**Case Problem**

**Predictive Analytics**

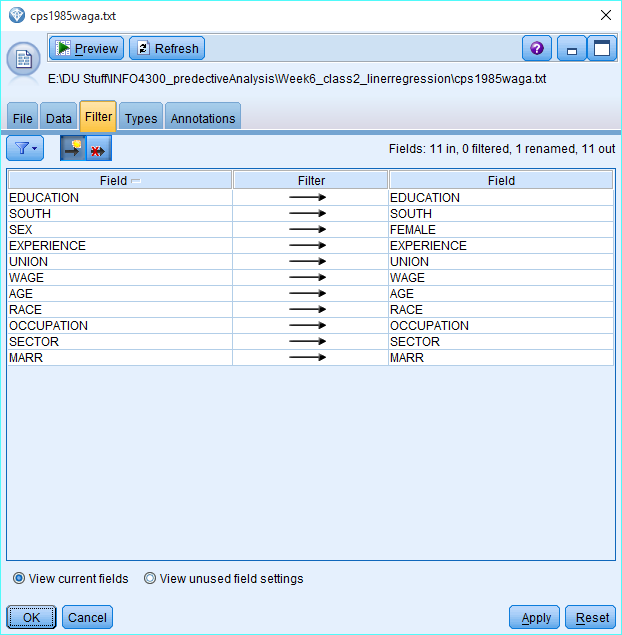
**Module 2 – Lesson 3**

**Professor: Lee**

**Make sure you have gone over PAM2L3inClassDemo.pdf from Canvas (I will hand out this in class too) before working on this case problem.**

**PART 1 (Data Prep.)**

1. Read in the CPC1985WageData.txt file. Change the name of the variable “SEX” to be “FEMALE” in the Filter Tab.



1. Note the description of the data below:

Determinants of Wages from the 1985 Current Population Survey

Description: There are 534 observations on 11 variables sampled from the Current Population Survey of 1985.

Variable Descriptions:

EDUCATION: Number of years of education.

SOUTH: Indicator variable for Southern Region (1=Person lives in South, 0=Person lives elsewhere).

SEX -> FEMALE: Indicator variable for sex (1=Female, 0=Male).

EXPERIENCE: Number of years of work experience.

UNION: Indicator variable for union membership (1=Union member, 0=Not union member).

WAGE: Wage (dollars per hour).

AGE: Age (years).

RACE: Race (1=Other, 2=Hispanic, 3=White).

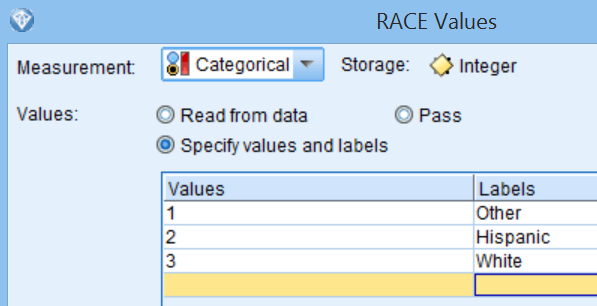
OCCUPATION: Occupational category (1=Management, 2=Sales, 3=Clerical, 4=Service, 5=Professional, 6=Other).

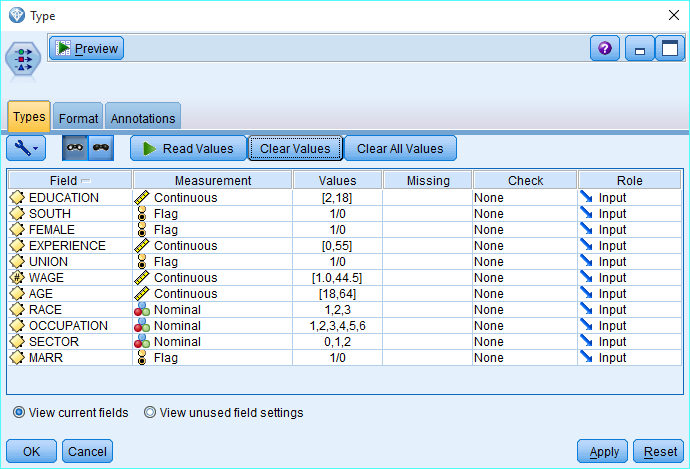
SECTOR: Sector (0=Other, 1=Manufacturing, 2=Construction).

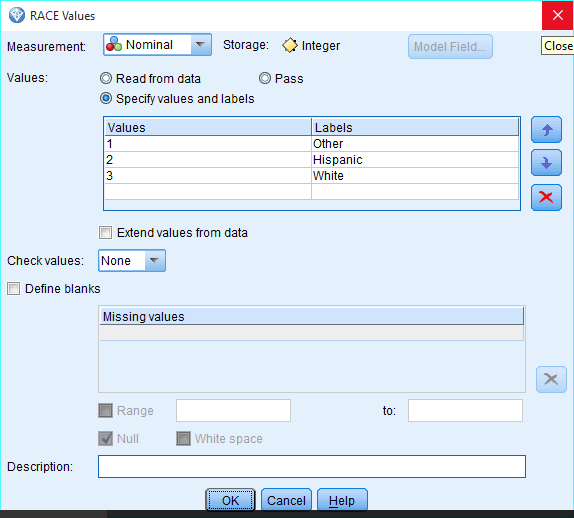
MARR: Marital Status (0=Unmarried, 1=Married)

1. Make sure that all variables have the correct type before continuing.

For RACE, OCCUPATION, and SECTOR – double click on the field name in the Type node and choose ‘Specify values and labels’ and specify labels for the values 0, 1, 2, 3, etc. as noted above. See example for RACE.



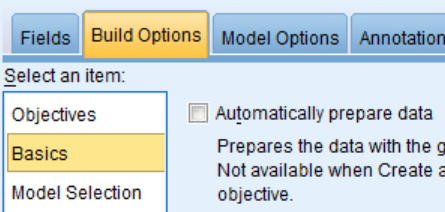
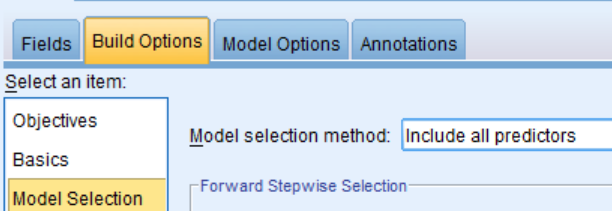




**PART 2 (Building a linear regression model)**

1. Create a linear regression model using all **10 variables** to predict **WAGE**.

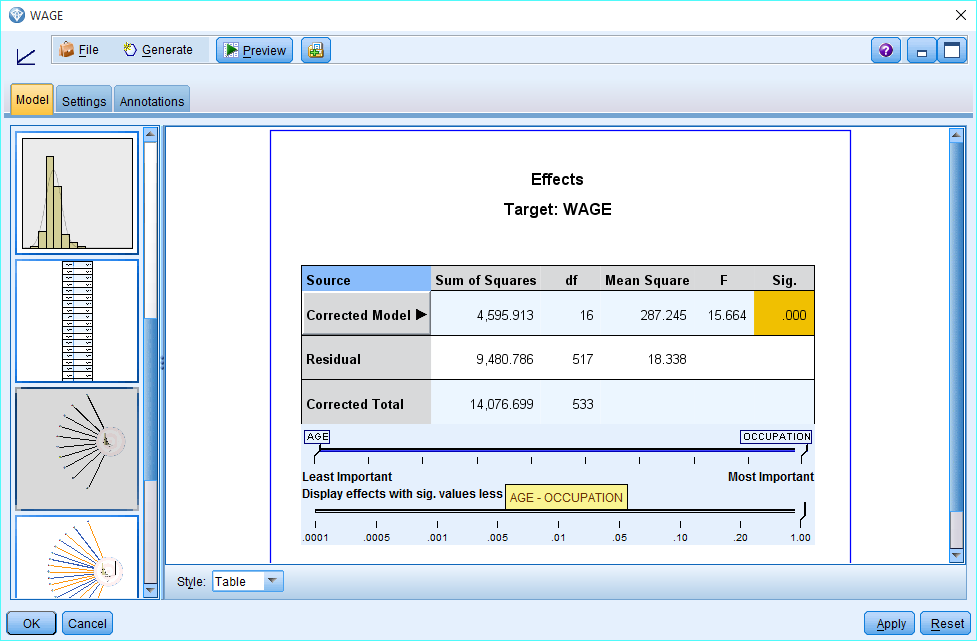
Make sure to uncheck ‘Automatically prepare data’ in Basics and for Model selection choose ‘include all predictors.’

1. What is your R2? Include a screenshot of your ANOVA table (Showing Sum of Squares values) and verify the calculation of the R2.

Answer:

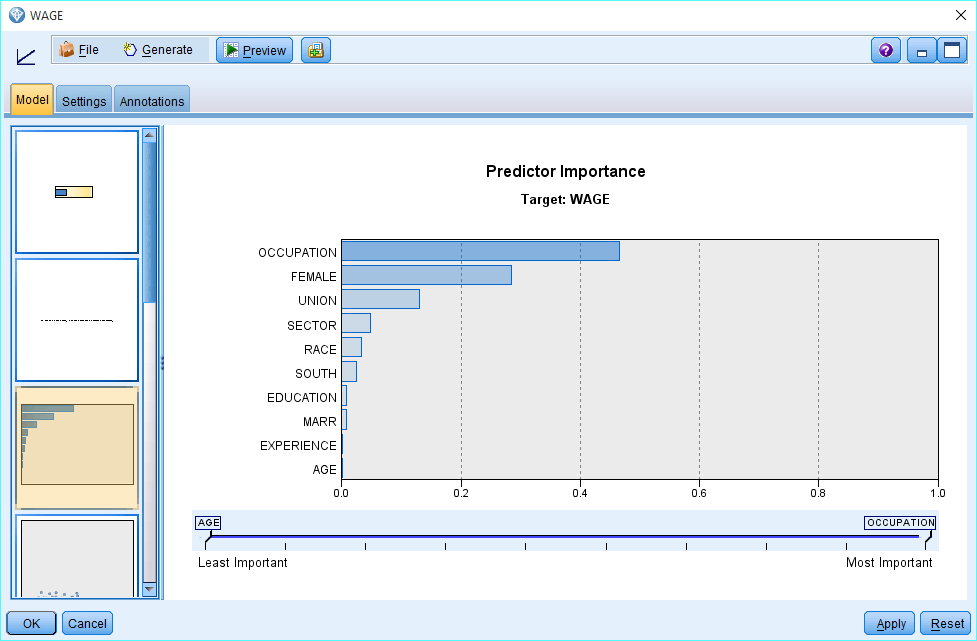
Adjusted R2 is .306



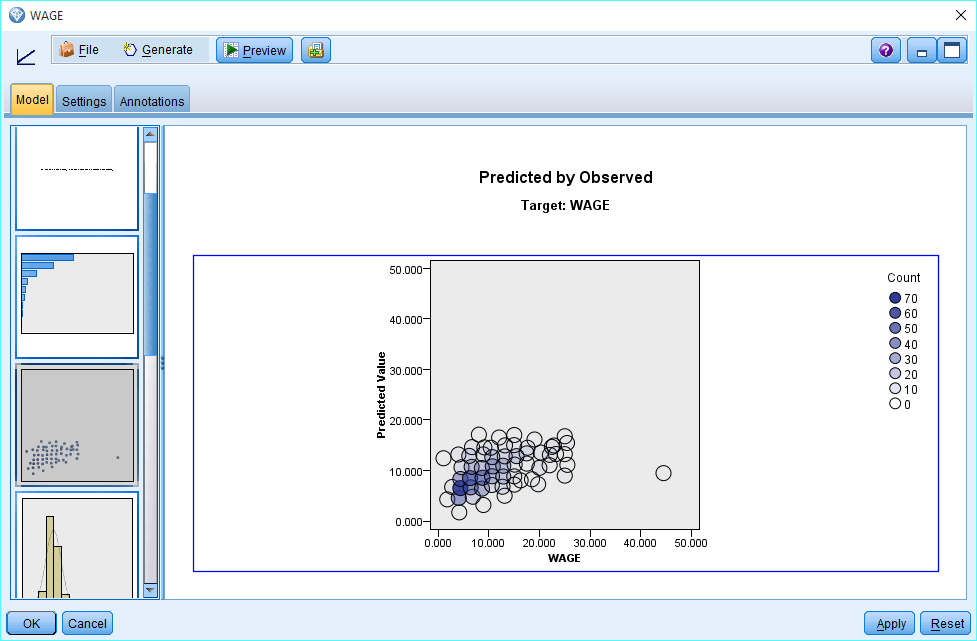
Rsquare = Sum of Squares( Corrected model) / Sum of squares(Corrected Total)

= 4595.913/14076.699 = .326

1. Include a screenshot of your Predictor Importance Graph and the Predicted by Observed Scatterplot. Describe/interpret what the graphs show.

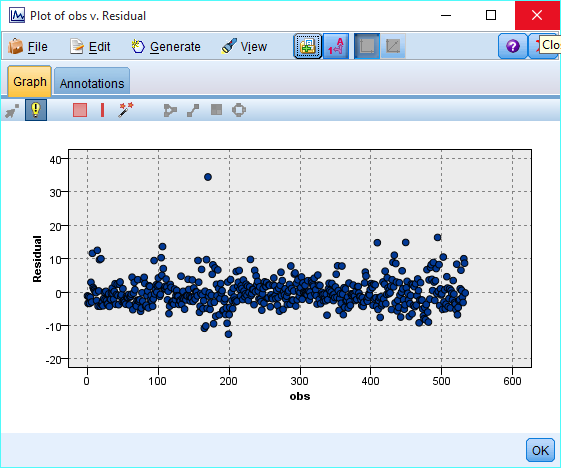


In the graph, we see that occupation,FEMALE and UNION have the highest predictor importance whereas Experience and Age have the lowest predictor importance.

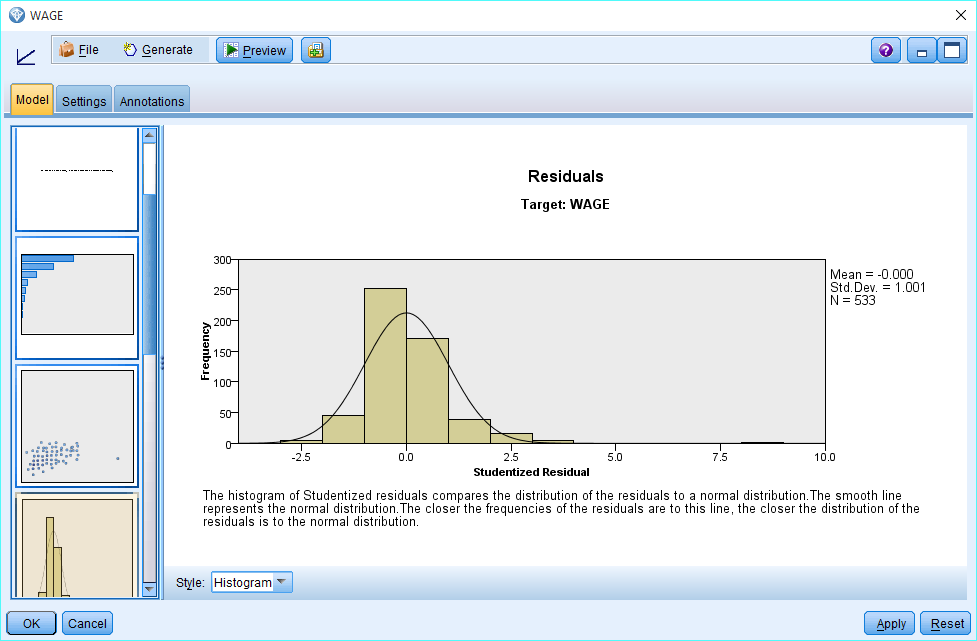


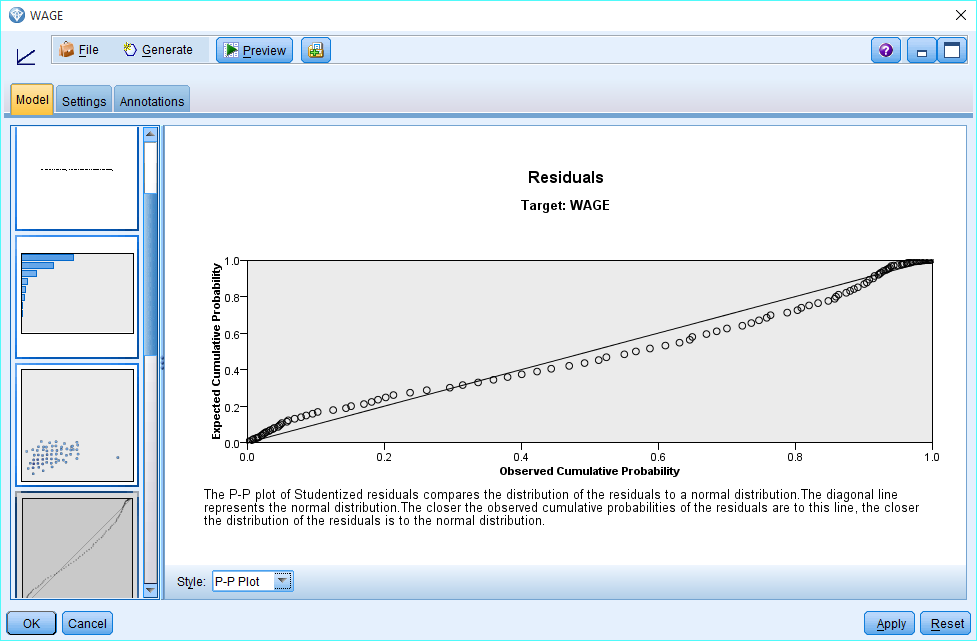
From the predicted, observed graph, we can see that our model is not predicting the wages very well. In a perfect situation, we would see all scatter plot circles on or around a 45 degree line

1. Check the 3 residual assumptions for regression using graphs and correlations as shown in the demo. Include screenshots of each of these and explain how you think about the assumptions for the model.
2. Constant Variance
3. Normality
4. Residuals are independent (no autocorrelation)

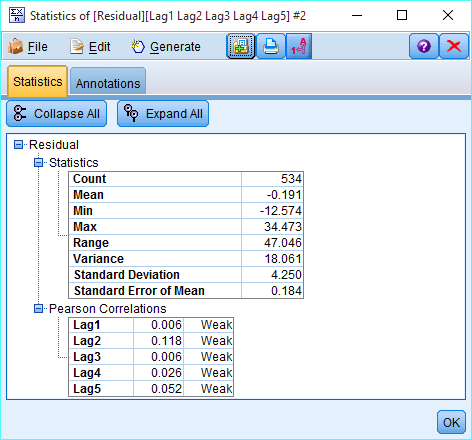


From the above graph, we can see that residuals are mostly centered around 0 as required by one of the assumption.





From the above graph, we can see that residuals are normally distributed as required by one of the assumption.



From the statistics, we can see that there is no strong correlation among the residuals so we can say that residuals are independent as required by one the assumptions.

1. Look at the table of Coefficients. Write your estimated regression. Which coefficients ARE helping to predict WAGE in this model?

**Answer:**

Wage(Predicted) = -.773+3.859\*Occupation(1)+(-1.766)\*Occupation(2)+(-.348)\*Occupation(3)+(-0.653)\*Occupation(4)+1.697\*Ocupation(5)+2.158\*Female(0)+(-1.695)\*Union(0)+0.858\*South(0) +(-0.894)\*Race(1)+.536\*Race(2)+(-0.557)\*Sector(0)+.338\*Sector(1) +(-0.474)\*marr(0)+0.866\*Education+0.341\*Experience+(-0.248)\*Age

Following coefficients are helping to predict Wage :

* 3.859\*Occupation(1)
* 2.158\*Female(0)
* (-1.695)\*Union(0)

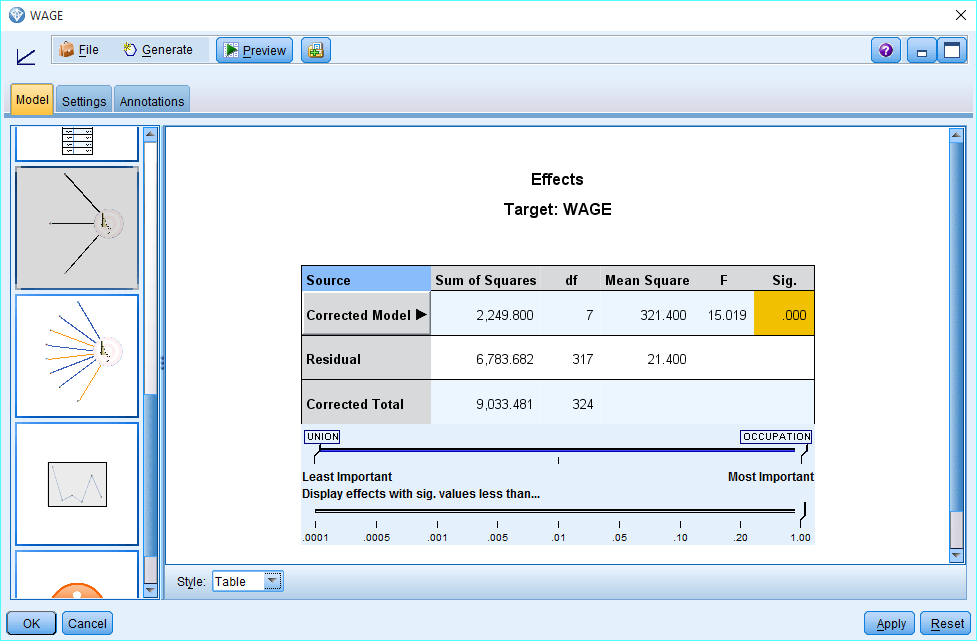
**PART 3 (Model Selection)**

1. Based on your p-values (Sig. values) from the previous step, one by one create new regression models each time removing the variable with the largest p-value. If any of your categorical values are good, keep that entire variable. Keep track of which variable you remove and the R2 for that model below. Keep each of these nodes in your stream.

|  |  |  |
| --- | --- | --- |
| Variables Removed | Variables In | R square |
| * AGE | * EDUCATION * SOUTH * FEMALE * EXPERIENCE * UNION * RACE * OCCUPATION * SECTOR * MARR | .30 |
| * AGE * SECTOR | * EDUCATION * SOUTH * FEMALE * EXPERIENCE * UNION * RACE * OCCUPATION * MARR | .301 |
| * EXPERIENCE * AGE * SECTOR | * EDUCATION * SOUTH * FEMALE * UNION * RACE * OCCUPATION * MARR | .266 |
| * EXPERIENCE * AGE * RACE * SECTOR | * EDUCATION * SOUTH * FEMALE * UNION * OCCUPATION * MARR | .269 |
| * EDUCATION * EXPERIENCE * AGE * RACE * SECTOR | * SOUTH * FEMALE * UNION * OCCUPATION * MARR | .244 |
| * EDUCATION * EXPERIENCE * AGE * RACE * SECTOR * MARR | * SOUTH * FEMALE * UNION * OCCUPATION | .240 |
| * EDUCATION * SOUTH * EXPERIENCE * AGE * RACE * SECTOR * MARR | * FEMALE * UNION * OCCUPATION | .232 |

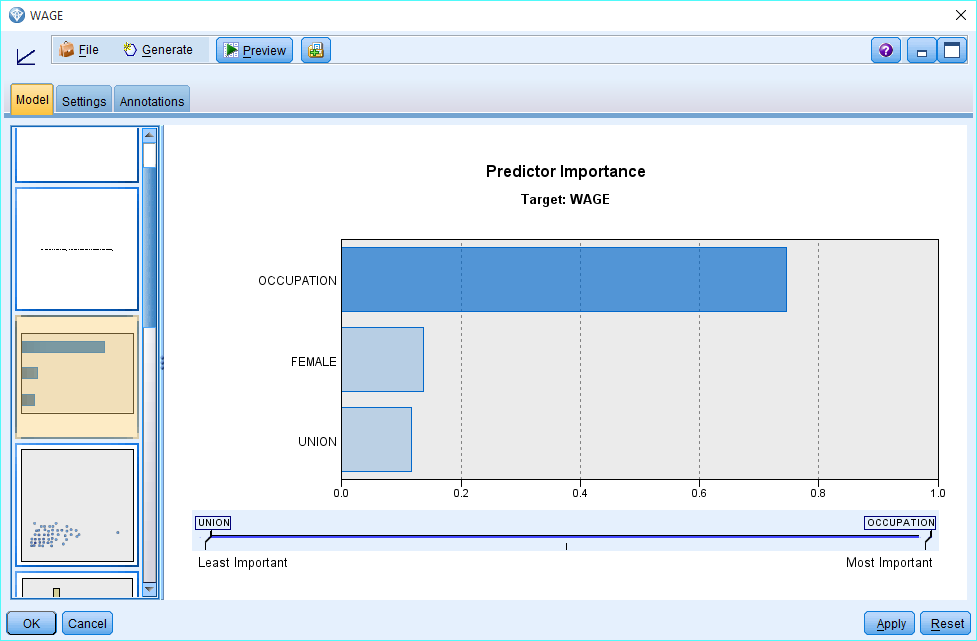
1. For your final model, what is your R2? Include a screenshot of your ANOVA table (Showing Sum of Squares values) and verify the calculation of the R2.

Adjusted R square for the final model is : .232

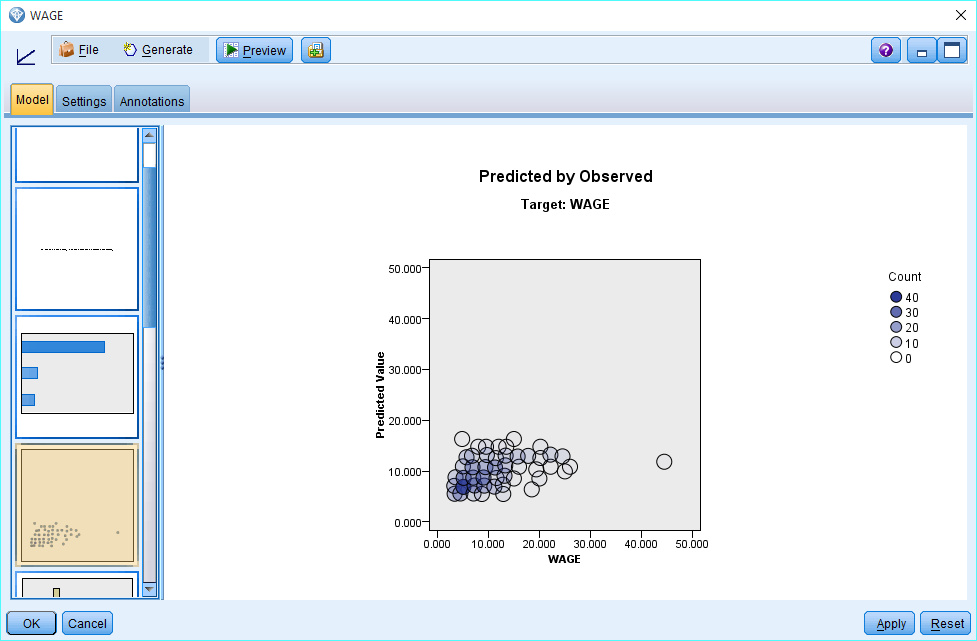


R-square= 2249.800/9033.480 = .249

1. Include a screenshot of your Predictor Importance Graph and the Predicted by Observed Scatterplot for the final model. Describe/interpret what the graphs show.



As per the graph, we can see that occupation variable has the most predicted importance.

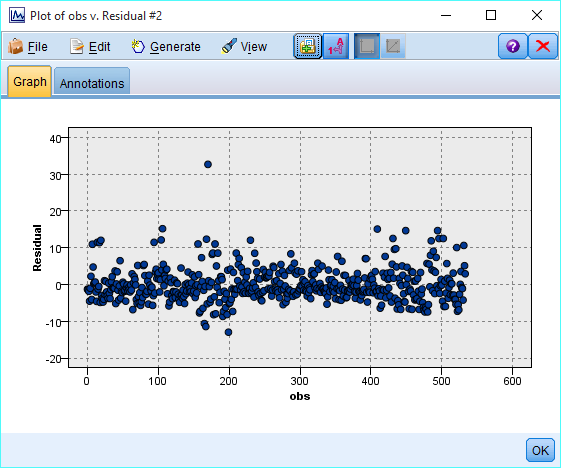


From the predicted, observed graph, we can see that our model is not predicting the wages very well. In a perfect situation, we would see all scatter plot circles on or around a 45 degree line.

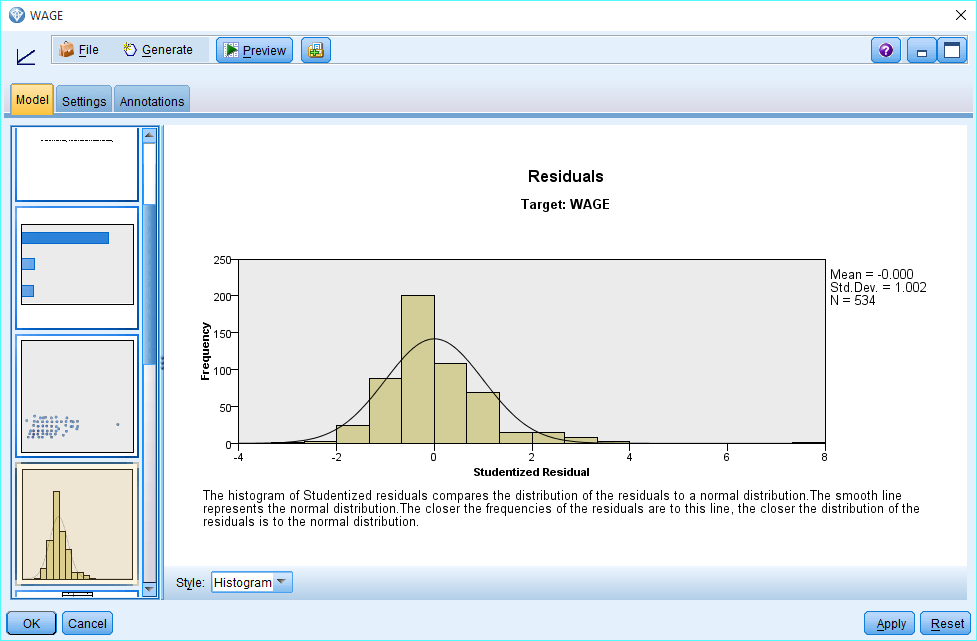
1. Check the 3 residual assumptions for regression using graphs and correlations as shown in the demo. (Note that you can copy the nodes from the Boston Housing Regression example which will be very helpful for the last 2 assumptions!)

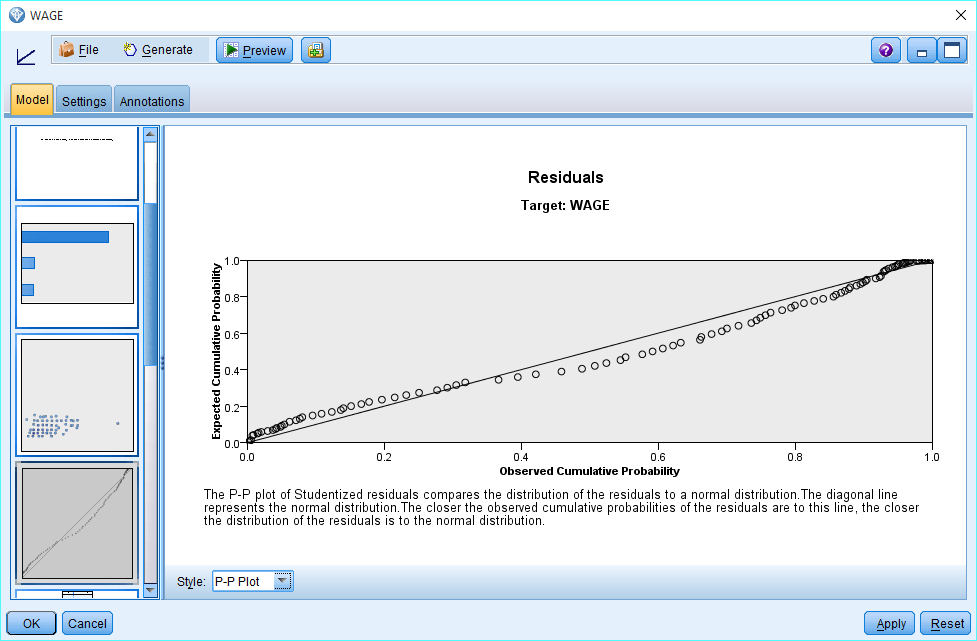
Include screenshots of each of these. Did anything change from **PART 2**?

* 1. Constant Variance
  2. Normality
  3. Residuals are independent (no autocorrelation)

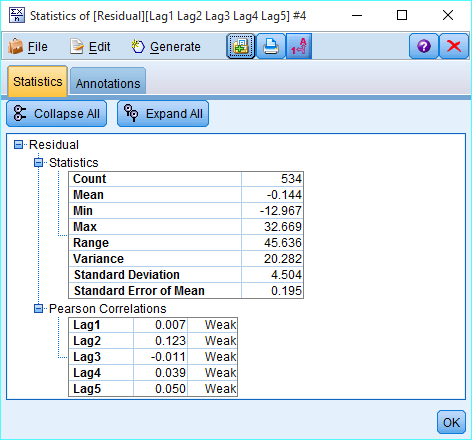


From the above graph, we can see that residuals are mostly centered around 0 as required by one of the assumption.





From the above graphs, we can see that residuals are normally distributed with mean around 0 as per the assumption.



From the statistics, we can see that there is no strong correlation among the residuals so we can say that residuals are independent as per the assumptions.

1. Look at the table of Coefficients. Make up a sample person (you make up) and predict their wage. Which coefficients ARE helping to predict WAGE in this model?

Answer:

Let’s consider a female with union membership and whose occupation is sales . Now, let’s use coefficients from our model to predict the wage for this female. As per our model, below is our equation for prediction:

Wage = 8.696+5.440\*occupation(1)+(-0.606)\*occupation(2)+.309\*Occupation(3)+(-.903)\*Occupation(4)+3.937\*Occupation(5)+2.136\*Female(0)+(-2.305)\*Union(0)

Wage = 8.696+(-0.606)\*2 =7.484

**Note:** In this example, we are considering only 3 variables of a record Female, union and occupation. Other variables are not part of our prediction equation so I have not included them for the purpose of simplicity.

**PART 4 (Forward and best subsets model selection)**

1. Create a linear regression model using all 10 variables to predict WAGE. Make sure to uncheck ‘Automatically prepare data’ and for Model selection choose ‘**Forward**.’ How does this list of variables differ from your final model in **PART 3**? What was the R2?

Answer:

R square is .305 and the most important predictor variables are occupation, age and Education. In the previous model we had removed education variable because of higher p value but in this model education variable is one of the most important predictor.

1. Do the same as above but choose ‘**best subsets**’ for Model selection. How does this list of variables differ from your final model in **PART 3** or the previous forward selection model above? What was the R2?

Answer:

R square Is .302. In this model, top 3 predictors are occupation , education and experience. In previous part, the top 3 predictors were occupation, age and Education. In part 3, the top 3 predictors were occupation , femal e and union.

Also, the R2 value is different in all three models.

1. Do you think there might be any interactions between variables that would help our prediction of WAGES?

From the correlation matrix, we see that Age and Experience variable have strong correlation. So these two variables seem to have some interaction with each other and some combination of these two variables can help improve the prediction.

**Submit your completed Word file and your stream file on Canvas. In the comments, put the names of the folks who worked on this assignment.**