CASE PROJECT – FINAL ASSIGMENT

LETTER A

Consider a linear model where the sale price of a house is the dependent variable and the explanatory variables are the other variables given above. Perform a test for linearity. What do you conclude based on the test result?

Regressing using the dataset provided, we obtain:

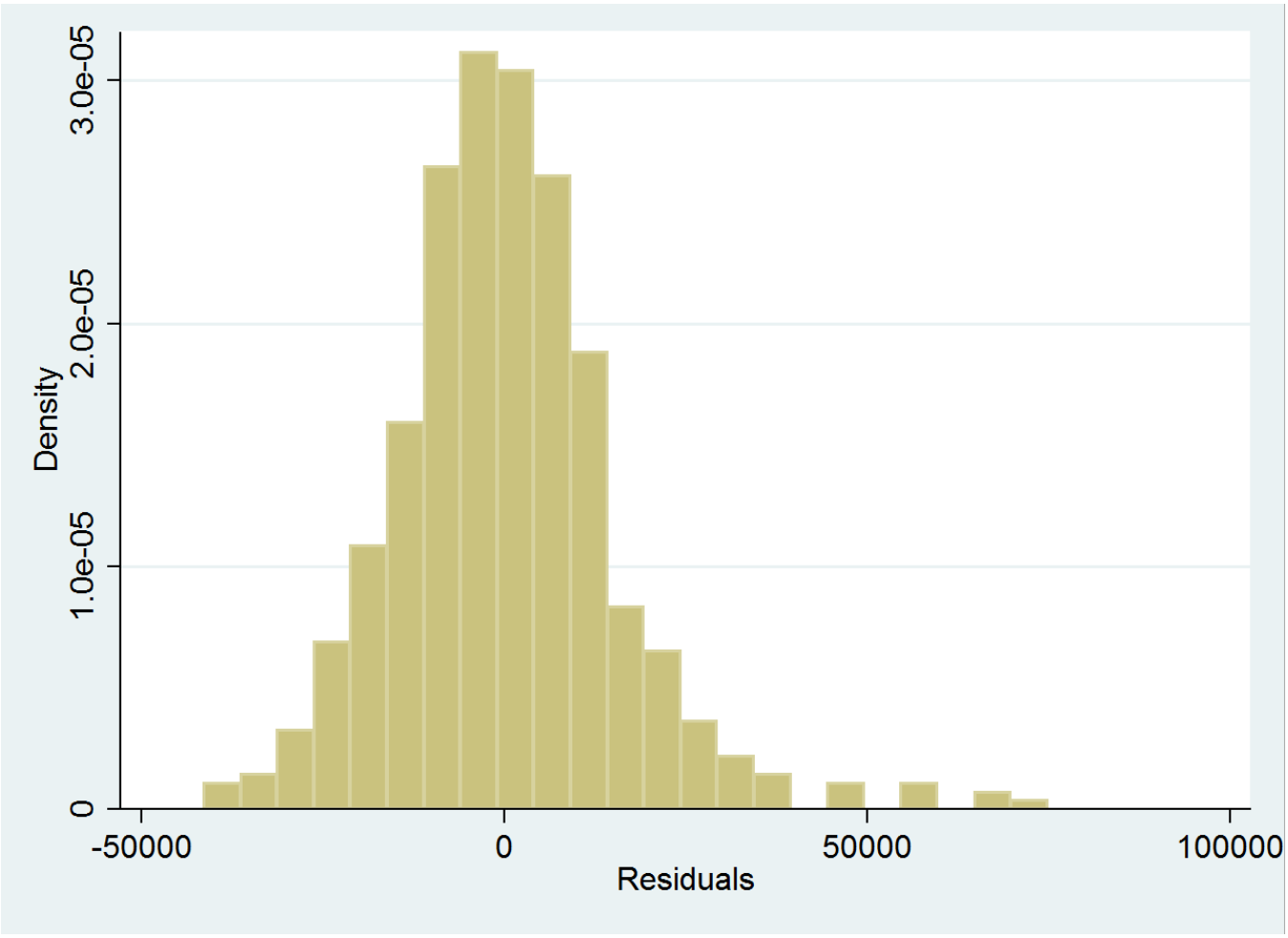
```
. regress sell lot bdms fb sty drv rec ffin ghw ca gar reg
```

Source	SS	df	MS	Number of obs =	546
				F(11, 534) =	99.97
Model	2.6158e+11	11	2.3780e+10	Prob > F =	0.0000
Residual	1.2703e+11	534	237874666	R-squared =	0.6731
				Adj R-squared =	0.6664
Total	3.8860e+11	545	713032635	Root MSE =	15423

sell	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lot	3.546303	.3503	10.12	0.000	2.858168	4.234438
bdms	1832.003	1047	1.75	0.081	-224.7409	3888.748
fb	14335.56	1489.921	9.62	0.000	11408.73	17262.38
sty	6556.946	925.2899	7.09	0.000	4739.291	8374.6
drv	6687.779	2045.246	3.27	0.001	2670.065	10705.49
rec	4511.284	1899.958	2.37	0.018	778.9759	8243.592
ffin	5452.386	1588.024	3.43	0.001	2332.845	8571.926
ghw	12831.41	3217.597	3.99	0.000	6510.706	19152.11
ca	12632.89	1555.021	8.12	0.000	9578.182	15687.6
gar	4244.829	840.5442	5.05	0.000	2593.65	5896.008
reg	9369.513	1669.091	5.61	0.000	6090.724	12648.3
_cons	-4038.35	3409.471	-1.18	0.237	-10735.97	2659.271

At the 5% level significance, we can see that the number of bedrooms and the inclusion of a recreational room are not statistically significant in defining the sale price of a house.

Plotting the residuals of such regression, we obtain:



Also, performing a Jarque-Bera test with the residuals, we obtain:

```
. sktest residuals
```

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
residuals	546	0.0000	0.0000	72.09	0.0000

We therefore reject the null hypothesis of normality of the residuals.

LETTER B

Now consider a linear model where the log of the sale price of the house is the dependent variable and the explanatory variables are as before. Perform again the test for linearity. What do you conclude now?

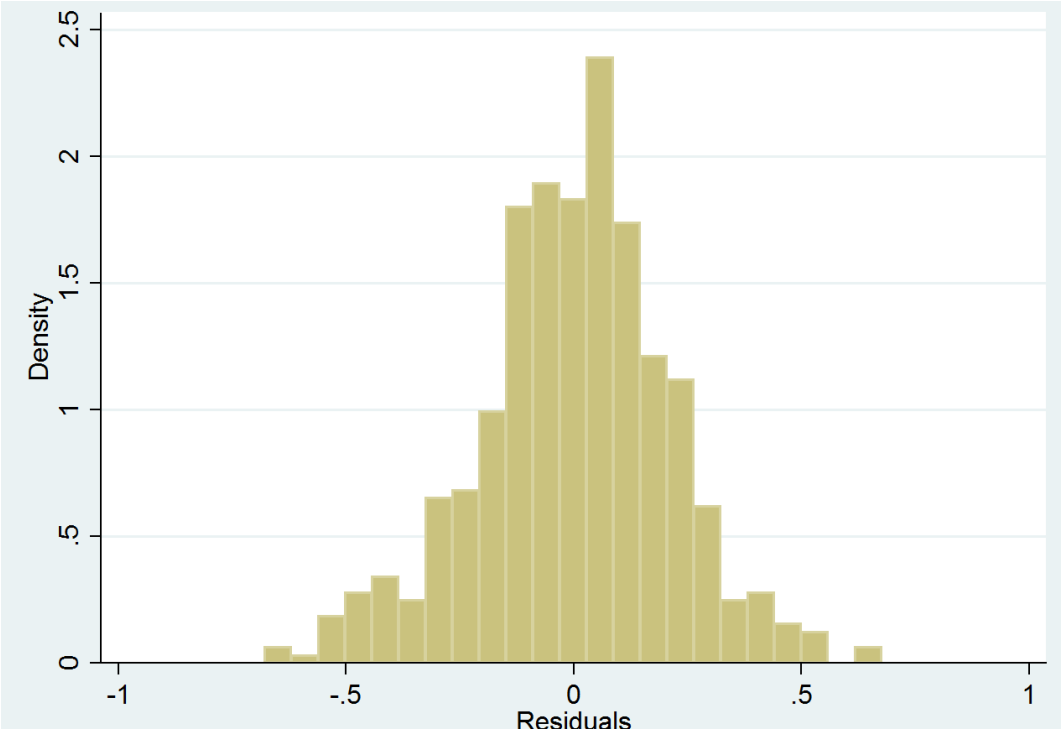
Performing the regression stated above, we obtain:

```
. regress logsell lot bdms fb sty drv rec ffin ghw ca gar reg
```

Source	SS	df	MS		
Model	51.0238728	11	4.63853389	Number of obs	= 546
Residual	24.3892974	534	.045672842	F(11, 534)	= 101.56
				Prob > F	= 0.0000
Total	75.4131702	545	.138372789	R-squared	= 0.6766
				Adj R-squared	= 0.6699
				Root MSE	= .21371

logsell	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lot	.0000506	4.85e-06	10.42	0.000	.000041	.0000601
bdms	.0340205	.0145078	2.34	0.019	.0055211	.0625199
fb	.1677687	.0206452	8.13	0.000	.127213	.2083244
sty	.0922745	.0128213	7.20	0.000	.0670881	.1174609
drv	.1306512	.02834	4.61	0.000	.0749796	.1863229
rec	.0735165	.0263268	2.79	0.005	.0217996	.1252334
ffin	.0993997	.0220045	4.52	0.000	.0561737	.1426257
ghw	.1783544	.0445848	4.00	0.000	.0907714	.2659375
ca	.1780197	.0215472	8.26	0.000	.135692	.2203474
gar	.0507568	.011647	4.36	0.000	.0278772	.0736365
reg	.1271134	.0231278	5.50	0.000	.0816807	.172546
_cons	10.02556	.0472435	212.21	0.000	9.93275	10.11836

The residuals of such regression in a histogram are as follows:



Performing the Jarque-Bera test, we obtain:

```
. sktest residualslog
```

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
residualslog	546	0.0567	0.0458	7.41	0.0246

We still reject the null hypothesis of normality.

LETTER C

Regressing the linear model including both the variable lot size and log(lot size), we obtain the following results:

```
. regress logsell loglot lot bdms fb sty drv rec ffin ghw ca gar reg
```

Source	SS	df	MS		
Model	51.8121891	12	4.31768243	Number of obs	546
Residual	23.6009811	533	.044279514	F(12, 533)	97.51
				Prob > F	0.0000
Total	75.4131702	545	.138372789	R-squared	0.6870
				Adj R-squared	0.6800
				Root MSE	.21043

logsell	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
loglot	.3826883	.0906977	4.22	0.000	.2045195	.5608571
lot	-.0000149	.0000162	-0.92	0.359	-.0000468	.000017
bdms	.0348924	.0142863	2.44	0.015	.0068281	.0629568
fb	.1659386	.0203324	8.16	0.000	.125997	.2058801
sty	.091213	.0126267	7.22	0.000	.0664087	.1160173
drv	.1068137	.0284706	3.75	0.000	.0508854	.162742
rec	.0546692	.0263042	2.08	0.038	.0029966	.1063418
ffin	.1052473	.0217106	4.85	0.000	.0625985	.1478961
ghw	.1790865	.0438998	4.08	0.000	.0928487	.2653243
ca	.1643383	.0214624	7.66	0.000	.1221771	.2064994
gar	.0482648	.0114832	4.20	0.000	.0257069	.0708226
reg	.1343625	.022837	5.88	0.000	.0895008	.1792241
_cons	7.150477	.682984	10.47	0.000	5.808806	8.492148

We can observe that we have two different p-values for both log of the lot size and the lot size itself:

$$p\ value\ (log\ of\ lot\ size) = 0.000$$

$$p\ value\ (lot\ size) = 0.359$$

Therefore, it is best to use the log of lot size instead of the untransformed variable.

LETTER D

To test whether the variables have significant effects or not, we defined the interaction variable as the variable log(lotsize) times the other explanatory variables we would like to test. Therefore, the result was as it follows:

Source	SS	df	MS
Model	52.4204317	21	2.49621103
Residual	22.9927385	524	.043879272
Total	75.4131702	545	.138372789

Number of obs	= 546
F(21, 524)	= 56.89
Prob > F	= 0.0000
R-squared	= 0.6951
Adj R-squared	= 0.6829
Root MSE	= .20947

logsell	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
loglot	.1526855	.1282938	1.19	0.235	-.0993479	.4047189
bdms	.019076	.3266998	0.06	0.953	-.6227263	.6608783
fb	-.3682336	.4290476	-0.86	0.391	-1.211098	.474631
sty	.4888853	.3096996	1.58	0.115	-.1195199	1.097291
drv	-1.46337	.7172247	-2.04	0.042	-2.872359	-.054381
rec	1.673993	.6559188	2.55	0.011	.3854396	2.962546
ffin	-.0318449	.4455434	-0.07	0.943	-.9071156	.8434259
ghw	-.5058876	.9027332	-0.56	0.575	-2.279308	1.267533
ca	-.3402737	.4960409	-0.69	0.493	-1.314747	.6341993
gar	.4019411	.2586464	1.55	0.121	-.1061703	.9100524
reg	.1184847	.479856	0.25	0.805	-.8241931	1.061162
loglotbdms	.0020695	.0386538	0.05	0.957	-.073866	.078005
loglotfb	.0620367	.0501454	1.24	0.217	-.036474	.1605474
loglotsty	-.0463612	.0359416	-1.29	0.198	-.1169685	.0242461
loglotdrv	.1915418	.0873606	2.19	0.029	.0199218	.3631618
loglotrec	-.1884625	.0763734	-2.47	0.014	-.3384982	-.0384267
loglotffin	.0159131	.0528514	0.30	0.763	-.0879135	.1197398
loglotghw	.0811352	.1069291	0.76	0.448	-.1289273	.2911976
loglotca	.0595486	.0580237	1.03	0.305	-.0544391	.1735362
loglotgar	-.0413586	.0301417	-1.37	0.171	-.1005721	.0178548
loglotreg	.0015151	.0559897	0.03	0.978	-.1084767	.1115069
_cons	8.966495	1.070667	8.37	0.000	6.863168	11.06982

We can observe that, at the 5% level, only the interaction of log(lotsize) with the dummy variable related to the driveway and the dummy variable related to the presence or not of a recreational room were significant.

LETTER E

Performing the F test, we obtain:

$$F = \frac{23.64 - 22.99}{10} * \frac{22 - 10}{22.9} = 1.47$$

We have $n = 546$ observations. Therefore, at the 1% level, the critical value of F is:

$$FCrit = 1.83$$

Therefore, we reject the hypothesis of joint significance of interaction effects.

LETTER F

By performing the general-to-specific approach model specification, we obtain a step-by-step of:

Source	SS	df	MS		
Model	52.4204317	21	2.49621103	Number of obs	= 546
Residual	22.9927385	524	.043879272	F(21, 524)	= 56.89
				Prob > F	= 0.0000
Total	75.4131702	545	.138372789	R-squared	= 0.6951
				Adj R-squared	= 0.6829
				Root MSE	= .20947

logsell	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
loglot	.1526855	.1282938	1.19	0.235	-.0993479	.4047189
bdms	.019076	.3266998	0.06	0.953	-.6227263	.6608783
fb	-.3682336	.4290476	-0.86	0.391	-1.211098	.474631
sty	.4888853	.3096996	1.58	0.115	-.1195199	1.097291
drv	-1.46337	.7172247	-2.04	0.042	-2.872359	-.054381
rec	1.673993	.6559188	2.55	0.011	.3854396	2.962546
ffin	-.0318449	.4455434	-0.07	0.943	-.9071156	.8434259
ghw	-.5058876	.9027332	-0.56	0.575	-2.279308	1.267533
ca	-.3402737	.4960409	-0.69	0.493	-1.314747	.6341993
gar	.4019411	.2586464	1.55	0.121	-.1061703	.9100524
reg	.1184847	.479856	0.25	0.805	-.8241931	1.061162
loglotbdms	.0020695	.0386538	0.05	0.957	-.073866	.078005
loglotfb	.0620367	.0501454	1.24	0.217	-.036474	.1605474
loglotsty	-.0463612	.0359416	-1.29	0.198	-.1169685	.0242461
loglotdrv	.1915418	.0873606	2.19	0.029	.0199218	.3631618
loglotrec	-.1884625	.0763734	-2.47	0.014	-.3384982	-.0384267
loglotffin	.0159131	.0528514	0.30	0.763	-.0879135	.1197398
loglotghw	.0811352	.1069291	0.76	0.448	-.1289273	.2911976
loglotca	.0595486	.0580237	1.03	0.305	-.0544391	.1735362
loglotgar	-.0413586	.0301417	-1.37	0.171	-.1005721	.0178548
loglotreg	.0015151	.0559897	0.03	0.978	-.1084767	.1115069
_cons	8.966495	1.070667	8.37	0.000	6.863168	11.06982

Source	SS	df	MS	Number of obs =	546
				F(13, 532) =	91.79
Model	52.1584308	13	4.01218698	Prob > F =	0.0000
Residual	23.2547394	532	.043711916	R-squared =	0.6916
				Adj R-squared =	0.6841
Total	75.4131702	545	.138372789	Root MSE =	.20907

logsell	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
loglot	.1790575	.0770663	2.32	0.021	.0276658	.3304491
bdms	.0388127	.0143012	2.71	0.007	.010719	.0669064
fb	.161451	.0202535	7.97	0.000	.1216643	.2012377
sty	.0908278	.0125416	7.24	0.000	.0661907	.1154648
drv	-1.189961	.6646176	-1.79	0.074	-2.495558	.1156359
rec	1.502532	.6255276	2.40	0.017	.2737249	2.731339
ffin	.1027617	.0215735	4.76	0.000	.0603821	.1451414
ghw	.1844812	.043682	4.22	0.000	.0986708	.2702916
ca	.16526	.0212085	7.79	0.000	.1235974	.2069226
gar	.0469019	.0114209	4.11	0.000	.0244663	.0693375
reg	.1326028	.02255	5.88	0.000	.0883049	.1769007
loglotdrv	.15943	.0812426	1.96	0.050	-.0001657	.3190257
loglotrec	-.1682589	.0727042	-2.31	0.021	-.3110815	-.0254363
_cons	8.741889	.6286296	13.91	0.000	7.506988	9.97679

Source	SS	df	MS	Number of obs =	546
				F(12, 533) =	98.59
Model	51.9900967	12	4.33250806	Prob > F =	0.0000
Residual	23.4230735	533	.043945729	R-squared =	0.6894
				Adj R-squared =	0.6824
Total	75.4131702	545	.138372789	Root MSE =	.20963

logsell	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
loglot	.3202423	.0276981	11.56	0.000	.2658316	.3746531
bdms	.0384222	.014338	2.68	0.008	.0102563	.0665881
fb	.1631818	.0202884	8.04	0.000	.1233269	.2030368
sty	.0907958	.012575	7.22	0.000	.0660931	.1154985
drv	.1131157	.0281545	4.02	0.000	.0578084	.168423
rec	1.443133	.6264636	2.30	0.022	.2124929	2.673774
ffin	.104499	.0216129	4.84	0.000	.0620421	.1469558
ghw	.1842855	.0437986	4.21	0.000	.0982465	.2703245
ca	.1659338	.0212623	7.80	0.000	.1241656	.207702
gar	.0480981	.0114351	4.21	0.000	.0256348	.0705614
reg	.1337299	.0226029	5.92	0.000	.0893283	.1781316
loglotrec	-.1611205	.0728071	-2.21	0.027	-.3041446	-.0180965
_cons	7.590715	.2265575	33.50	0.000	7.145659	8.03577

In the end, the only interaction effect that is significant is the interaction with the dummy variable related to the presence of recreational rooms (one might argue that the driveway dummy variable is also significant, as its *p value* was exactly 0.05 in STATA. However, it is actualy 5.0505 … %, which is greater than the 5% significance level chosen for this project).

LETTER G

It will be overestimated, as the "condition" variable is included in the air-conditioning variable and is probably positive, that is, a house in better condition is more valuable. Therefore, it is straightforward to see that the addition of a variable that can accurately captures how well maintained is the house will definitely reduces the effect of the air-conditioning variable.

LETTER F

The regression suing the first 400 observations is as it follows:

Source	SS	df	MS
Model	39.5418672	11	3.5947152
Residual	19.4329558	388	.050084938
Total	58.974823	399	.147806574

Number of obs	= 400
F(11, 388)	= 71.77
Prob > F	= 0.0000
R-squared	= 0.6705
Adj R-squared	= 0.6611
Root MSE	= .2238

logsell	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
loglot	.3137757	.0361499	8.68	0.000	.2427015	.38485
bdms	.0378721	.0174374	2.17	0.030	.0035885	.0721556
fb	.1523751	.0246945	6.17	0.000	.1038234	.2009269
sty	.0882383	.0181935	4.85	0.000	.052468	.1240086
drv	.0864138	.0314094	2.75	0.006	.0246599	.1481676
rec	.0546501	.0339213	1.61	0.108	-.0120425	.1213427
ffin	.1147108	.0267323	4.29	0.000	.0621526	.167269
ghw	.1986973	.0530146	3.75	0.000	.0944655	.3029292
ca	.1776342	.0272387	6.52	0.000	.1240803	.2311881
gar	.0530146	.014797	3.58	0.000	.0239222	.0821069
reg	.1511603	.0421485	3.59	0.000	.0682922	.2340284
_cons	7.673094	.2924037	26.24	0.000	7.0982	8.247988

Using Excel's spreadsheet to calculate the MAE and the STDEV, we obtain:

VARIABLES MODEL			logsellfitted		obs	lnsell	lnlot	bdms	fb	sty	drv	rec	ffin	ghw	ca	gar	reg	error
LOGLOT	0.313		11.49801333		401	11.43496	8.910586	3	1	1	1	1	1	0	1	2	1	0.063049
BDMS	0.037		11.46006813		402	11.23849	8.961879	3	1	1	1	0	1	0	1	2	1	0.22158
FB	0.152		11.36627281		403	11.25803	8.828348	3	1	1	1	1	1	0	1	0	1	0.10824
STY	0.088		11.17518628		404	11.28978	8.757784	3	1	3	1	0	0	0	0	0	1	0.114596
DRV	0.086		11.3137802		405	11.28978	8.794825	4	2	1	1	0	1	0	0	0	1	0.023998
REC	0.054		11.3446936		406	11.3621	8.839277	3	2	1	1	1	1	0	0	0	1	0.017409
FFIN	0.114		11.2847802		407	11.37366	8.794825	3	1	1	1	1	1	0	0	2	1	0.088883
GHW	0.198		11.29212527		408	11.37939	8.767173	3	1	3	1	0	1	0	0	0	1	0.087269
CA	0.177		11.4537802		409	11.39639	8.794825	3	2	1	1	0	1	0	1	0	1	0.057389
GAR	0.053		11.5157802		410	11.40645	8.794825	3	2	3	1	0	0	0	1	0	1	0.109327
REG	0.151		11.2748587		411	11.40756	9.10498	3	1	1	1	0	1	0	0	1	1	0.132706
CONST	7.67		11.51100148		412	11.46163	8.779557	3	2	3	1	0	0	0	1	0	1	0.049369
			11.59218628		413	11.62625	8.757784	3	2	4	1	0	0	0	1	0	1	0.034068
MAE	0.128921		11.03311964		414	10.37036	8.575462	3	1	1	0	0	0	0	1	0	1	0.662758
STDEV	0.256912		11.01593825		415	10.859	7.955074	3	2	2	0	0	1	0	0	0	1	0.156939
			11.39914867		416	11.40756	8.764053	3	1	1	1	1	1	0	1	1	1	0.008416
			11.51960897		417	11.51293	9.321434	3	1	1	1	0	1	0	1	1	1	0.006684
			11.25481421		418	11.42628	8.817298	2	1	1	1	1	1	0	0	2	1	0.171463
			11.77779205		419	12.06968	8.922658	4	2	2	1	0	1	0	1	3	1	0.291888
			11.31994812		420	11.45847	8.699515	3	1	2	1	0	0	1	0	1	1	0.138521
			11.39425981		421	11.12726	9.234057	2	1	1	1	0	0	0	1	2	1	0.266997
			11.36528134		422	11.28978	8.54403	3	1	2	1	1	1	0	1	0	1	0.075499
			11.11116911		423	11.02027	8.131531	3	1	2	1	0	1	0	0	2	1	0.090902
			10.83921576		424	11.0493	7.965546	3	1	2	1	0	0	0	0	0	1	0.210086
			10.98226025		425	11.08981	8.253228	3	1	2	1	0	0	0	0	1	1	0.107545
			10.96912707		426	11.09741	7.962067	2	1	2	1	1	1	0	0	0	1	0.128283
			10.95929855		427	10.80973	8.579229	2	1	1	1	0	0	0	0	1	1	0.149571
			10.82931352		428	10.81978	8.163941	2	1	1	1	0	0	0	0	1	1	0.009535
			10.77344851		429	10.88744	8.154788	2	1	1	1	0	0	0	0	0	1	0.113988
			11.1120597		430	10.97764	8.188689	3	1	1	1	0	1	0	1	0	1	0.134423
			10.86502569		431	11.07442	8.166216	2	1	2	1	0	0	0	0	0	1	0.209395
			11.19929855		432	11.08214	8.579229	3	1	2	1	1	1	0	0	0	1	0.117156
			11.08902786		433	11.14186	8.706159	3	1	1	1	0	0	0	0	2	1	0.052834
			11.23312282		434	11.19821	9.342245	2	1	2	1	0	0	0	0	0	1	0.034908
			11.62526393		435	11.22524	9.035987	3	1	2	1	1	1	0	1	2	1	0.400021
			11.30811964		436	11.22524	8.575462	4	2	1	1	0	0	0	1	0	1	0.082876
			11.30306813		437	11.79056	8.961879	3	2	2	1	0	0	0	0	0	1	0.487489

We can see that:

$$MAE = 0.128$$
$$STDEV = 0.256$$

Therefore, the model sure has a good predictive power (the MAE is about half of the STDEV of the predicted values).