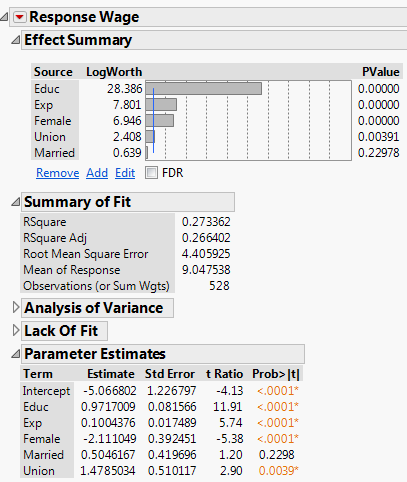
I)

1)

a)



The fitted equation is:

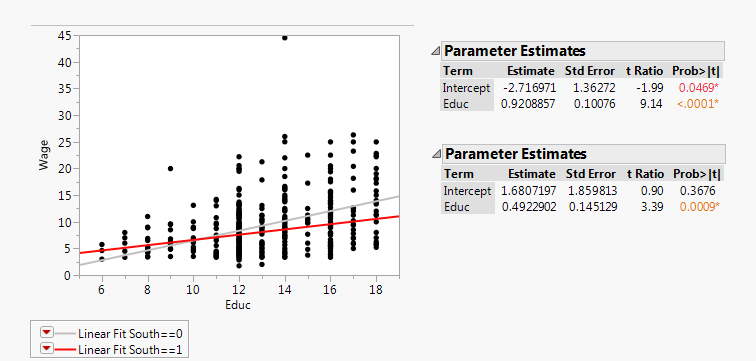
ANS

b)

Based on the model, women seem to significantly have less wage than men because the estimated coefficient of Female variable is significant and it is negative, implies a lower intercept term, and thus lower wage for female, holding other variables constant.

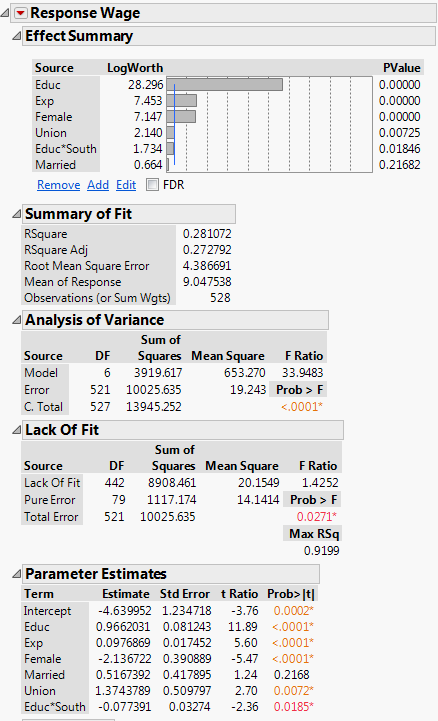
ANS

c)



The graph suggests there is an interaction effect between Education and Region (whether from the South or not). As Education increases, Wage of people in the South seems to increase at a slower rate. Educ in both group is significant so it is worth adding an interaction term Education \* South to the model. ANS

d)



The new fitted equation is:

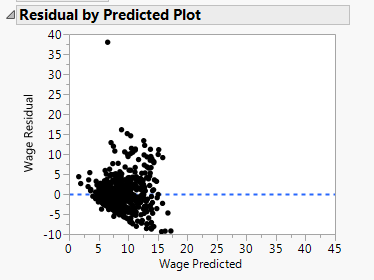
For South = 0:

For South = 1:

The model is improved a little bit as it has higher Adjusted R-Square and lower RMSE ANS

2)

a)

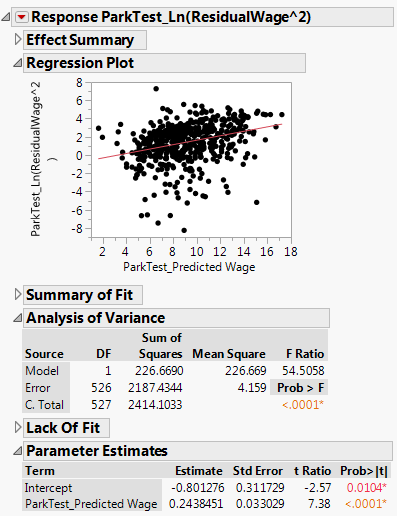


Yes, there might be a heteroscedasticity issue. ANS

b)

I use Park Test to test for heteroscedasticity, I collect the predicted Wage values and residuals and model a regression line between:

I got the following output:

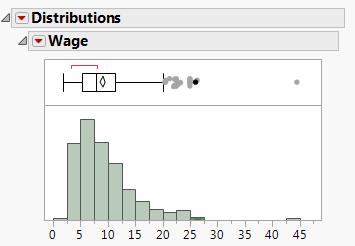


The coefficient of Predicted\_Wage is statistically significant so we can conclude that there is heteroscedasticity. ANS

c)

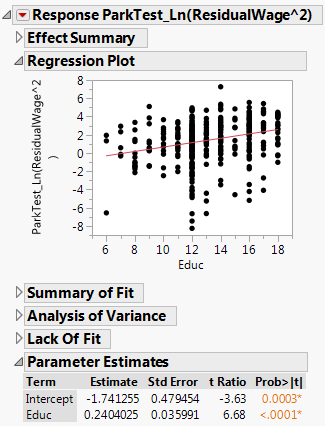
There are several possible sources:

1. Wage Variable



Heteroscedasticity may be caused the right skewed distribution of the predicted value, Wage. As shown in the histogram above.

1. Educ Variable

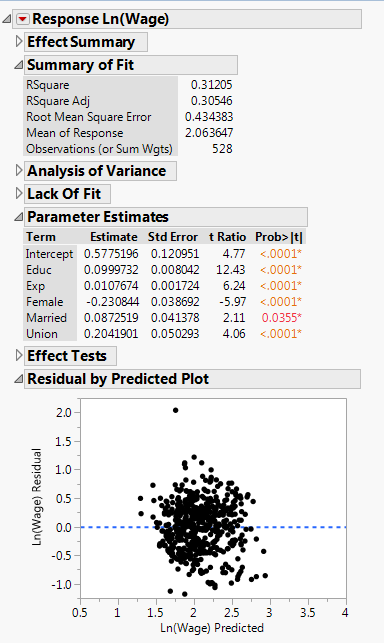


The model: shows that the coefficient of Educ is significant. So, it can be the source of heteroscedasticity.

ANS

d)

For Educ variable(in case c.2), when I plot Wage against Educ, I can’t see any curvature patterns there. So, I decided to consider a remedy for the case c.1, I take Log Transformation to Wage variable and make a new regression model. This is the final output:

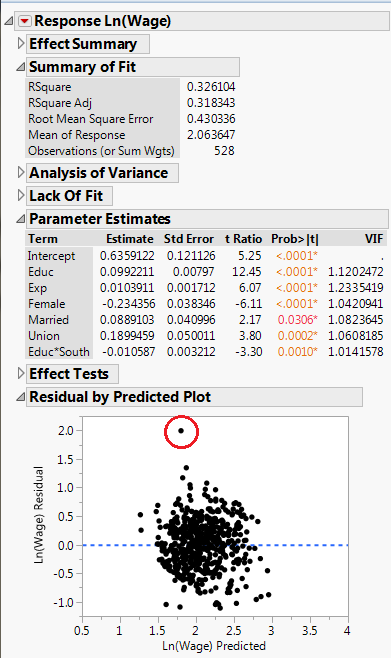


ANS

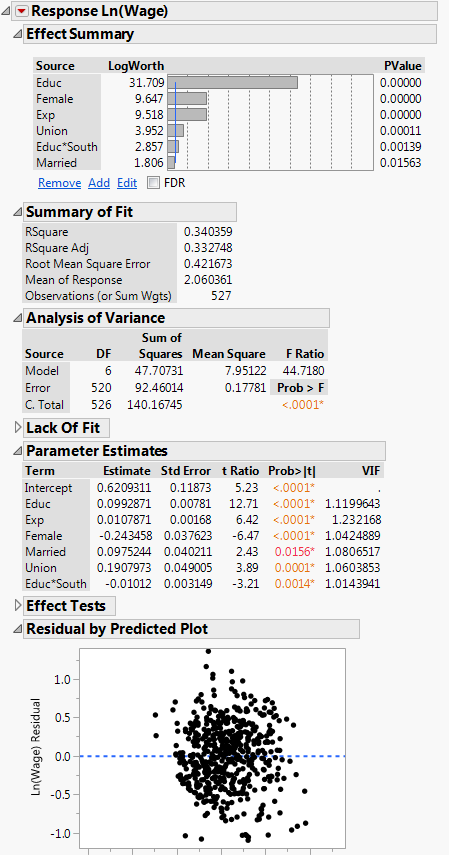
3)

To create a better model, I add the interaction term Educ\*South to the model in 2.d). I tried to find a curvurture between Ln(Wage) and independent variables but I don’t see any obvoius patterns. I also experiemented with other interaction terms but they seem to either worsen the performance or not significant. I add South variable to the model but this one is not significant as well.

Then I get the following model:



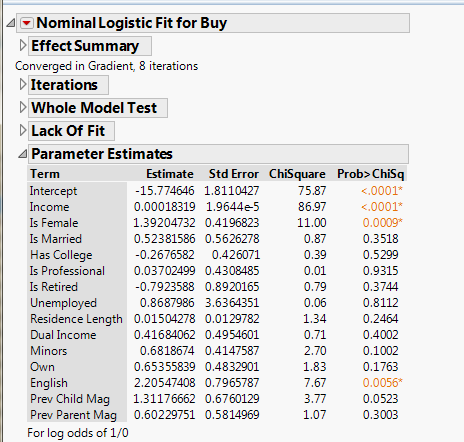
I then calculate Cook’s distance and remove one obvious outlier on the middle top(this one also has the highest Cook’s distance and exceeds 4/n criterior) and refit the model:



As this model has a better Adjusted R-Square and I don’t see any obvious outliers here, I decided it to be my final model. ANS

II)

a)



The regression equation is:

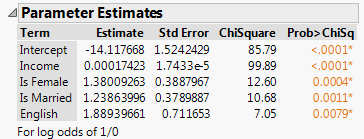
Where π is the probability of buying “Kid Creative”. ANS

b)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Removed Variable | -Loglilkelihood(Before removal) | -Loglilkelihood(After removal) | Change in -LogLikelihood | Compare to Threshold(3.84) |
| IsProfessional | 98.79 | 98.79 | 0 | 0 |
| Unemployed | 98.79 | 98.82 | -0.03 | 0.06 |
| HasCollege | 98.82 | 99.02 | -0.2 | 0.4 |
| DualIncome | 99.02 | 99.34 | -0.32 | 0.64 |
| ResidenceLength | 99.34 | 99.9 | -0.56 | 1.12 |
| IsRetired | 99.9 | 100.41 | -0.51 | 1.02 |
| PrevParentMag | 100.41 | 101.2 | -0.79 | 1.58 |
| Own | 101.2 | 102.18 | -0.98 | 1.96 |
| Minors | 102.18 | 103.67 | -1.49 | 2.98 |
| PrevChildMag | 103.67 | 105.33 | -1.66 | 3.32 |

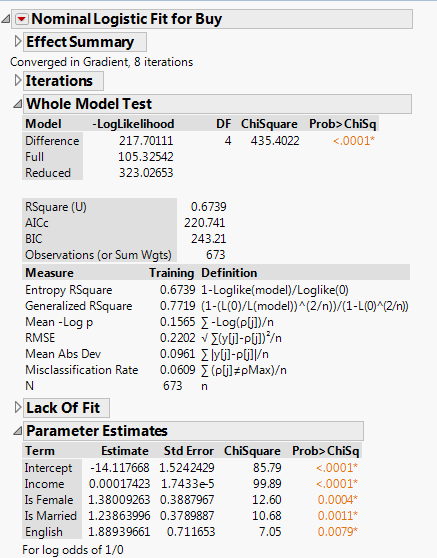
I removed each variable that has P-Value exceeds 0.05 significant level and the Loglikelihood error after removal doesn’t exceeds the 3.84 threshold criterion, as shown in the above table.

I stopped at English variable as the P-Value of this one doesn’t exceed 0.05 significant level.



ANS

c)



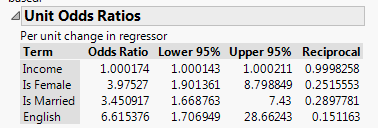
As shown in the picture above, the final model is:

Where π is the probability of buying “Kid Creative”. ANS

d)

Yes, at 0.05 significant level, the model is statistically significant overall. Based on P-Value, at the Whole Model Test dialog, we can reject the null hypothesis that all coefficients have 0 value and at Parameters Estimates dialog, we can reject the null hypothesis that each parameter has 0 value. ANS

e)



From the dialog above, being female increases the odd of buying “Kid Creative” by nearly four times. ANS

f)

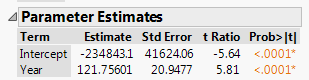
94 % ANS

g)

Probability(Purchase) = 0.0001923417 ANS

III)

a)

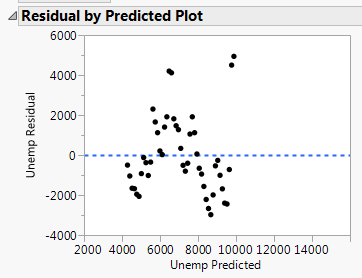


The fitted equation is:

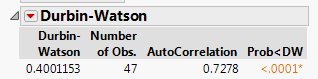
ANS

b)

Autocorrelation can be shown by the residual plot



It can be also shown by Durbin-Watson Statistics



With 0.05 significant level, we can reject the null hypothesis that there is no autocorrelation. Durbin-Watson Statistics is also close to 0 means there is a positive autocorrelation. ANS

c)

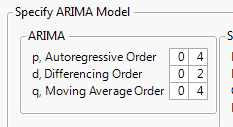
1. I tried encoding Year to start from 1 and add a quadratic term of Year to the model. Autocorrelation still presents.

2. I take a look at Time Series Modelling of Unemp and found that the Dicky-Fuller Test Statistics of all random walk models(Random Walk, Random Walk with Drift, Random Walk with Drift + Trend) indicate Unemp series is not stationary.

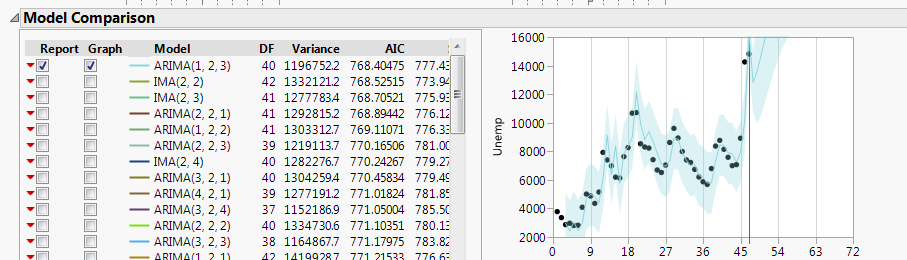
3. I repeat the step 2 again, by using a dependent variable Unemp(t) – Unemp(t-1). Now, all Test Statistics show that the series is stationary.

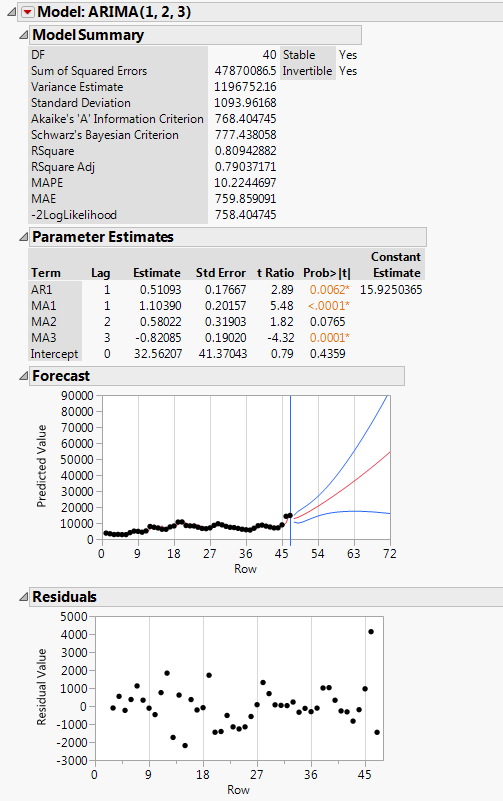
4. I look at PACF plot of 3. and noted that some lags at later intervals spike up again and nearly touch the significant threshold, so Integrated with order higher than 1 may be needed.

5. I go back to 2. and use ARIMA Model Group to let JMP discovers which model is the best predictor.



6. I get ARIMA(1,2,3) as the best prediction model, as shown below:





ANS

d)

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Prediction | Lower | Upper |
| 2011 | 12830.25 | 10686.12 | 14974.38 |
| 2012 | 13663.15 | 9961.98 | 17364.32 |

ANS