | 281 | -4 | 3227 |
| **CV\_HH\_SIZE\_** | 3.32405494104777 | 1.62372065914759 | 3 | 1 | 19 |
| **CVC\_FIRST\_MARRY\_DATE\_Y\_XRND** | 1092.40269843199 | 1001.11684121024 | 2001 | -4 | 2014 |
| **CVC\_MARRIAGES\_TTL\_XRND** | 0.600370730521454 | 0.588725099729283 | 1 | 0 | 5 |
| **CHILD\_BIRTH\_01\_Y** | 1166.62863133585 | 991.073969451529 | 2000 | -4 | 2014 |
| **CV\_BIO\_CHILD\_HH\_** | -2.22681414853531 | 2.59804585378014 | -4 | -4 | 8 |
| **WAGE\_** | 24810.5953567522 | 22096.148491092 | 30000 | 4 | 99999 |
| **PARTNERS\_UID\_01\_** | 87088.0078704266 | 99380.2148910325 | -4 | -4 | 201304 |
| **PARTNERS\_UID\_02\_** | 2193.13467849763 | 20891.7422359365 | -4 | -4 | 201302 |
| **AGE\_FIRST\_CHILD** | 11.9178619180746 | 13.8857072552609 | 18 | -4 | 34 |
| **AGE\_** | 23.7349580649082 | 4.43463380368045 | 24 | 15 | 34 |

KEY\_Race and Wage\_: Table 2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **KEY\_RACE** | **mean** | **median** | **sd** | **min** | **max** |
| **1** | American Indian, Eskimo, or Aleut | 21908.1314285714 | 8000 | 21079.2607102769 | 35 | 90000 |
| **2** | Asian or Pacific Islander | 31144.8300455235 | 33000 | 25230.1611189182 | 6 | 95000 |
| **3** | Black or African American | 20896.3702888583 | 8000 | 20390.5215307895 | 4 | 99800 |
| **4** | Something else? (SPECIFY) | 25629.6058839856 | 30000 | 21926.1596038487 | 5 | 99800 |
| **5** | White | 25840.6973255978 | 30000 | 22415.6492558767 | 4 | 99999 |

Year vs Wage: Table 3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **YEAR** | **mean** | **median** | **sd** | **min** | **max** |
| **1** | 1997 | 1426.48536585366 | 600 | 1766.21711459077 | 5 | 9000 |
| **2** | 1998 | 2538.9686450168 | 900 | 2357.41017859363 | 6 | 8900 |
| **3** | 1999 | 3458.36607142857 | 3500 | 2607.94525720597 | 5 | 9900 |
| **4** | 2000 | 5174.66045428072 | 4000 | 6081.96361003219 | 6 | 30623 |
| **5** | 2001 | 5534.10216572505 | 4200 | 6786.18463217055 | 5 | 35558 |
| **6** | 2002 | 7174.87959183674 | 5000 | 9125.25031026011 | 4 | 42458 |
| **7** | 2003 | 9297.03138297872 | 5000 | 11502.4007272497 | 5 | 44091 |
| **8** | 2004 | 14223.2109283196 | 6500 | 17485.8160564518 | 30 | 70712 |
| **9** | 2005 | 19412.3267605634 | 8000 | 18454.0525598245 | 50 | 69342 |
| **10** | 2006 | 26221.9381642512 | 30000 | 20367.1312304834 | 5 | 80471 |
| **11** | 2007 | 30032.1139575972 | 34000 | 17848.5699155758 | 40 | 70000 |
| **12** | 2008 | 33572.8372177713 | 35000 | 17678.1604788179 | 30 | 74000 |
| **13** | 2009 | 36672.7359364659 | 37750 | 18594.4895206445 | 30 | 84000 |
| **14** | 2010 | 38501.7335854245 | 38500 | 19271.700645215 | 4 | 88000 |
| **15** | 2011 | 40485.6453621763 | 40000 | 20241.5172675831 | 5 | 94000 |
| **16** | 2013 | 43671.9092927371 | 43000 | 21376.1191455325 | 50 | 99999 |

Descriptive table with new variables: Table 4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **mean** | **sd** | **median** | **min** | **max** |
| **YEAR** | 2006.20496535797 | 4.40622868298385 | 2007 | 1997 | 2013 |
| **KEY\_SEX\*** | 1.53856205178072 | 0.498518324779857 | 2 | 1 | 2 |
| **KEY\_RACE\*** | 4.36732709371581 | 0.92640663582745 | 5 | 1 | 5 |
| **CV\_SAMPLE\_TYPE\*** | 1.20900692840647 | 0.406605529022681 | 1 | 1 | 2 |
| **KEY\_RACE\_ETHNICITY\*** | 2.96812325270451 | 1.27932002115855 | 4 | 1 | 4 |
| **KEY\_BDATE\_Y** | 1981.83192536769 | 1.37820731784712 | 1982 | 1980 | 1984 |
| **CV\_MARSTAT\_\*** | 5.22085814999392 | 1.1170421820912 | 6 | 1 | 10 |
| **CV\_HH\_POV\_RATIO\_** | 346.189133341437 | 369.50124655369 | 281 | -4 | 3227 |
| **CV\_HH\_SIZE\_** | 3.32405494104777 | 1.62372065914759 | 3 | 1 | 19 |
| **CVC\_FIRST\_MARRY\_DATE\_Y\_XRND** | 1092.40269843199 | 1001.11684121024 | 2001 | -4 | 2014 |
| **CVC\_MARRIAGES\_TTL\_XRND** | 0.600370730521454 | 0.588725099729283 | 1 | 0 | 5 |
| **CHILD\_BIRTH\_01\_Y** | 1166.62863133585 | 991.073969451529 | 2000 | -4 | 2014 |
| **CV\_BIO\_CHILD\_HH\_** | -2.22681414853531 | 2.59804585378014 | -4 | -4 | 8 |
| **WAGE\_** | 24810.5953567522 | 22096.148491092 | 30000 | 4 | 99999 |
| **PARTNERS\_UID\_01\_** | 87088.0078704266 | 99380.2148910325 | -4 | -4 | 201304 |
| **PARTNERS\_UID\_02\_** | 2193.13467849763 | 20891.7422359365 | -4 | -4 | 201302 |
| **AGE\_FIRST\_CHILD** | 11.9178619180746 | 13.8857072552609 | 18 | -4 | 34 |
| **AGE\_** | 23.7349580649082 | 4.43463380368045 | 24 | 15 | 34 |
| **YEAR97\_04** | 0.297982253555366 | 0.457378603666578 | 0 | 0 | 1 |
| **YEAR05\_13** | 0.583900571289656 | 0.49291792155479 | 1 | 0 | 1 |

Descriptive table of train dataset :Table 5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **mean** | **sd** | **median** | **min** | **max** |
| **YEAR** | 2006.19524424523 | 4.40667905649069 | 2007 | 1997 | 2013 |
| **KEY\_SEX\*** | 1.53631391020284 | 0.498689027860924 | 2 | 1 | 2 |
| **KEY\_RACE\*** | 4.36621590822761 | 0.927501364807593 | 5 | 1 | 5 |
| **CV\_SAMPLE\_TYPE\*** | 1.20907088049837 | 0.406652835864801 | 1 | 1 | 2 |
| **KEY\_RACE\_ETHNICITY\*** | 2.9670667780901 | 1.27928112078849 | 4 | 1 | 4 |
| **KEY\_BDATE\_Y** | 1981.83092760009 | 1.37921704185644 | 1982 | 1980 | 1984 |
| **CV\_MARSTAT\_\*** | 5.22236572209983 | 1.11551726468877 | 6 | 1 | 10 |
| **CV\_HH\_POV\_RATIO\_** | 343.06366329864 | 365.798855635768 | 278.5 | -4 | 3227 |
| **CV\_HH\_SIZE\_** | 3.32336853300919 | 1.62373379615578 | 3 | 1 | 15 |
| **CVC\_FIRST\_MARRY\_DATE\_Y\_XRND** | 1093.96577527919 | 1000.98131752341 | 2001 | -4 | 2014 |
| **CVC\_MARRIAGES\_TTL\_XRND** | 0.600622958292183 | 0.587910807930135 | 1 | 0 | 4 |
| **CHILD\_BIRTH\_01\_Y** | 1167.47971587024 | 990.943981168744 | 2000 | -4 | 2014 |
| **CV\_BIO\_CHILD\_HH\_** | -2.23019068601383 | 2.59697423793679 | -4 | -4 | 8 |
| **WAGE\_** | 24741.6148674314 | 22046.6998369178 | 30000 | 4 | 99999 |
| **PARTNERS\_UID\_01\_** | 87142.4517967029 | 99387.5752562842 | -4 | -4 | 201304 |
| **PARTNERS\_UID\_02\_** | 2200.61912178075 | 20926.1261594599 | -4 | -4 | 201302 |
| **AGE\_FIRST\_CHILD** | 11.9407050064575 | 13.8953467933584 | 18 | -4 | 34 |
| **AGE\_** | 23.7251386462053 | 4.43472913889703 | 24 | 15 | 34 |
| **YEAR97\_04** | 0.297690496087518 | 0.457251360347754 | 0 | 0 | 1 |
| **YEAR05\_13** | 0.583947428397782 | 0.492911815888775 | 1 | 0 | 1 |

Descriptive table of test dataset: Table 6

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **mean** | **sd** | **median** | **min** | **max** |
| **YEAR** | 2006.24384685506 | 4.40454700401016 | 2007 | 1997 | 2013 |
| **KEY\_SEX\*** | 1.54755393497417 | 0.497771300693518 | 2 | 1 | 2 |
| **KEY\_RACE\*** | 4.37177149802492 | 0.922071881348551 | 5 | 1 | 5 |
| **CV\_SAMPLE\_TYPE\*** | 1.20875113947129 | 0.406447044365364 | 1 | 1 | 2 |
| **KEY\_RACE\_ETHNICITY\*** | 2.97234883014281 | 1.2795640763005 | 4 | 1 | 4 |
| **KEY\_BDATE\_Y** | 1981.83591613491 | 1.37425861963469 | 1982 | 1980 | 1984 |
| **CV\_MARSTAT\_\*** | 5.21437253114555 | 1.1214449211719 | 6 | 1 | 9 |
| **CV\_HH\_POV\_RATIO\_** | 358.690063810392 | 383.727026143118 | 289 | -4 | 2627 |
| **CV\_HH\_SIZE\_** | 3.32680036463081 | 1.62378857001623 | 3 | 1 | 19 |
| **CVC\_FIRST\_MARRY\_DATE\_Y\_XRND** | 1086.15086599818 | 1001.71039983952 | 2001 | -4 | 2014 |
| **CVC\_MARRIAGES\_TTL\_XRND** | 0.599361896080219 | 0.592014596504449 | 1 | 0 | 5 |
| **CHILD\_BIRTH\_01\_Y** | 1163.22455180796 | 991.661723283389 | 2000 | -4 | 2014 |
| **CV\_BIO\_CHILD\_HH\_** | -2.21330902461258 | 2.60248131598816 | -4 | -4 | 6 |
| **WAGE\_** | 25086.4963536919 | 22292.3832122699 | 30000 | 4 | 99800 |
| **PARTNERS\_UID\_01\_** | 86870.2487085992 | 99358.0210164321 | -4 | -4 | 201303 |
| **PARTNERS\_UID\_02\_** | 2163.19917958067 | 20755.2024135432 | -4 | -4 | 201302 |
| **AGE\_FIRST\_CHILD** | 11.8264965056214 | 13.8477615986494 | 18 | -4 | 33 |
| **AGE\_** | 23.7742327560012 | 4.43437193307891 | 24 | 16 | 33 |
| **YEAR97\_04** | 0.299149194773625 | 0.457920093751263 | 0 | 0 | 1 |
| **YEAR05\_13** | 0.583713157095108 | 0.492979746652355 | 1 | 0 | 1 |

**Bar Chart of Race & Wage:**

* White has maximum wages than another race
* America and Indian has minimum wages
* Black or African has average wages

**Graph 1: Bar Chart of Race & Wage**

Chart, bar chart

Description automatically generated

**Bar Chart of Sex & Wage:**

* Males have higher wages than the females

**Graph 2: Bar Chart of Sex & Wage**

Chart, bar chart

Description automatically generated

**Bar Char of Marital Status and Wage:**

* “Never married" and "not cohabiting” have the highest wages in comparison to other groups.

Chart, histogram

Description automatically generated

**Graph 3: Bar Chart of Marital Status & Wage**

**Comparison Chart of Gender and Year:**

* Graph​: Gender and Year​
* Group​: Between the age group of 10-15, wages are increasing gradually.​
* Age​: On the other hand, age group of 18- 25, wages are stagnant.​
* Result​: As age increases, the wages decreases ​

Chart, histogram

Description automatically generated

**Graph 4: Comparison of Gender and Year**

**Correlation Model:**The correlation matrix demonstrates that the variables we are using have a linear relationship with each other. Variables used:

* Wages with year
* Wages with partners id 01
* Wages with CV poverty ratio
* Year with partners id 01

**Correlation Table: Correlation Between Variables**

Table

Description automatically generated with low confidence

**Correlation MATRIX: Correlation Between Variables**

Chart, bubble chart

Description automatically generated

**Stepwise Selection Methods:**

**Step1: Forward Selection MethodTable

Description automatically generated**

**Step2: Backward Selection Model**

**Table

Description automatically generated**

**Prediction Model**

**Linear Regression Analysis:**

In this linear regression model using dependent variable Wage on the other hand independent variable majorly focus on Year, and Marital Status. This model will help in predict the wages in the future as per their marital status whether the person is married, divorced, cohabitating, etc. depending on various factor this model will suggest the wages of youth.

**Model 1.**

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

**Linear Regression Wages with Year 1997-2004:**

On the other hand, we can check the individual interpretation of the independent variable which help in determining the variability of the dummy variable while predicting the wages in the future years. In this model we have taken wage as dependent variable whereas, independent variable was dummy from year 1997 to 2004. By which we can see the p-value is significant, whereas the variability is 31 % and coefficient 32675.29, standard error 135.12, and the t value is 241.82 other the other hand p – value is approaching to zero.

**Graph 5: Linear Regression B/w Wage & Year**

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, line chart

Description automatically generated

**Linear Regression Wages with Year 2005-2013:**

On the other hand, we can check the individual interpretation of the independent variable which help in determining the variability of the dummy variable while predicting the wages in the future years. In this model we have taken wage as dependent variable whereas, independent variable was dummy from year 2005 to 2013. By which we can see the p-value is significant, whereas the variability is 38 % and coefficient 8561.93, standard error 165.37, and the t value is 51.77 other the other hand p – value is approaching to zero. While implementing a linear regression model with the dummy variable, wages, and dependent variable between the years 2005 and 2013, we found the accuracy of the model to 58%.

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, line chart

Description automatically generated

**Linear Regression Wages with Marital Status:**

• The goal is to create an equation for the respondents' marital status so that you can use it to forecast their wages.

• As a result, it is preferable to construct a linear regression model with Wages as the response variable and the predictor.

• It's a good idea to evaluate and comprehend the variables before we start developing the regression model. The accuracy rate is 58%, whereas the absolute percentage error is 8.6

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated with medium confidence

Chart

Description automatically generated

**Conclusion**

In this final project report, we have included the introduction of the data set, the methods that we used to analyze and create the answers for the questions we asked to work on the data set. We have also included linear regression models that we have created for the dataset. We have added the summary statistics and exploratory analysis for examine the dataset at it fullest.

As we wanted to work on the dependent variables, we also added the correlation matrix for analysis the dependent and independent variables to check the collinearity of predict variables, Step wise selection model for the selection of the predictive variables with their significance level, linear regression and further also worked on prediction models such as a scatterplot to understand the relation between predictive variables.

The goal is to build an equation that may be used to forecast the respondents' income based on their marital status. As a result, using a linear regression model with Wages as the response variable and predictor is preferable. Before starting to create the regression model, it's a good idea to examine and understand the variables. On the other hand, we can predict whether salaries will rise or fall depending on marital status and the year.

For further, about the project we will cleaning refine our model to increase the accuracy of the model atleast upto 90%. High the accuracy higher the acceptability to predict the outcome from the predictive variable.

**References**

1. Jim Frost (Feb 27, 2022) How to Interpret P-values and Coefficients in Regression Analysis Retrieved on (May 17, 2022) from <https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/>
2. Adam Hayes (Jan 02, 2022) Multiple Linear Regression (MLR) Definition Retrieved on (May 17, 2022) from <https://www.investopedia.com/terms/m/mlr.asp>
3. Rebecca Bevans(Feb 20, 2020) An introduction to multiple linear regression Retrieved on (May 17, 2022) from <https://www.scribbr.com/statistics/multiple-linear-regression/>
4. Linear Regression Calculator Retrieved on (May 17, 2022) from <https://www.socscistatistics.com/tests/regression/default.aspx>
5. Linear regression - A complete introduction in R with examples*.* (2017, March 12). Machine Learning Plus.   
   [Https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/](https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/)
6. Linear Regression - A Complete Introduction in R with Examples

Prabhakaran

[Https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/](https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/)