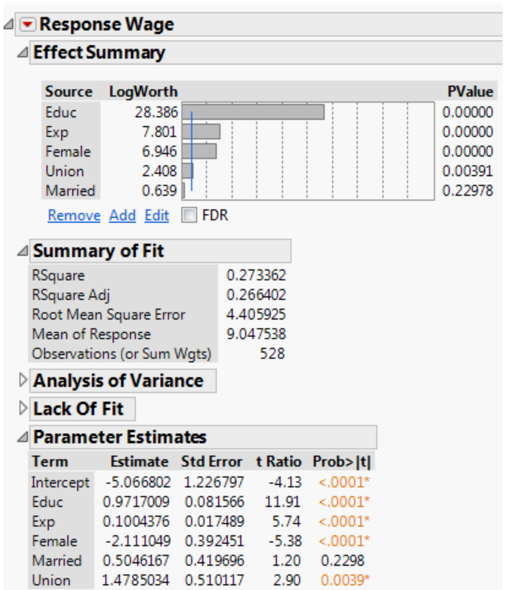
I)

1)

a)



Response Wage

Effect Summary

Source	LogWorth	PValue
Educ	28.386	0.00000
Exp	7.801	0.00000
Female	6.946	0.00000
Union	2.408	0.00391
Married	0.639	0.22978

Remove Add Edit FDR

Summary of Fit

RSquare	0.273362
RSquare Adj	0.266402
Root Mean Square Error	4.405925
Mean of Response	9.047538
Observations (or Sum Wgts)	528

Analysis of Variance

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-5.066802	1.226797	-4.13	<.0001*
Educ	0.9717009	0.081566	11.91	<.0001*
Exp	0.1004376	0.017489	5.74	<.0001*
Female	-2.111049	0.392451	-5.38	<.0001*
Married	0.5046167	0.419696	1.20	0.2298
Union	1.4785034	0.510117	2.90	0.0039*

The fitted equation is:

$$\text{Wage} = -5.07 + 0.97 * \text{Educ} + 0.1 * \text{Exp} - 2.11 * \text{Female} + 0.5 * \text{Married} + 1.48 * \text{Union}$$

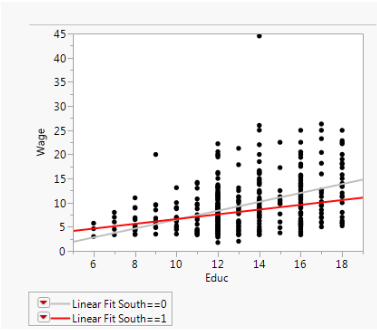
ANS

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)



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.716971	1.36272	-1.99	0.0469*
Educ	0.9208857	0.10076	9.14	<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.6807197	1.859813	0.90	0.3676
Educ	0.4922902	0.145129	3.39	0.0009*

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)

Response Wage

Effect Summary

Source	LogWorth	PValue
Educ	28.296	0.00000
Exp	7.453	0.00000
Female	7.147	0.00000
Union	2.140	0.00725
Educ*South	1.734	0.01846
Married	0.664	0.21682

Remove Add Edit FDR

Summary of Fit

RSquare	0.281072
RSquare Adj	0.272792
Root Mean Square Error	4.386691
Mean of Response	9.047538
Observations (or Sum Wgts)	528

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	3919.617	653.270	33.9483
Error	521	10025.635	19.243	Prob > F
C. Total	527	13945.252		<.0001*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	442	8908.461	20.1549	1.4252
Pure Error	79	1117.174	14.1414	Prob > F
Total Error	521	10025.635		0.0271*
				Max RSq
				0.9199

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-4.639952	1.234718	-3.76	0.0002*
Educ	0.9662031	0.081243	11.89	<.0001*
Exp	0.0976869	0.017452	5.60	<.0001*
Female	-2.136722	0.390889	-5.47	<.0001*
Married	0.5167392	0.417895	1.24	0.2168
Union	1.3743789	0.509797	2.70	0.0072*
Educ*South	-0.077391	0.03274	-2.36	0.0185*

The new fitted equation is:

For South = 0:

$$\text{Wage} = -4.64 + 0.97 * \text{Educ} + 0.1 * \text{Exp} - 2.14 * \text{Female} + 0.52 * \text{Married} + 1.37 * \text{Union}$$

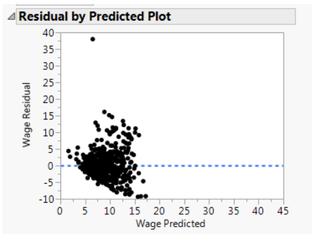
For South = 1:

$$\text{Wage} = -4.64 + (0.97 - 0.08) * \text{Educ} + 0.1 * \text{Exp} - 2.14 * \text{Female} + 0.52 * \text{Married} + 1.37 * \text{Union}$$

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. <u>ANS</u>

b)

I use Park Test to test for heteroscedasticity, I sorted the data with predicted Wages and divided them into 2 groups. I ran a regression and calculate MSE for each group.

The lower Wage group has MSE = 13.66

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	237.4692	47.4938	3.4751
Error	258	3526.0950	13.6670	Prob > F
C. Total	263	3763.5641		0.0047*

The higher Wage group has MSE = 25.20

Observations (or Sum Wgts) 264

Analysis of Variance

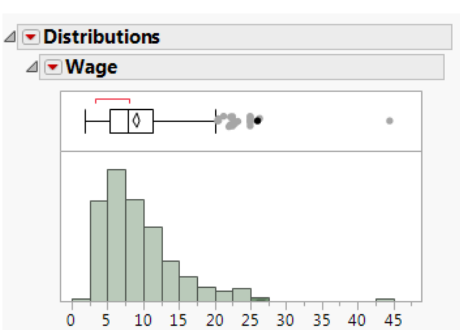
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	1000.4971	200.099	7.9382
Error	258	6503.4189	25.207	Prob > F
C. Total	263	7503.9159		<.0001*

F-Statistics = 25.20/13.66 = 1.84

According to the F-table, the cutoff F-Statistics at df = (263,263) is around 1.00.

So, we reject the null Hypothesis and conclude that there is heteroscedasticity. <u>ANS</u>

c)



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

d)

To remedy this, I take Log Transformation to Wage variable and make a new regression model. This is the final output:

Response Ln(Wage)

Effect Summary

Summary of Fit

RSquare	0.31205
RSquare Adj	0.30546
Root Mean Square Error	0.434383
Mean of Response	2.063647
Observations (or Sum Wgts)	528

Analysis of Variance

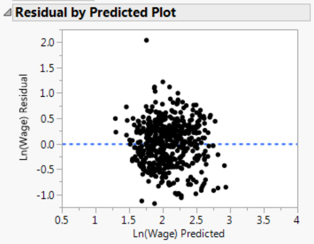
Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.5775196	0.120951	4.77	<.0001*
Educ	0.0999732	0.008042	12.43	<.0001*
Exp	0.0107674	0.001724	6.24	<.0001*
Female	-0.230844	0.038692	-5.97	<.0001*
Married	0.0872519	0.041378	2.11	0.0355*
Union	0.2041901	0.050293	4.06	<.0001*

Effect Tests

Residual by Predicted Plot

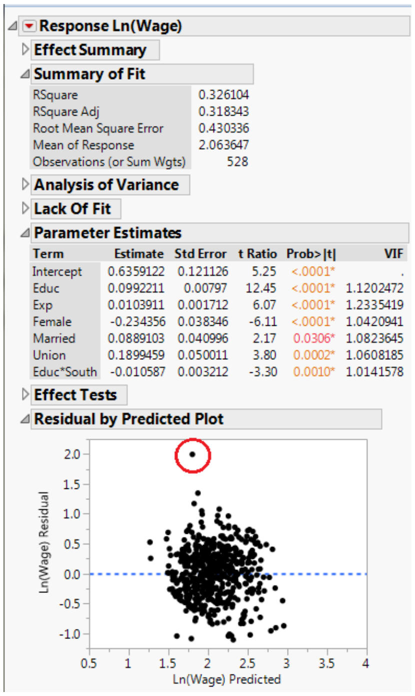


ANS

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:



Response Ln(Wage)

Effect Summary

Summary of Fit

RSquare	0.326104
RSquare Adj	0.318343
Root Mean Square Error	0.430336
Mean of Response	2.063647
Observations (or Sum Wgts)	528

Analysis of Variance

Lack Of Fit

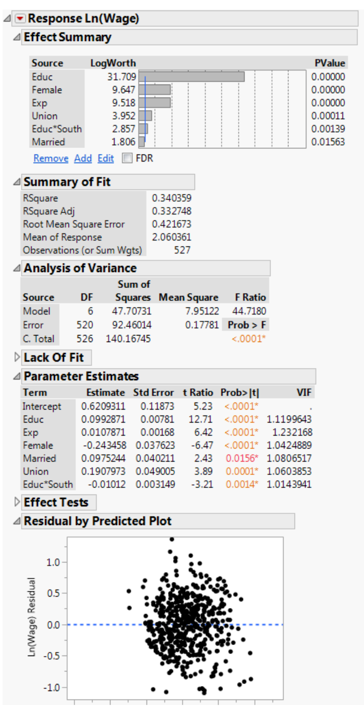
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.6359122	0.121126	5.25	<.0001*	.
Educ	0.0992211	0.00797	12.45	<.0001*	1.1202472
Exp	0.0103911	0.001712	6.07	<.0001*	1.2335419
Female	-0.234356	0.038346	-6.11	<.0001*	1.0420941
Married	0.0889103	0.040996	2.17	0.0306*	1.0823645
Union	0.1899459	0.050011	3.80	0.0002*	1.0608185
Educ*South	-0.010587	0.003212	-3.30	0.0010*	1.0141578

Effect Tests

Residual by Predicted Plot

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:



Response Ln(Wage)

Effect Summary

Source	LogWorth	PValue
Educ	31.709	0.00000
Female	9.647	0.00000
Exp	9.518	0.00000
Union	3.952	0.00011
Educ*South	2.857	0.00139
Married	1.806	0.01563

Remove Add Edit FDR

Summary of Fit

RSquare	0.340359
RSquare Adj	0.332748
Root Mean Square Error	0.421673
Mean of Response	2.060361
Observations (or Sum Wgts)	527

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	47.70731	7.95122	44.7180
Error	520	92.46014	0.17781	Prob > F
C. Total	526	140.16745		<.0001*

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.6209311	0.11873	5.23	<.0001*	.
Educ	0.0992871	0.00781	12.71	<.0001*	1.1199643
Exp	0.0107871	0.00168	6.42	<.0001*	1.232168
Female	-0.243458	0.037623	-6.47	<.0001*	1.0424889
Married	0.0975244	0.040211	2.43	0.0156*	1.0806517
Union	0.1907973	0.049005	3.89	0.0001*	1.0603853
Educ*South	-0.01012	0.003149	-3.21	0.0014*	1.0143941

Effect Tests

Residual by Predicted Plot

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)

Nominal Logistic Fit for Buy

Effect Summary

Converged in Gradient, 8 iterations

Iterations

Whole Model Test

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-15.774646	1.8110427	75.87	<.0001*
Income	0.00018319	1.9644e-5	86.97	<.0001*
Is Female	1.39204732	0.4196823	11.00	0.0009*
Is Married	0.52381586	0.5626278	0.87	0.3518
Has College	-0.2676582	0.426071	0.39	0.5299
Is Professional	0.03702499	0.4308485	0.01	0.9315
Is Retired	-0.7923588	0.8920165	0.79	0.3744
Unemployed	0.8687986	3.6364351	0.06	0.8112
Residence Length	0.01504278	0.0129782	1.34	0.2464
Dual Income	0.41684062	0.4954601	0.71	0.4002
Minors	0.6818674	0.4147587	2.70	0.1002
Own	0.65355839	0.4832901	1.83	0.1763
English	2.20547408	0.7965787	7.67	0.0056*
Prev Child Mag	1.31176662	0.6760129	3.77	0.0523
Prev Parent Mag	0.60229751	0.5814969	1.07	0.3003

For log odds of 1/0

The regression equation is:

$$\log\frac{\pi}{1-\pi} = -15.78 + 0.0002 * Income + 1.39 * IsFemale + 0.52 * IsMArried - 0.27 * HasCollege + 0.04 * IsProfessional - 0.79 * IsRetired + 0.87 * Unemployed + 0.02 * ResidenceLength + 0.42 * DualIncome + 0.68 * Minors + 0.65 * Own + 2.21 * English + 1.31 * PrevChildMag + 0.6 * PrevParentMag$$

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.

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-14.117668	1.5242429	85.79	<.0001*
Income	0.00017423	1.7433e-5	99.89	<.0001*
Is Female	1.38009263	0.3887967	12.60	0.0004*
Is Married	1.23863996	0.3789887	10.68	0.0011*
English	1.88939661	0.711653	7.05	0.0079*

For log odds of 1/0

ANS

c)

Nominal Logistic Fit for Buy

Effect Summary

Converged in Gradient, 8 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	217.70111	4	435.4022	<.0001*
Full	105.32542			
Reduced	323.02653			

RSquare (U)	0.6739
AICc	220.741
BIC	243.21
Observations (or Sum Wgts)	673

Measure	Training	Definition
Entropy RSquare	0.6739	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7719	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1565	∑ -Log(ρ[j])/n
RMSE	0.2202	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0961	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0609	∑ (ρ[j]≠ρMax)/n
N	673	n

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-14.117668	1.5242429	85.79	<.0001*
Income	0.00017423	1.7433e-5	99.89	<.0001*
Is Female	1.38009263	0.3887967	12.60	0.0004*
Is Married	1.23863996	0.3789887	10.68	0.0011*
English	1.88939661	0.711653	7.05	0.0079*

For log odds of 1/0

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

$$\log\frac{\pi}{1-\pi} = -14.12 + 0.0002 * Income + 1.38 * IsFemale + 1.24 * IsMarried + 1.89 * English$$

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 parameters have 0 value and at Parameters Estimates dialog, we can reject the null hypothesis that each parameter has 0 value. ANS

e)

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Income	1.000174	1.000143	1.000211	0.9998258
Is Female	3.97527	1.901361	8.798849	0.2515553
Is Married	3.450917	1.668763	7.43	0.2897781
English	6.615376	1.706949	28.66243	0.151163

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)

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-234843.1	41624.06	-5.64	<.0001*
Year	121.75601	20.9477	5.81	<.0001*

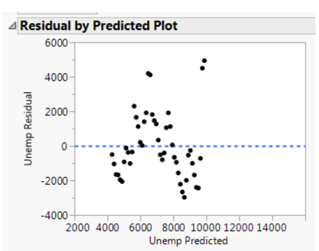
The fitted equation is:

$$Unemp = -234843.1 + 121.75601 * Year$$

ANS

b)

Autocorrelation can be shown by the residual plot



It can be also shown by Durbin-Watson Statistics

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
0.4001153	47	0.7278	<.0001*

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)

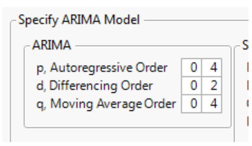
1. I tried encoding Year to start from 1 and add a quadratic term of Year to the model. Autocorrelation still presents.

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.

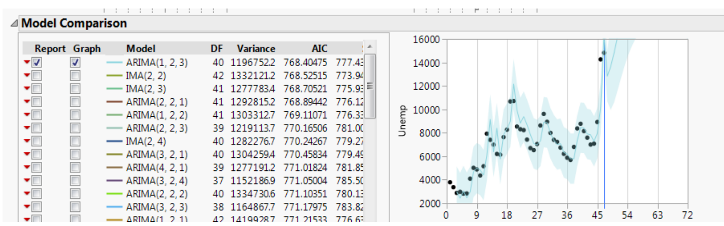
3. I repeat the step 2 again, by using a dependent variable Unemp(t) – Unemp(t-1). Now, all Test Statistics that the series is stationary.

4. I look at PACF plot of 3. and noted that some lags at later interval 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.

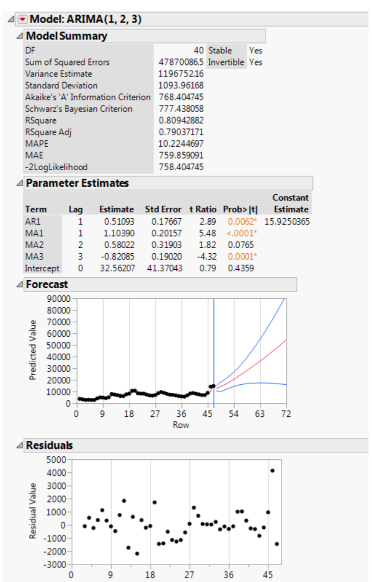


6. I get ARIMA(1,2,3) as the best prediction, as shown below:



Model Comparison

Report	Graph	Model	DF	Variance	AIC	
☑	☑	ARIMA(1, 2, 3)	40	1196752.2	768.40475	777.43
☐	☐	IMA(2, 2)	42	1332121.2	768.52515	773.94
☐	☐	IMA(2, 3)	41	1277783.4	768.70521	775.93
☐	☐	ARIMA(2, 2, 1)	41	1292815.2	768.89442	776.12
☐	☐	ARIMA(1, 2, 2)	41	1303312.7	769.11071	776.33
☐	☐	ARIMA(2, 2, 3)	39	1219113.7	770.16506	781.00
☐	☐	IMA(2, 4)	40	1282276.7	770.24267	779.27
☐	☐	ARIMA(3, 2, 1)	40	1304259.4	770.45834	779.49
☐	☐	ARIMA(4, 2, 1)	39	1277191.2	771.01824	781.85
☐	☐	ARIMA(3, 2, 4)	37	1152186.9	771.05004	785.50
☐	☐	ARIMA(2, 2, 2)	40	1334730.6	771.10351	780.13
☐	☐	ARIMA(3, 2, 3)	38	1164867.7	771.17975	783.82
☐	☐	ARIMA(1, 2, 1)	42	1419928.7	771.21533	776.6



Model: ARIMA(1, 2, 3)

Model Summary

DF	40	Stable	Yes
Sum of Squared Errors	47870086.5	Invertible	Yes
Variance Estimate	119675216		
Standard Deviation	1093.96168		
Akaike's 'A' Information Criterion	768.404745		
Schwarz's Bayesian Criterion	777.438058		
RSquare	0.80942882		
RSquare Adj	0.79037171		
MAPE	10.2244697		
MAE	759.859091		
-2LogLikelihood	758.404745		

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate
AR1	1	0.51093	0.17667	2.89	0.0062*	15.9250365
MA1	1	1.10390	0.20157	5.48	<.0001*	
MA2	2	0.58022	0.31903	1.82	0.0765	
MA3	3	-0.82085	0.19020	-4.32	0.0001*	
Intercept	0	32.56207	41.37043	0.79	0.4359	

<u>ANS</u>

d)

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

ANS