

## Appendix to Practice Paper

### Question 1:

Descriptive Statistics table

	Bedrooms	Bathrooms	Living area	Lot size	Year built	Property tax	Sales price
<b>count</b>	85.000000	85.000000	85.000000	85.000000	85.000000	85.000000	85.000000
<b>mean</b>	3.917647	1.911765	1562.235294	6848.211765	1950.105882	8425.964706	345265.176471
<b>std</b>	0.621284	0.470413	417.003673	1277.568925	2.935917	2081.204702	64991.932441
<b>min</b>	3.000000	1.000000	941.000000	5922.000000	1948.000000	2448.000000	160000.000000
<b>25%</b>	4.000000	2.000000	1271.000000	6000.000000	1948.000000	7680.000000	315000.000000
<b>50%</b>	4.000000	2.000000	1504.000000	6000.000000	1949.000000	8623.000000	340000.000000
<b>75%</b>	4.000000	2.000000	1761.000000	7400.000000	1951.000000	9264.000000	375000.000000
<b>max</b>	5.000000	3.000000	3336.000000	11325.000000	1962.000000	14358.000000	555000.000000

Correlation Matrix

	Bedrooms	Bathrooms	Living area	Lot size	Year built	Property tax	Sales price
<b>Bedrooms</b>	1.000000	0.158141	0.424934	-0.067096	-0.132221	0.208949	0.109961
<b>Bathrooms</b>	0.158141	1.000000	0.394730	0.013284	0.343018	0.293705	0.594456
<b>Living area</b>	0.424934	0.394730	1.000000	0.096276	0.010646	0.456550	0.542321
<b>Lot size</b>	-0.067096	0.013284	0.096276	1.000000	0.193871	-0.029970	0.067451
<b>Year built</b>	-0.132221	0.343018	0.010646	0.193871	1.000000	-0.005848	0.314592
<b>Property tax</b>	0.208949	0.293705	0.456550	-0.029970	-0.005848	1.000000	0.325209
<b>Sales price</b>	0.109961	0.594456	0.542321	0.067451	0.314592	0.325209	1.000000

Summary output of the regression model:

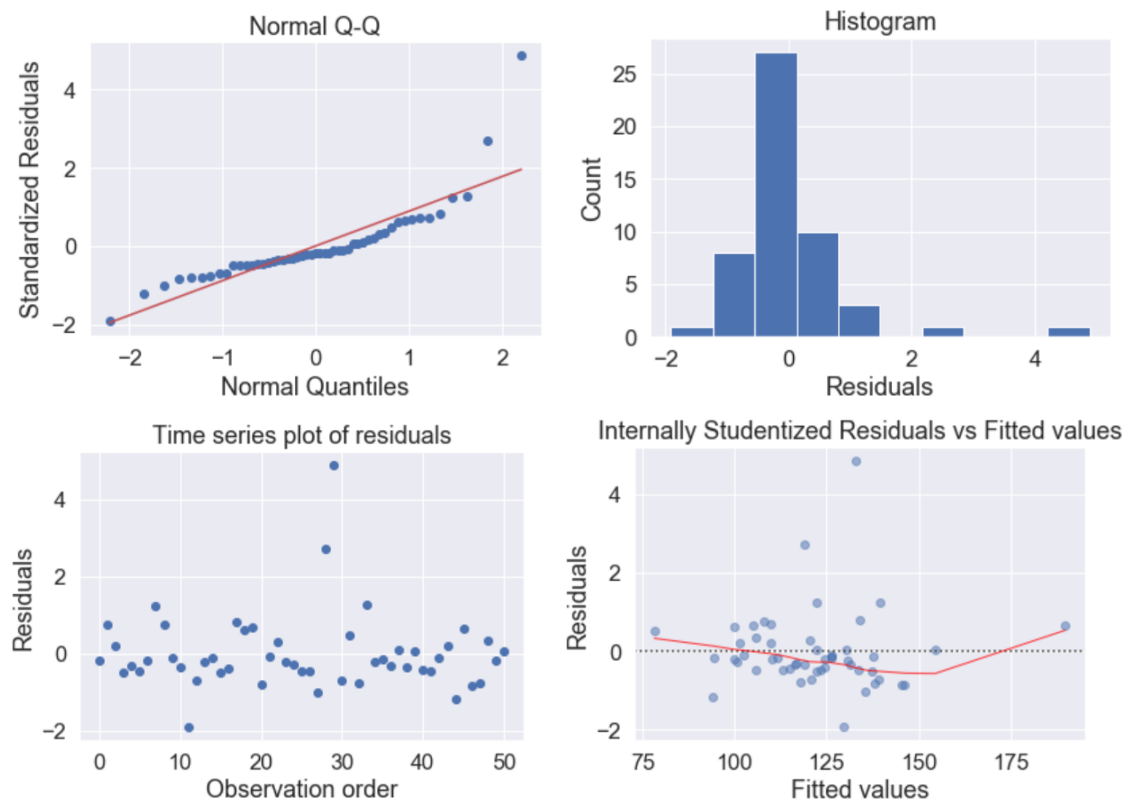
OLS Regression Results						
=====						
Dep. Variable:	Sales price		R-squared:	0.506		
Model:	OLS		Adj. R-squared:	0.469		
Method:	Least Squares		F-statistic:	13.34		
Date:	Fri, 15 May 2020		Prob (F-statistic):	2.42e-10		
Time:	20:54:17		Log-Likelihood:	-1032.1		
No. Observations:	85		AIC:	2078.		
Df Residuals:	78		BIC:	2095.		
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	-7.149e+06	3.82e+06	-1.871	0.065	-1.48e+07	4.56e+05
Bedrooms	-1.229e+04	9346.727	-1.315	0.192	-3.09e+04	6316.893
Bathrooms	5.17e+04	1.31e+04	3.948	0.000	2.56e+04	7.78e+04
Living area	65.9030	15.979	4.124	0.000	34.091	97.715
Lot size	-0.8971	4.194	-0.214	0.831	-9.247	7.453
Year built	3760.8978	1962.504	1.916	0.059	-146.148	7667.944
Property tax	1.4761	2.832	0.521	0.604	-4.163	7.115
=====						
Omnibus:	0.196		Durbin-Watson:	2.144		
Prob(Omnibus):	0.907		Jarque-Bera (JB):	0.123		
Skew:	-0.090		Prob(JB):	0.941		
Kurtosis:	2.956		Cond. No.	8.38e+06		

## Question 2:

Summary output of the regression model:

OLS Regression Results						
=====						
Dep. Variable:	Sales	R-squared:	0.321			
Model:	OLS	Adj. R-squared:	0.228			
Method:	Least Squares	F-statistic:	3.468			
Date:	Fri, 08 May 2020	Prob (F-statistic):	0.00682			
Time:	18:18:08	Log-Likelihood:	-238.85			
No. Observations:	51	AIC:	491.7			
Df Residuals:	44	BIC:	505.2			
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	104.8152	244.110	0.429	0.670	-387.156	596.787
Age	4.6844	3.135	1.494	0.142	-1.634	11.003
HS	0.1038	0.732	0.142	0.888	-1.372	1.579
Income	0.0168	0.014	1.245	0.220	-0.010	0.044
Black	0.3985	0.419	0.952	0.346	-0.445	1.242
Female	-1.2116	5.650	-0.214	0.831	-12.598	10.174
Price	-3.2333	1.023	-3.159	0.003	-5.296	-1.171
=====						
Omnibus:	55.944	Durbin-Watson:	1.671			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	351.282			
Skew:	2.828	Prob(JB):	5.25e-77			
Kurtosis:	14.546	Cond. No.	2.36e+05			
=====						

Residual plots:



VIF table:

	<b>VIF</b>	<b>Predictors</b>
1	2.182	Age
2	3.574	HS
3	4.067	Income
4	1.767	Black
5	2.485	Female
6	1.125	Price

### Question 3:

Model 1 summary:

OLS Regression Results						
=====						
Dep. Variable:	medv	R-squared:	0.648			
Model:	OLS	Adj. R-squared:	0.647			
Method:	Least Squares	F-statistic:	928.1			
Date:	Fri, 08 May 2020	Prob (F-statistic):	2.23e-116			
Time:	18:34:22	Log-Likelihood:	-0.57634			
No. Observations:	506	AIC:	5.153			
Df Residuals:	504	BIC:	13.61			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	3.6176	0.022	164.654	0.000	3.574	3.661
lstat	-0.0461	0.002	-30.465	0.000	-0.049	-0.043
=====						
Omnibus:	27.562	Durbin-Watson:	0.909			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	50.719			
Skew:	0.351	Prob(JB):	9.69e-12			
Kurtosis:	4.383	Cond. No.	29.7			
=====						

Model 2 summary:

OLS Regression Results						
Dep. Variable:	medv	R-squared:	0.641			
Model:	OLS	Adj. R-squared:	0.639			
Method:	Least Squares	F-statistic:	448.5			
Date:	Fri, 08 May 2020	Prob (F-statistic):	1.56e-112			
Time:	18:35:00	Log-Likelihood:	-1581.3			
No. Observations:	506	AIC:	3169.			
Df Residuals:	503	BIC:	3181.			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	42.8620	0.872	49.149	0.000	41.149	44.575
lstat	-2.3328	0.124	-18.843	0.000	-2.576	-2.090
lstat_square	0.0435	0.004	11.628	0.000	0.036	0.051
Omnibus:	107.006	Durbin-Watson:	0.921			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	228.388			
Skew:	1.128	Prob(JB):	2.55e-50			
Kurtosis:	5.397	Cond. No.	1.13e+03			

VIF table for model 2:

VIF Factor	Predictors
1 12.936566	lstat
2 12.936566	lstat_square

#### Question 4:

##### Summary output of the regression model

OLS Regression Results						
=====						
Dep. Variable:	Balance	R-squared:	0.991			
Model:	OLS	Adj. R-squared:	0.990			
Method:	Least Squares	F-statistic:	821.8			
Date:	Sat, 09 May 2020	Prob (F-statistic):	1.14e-42			
Time:	19:50:18	Log-Likelihood:	-155.06			
No. Observations:	50	AIC:	324.1			
Df Residuals:	43	BIC:	337.5			
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	14.3475	4.363	3.289	0.002	5.549	23.146
Purchase	13.9366	1.390	10.029	0.000	11.134	16.739
Expense	-4.9187	0.547	-8.988	0.000	-6.022	-3.815
Renter	13.1473	5.363	2.452	0.018	2.332	23.962
Male	-5.3698	1.720	-3.122	0.003	-8.838	-1.902
Renter*Expense	12.6451	0.771	16.405	0.000	11.091	14.200
Purchase_sq	-0.5091	0.127	-4.006	0.000	-0.765	-0.253
=====						
Omnibus:	2.477	Durbin-Watson:	2.099			
Prob(Omnibus):	0.290	Jarque-Bera (JB):	1.557			
Skew:	0.166	Prob(JB):	0.459			
Kurtosis:	2.202	Cond. No.	371.			
=====						

#### Question 5:

##### Summary of all possible models:

	numb_features	SSE	R_squared	adj_R2	AIC	BIC	features
0	1	35797.217408	0.366024	0.339608	265.700555	268.216748	(X1,)
1	1	55951.330756	0.009090	-0.032198	277.312469	279.828662	(X2,)
2	1	53093.903632	0.059696	0.020517	275.949546	278.465739	(X3,)
3	1	46806.597780	0.171045	0.136506	272.672559	275.188753	(X4,)
4	2	28802.072515	0.489909	0.445554	262.047562	265.821852	(X1, X2)
5	2	31061.323721	0.449898	0.402063	264.010980	267.785270	(X1, X3)
6	2	35266.830049	0.375417	0.321106	267.312445	271.086735	(X1, X4)
7	2	53011.176300	0.061161	-0.020477	277.909003	281.683292	(X2, X3)
8	2	43830.546780	0.223752	0.156252	272.964537	276.738826	(X2, X4)
9	2	32251.674479	0.428816	0.379148	264.988750	268.763040	(X3, X4)
10	3	26189.842672	0.536172	0.472923	261.575594	266.607980	(X1, X2, X3)
11	3	27713.191707	0.509194	0.442266	263.045553	268.077939	(X1, X2, X4)
12	3	26098.477656	0.537791	0.474762	261.484733	266.517119	(X1, X3, X4)
13	3	30544.017741	0.459059	0.385294	265.574321	270.606707	(X2, X3, X4)
14	4	21282.821658	0.623077	0.551282	258.181338	264.471820	(X1, X2, X3, X4)

# Question 6:

Trained model in leave one out CV for model A:

1<sup>st</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	-50.7580	25.297	-2.007	0.294	-372.183	270.667
X1	1.1922	0.078	15.317	0.042	0.203	2.181
X2	-0.3906	0.011	-34.240	0.019	-0.535	-0.246

2<sup>nd</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	-444.8126	458.900	-0.969	0.510	-6275.691	5386.065
X1	2.2504	1.520	1.481	0.378	-17.058	21.559
X2	-0.1820	0.215	-0.846	0.553	-2.915	2.552

3<sup>rd</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	-231.1106	123.629	-1.869	0.313	-1801.970	1339.749
X1	0.7293	0.519	1.406	0.394	-5.862	7.321
X2	0.3652	0.135	2.696	0.226	-1.356	2.086

4<sup>th</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	-612.9171	332.382	-1.844	0.316	-4836.233	3610.399
X1	2.8360	1.113	2.547	0.238	-11.311	16.983
X2	-0.2216	0.152	-1.461	0.382	-2.149	1.706

5<sup>th</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	-1096.0094	926.966	-1.182	0.447	-1.29e+04	1.07e+04
X1	4.3500	3.014	1.443	0.386	-33.949	42.649
X2	-0.3280	0.261	-1.257	0.428	-3.644	2.988

Trained model in leave one out CV for model B:

1<sup>st</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	239.4406	1909.203	0.125	0.921	-2.4e+04	2.45e+04
X1	0.5940	3.715	0.160	0.899	-46.605	47.793
X3	-0.8029	2.902	-0.277	0.828	-37.676	36.070

2<sup>nd</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	-85.6243	789.062	-0.109	0.931	-1.01e+04	9940.360
X1	1.5472	1.606	0.963	0.512	-18.860	21.955
X3	-0.6464	1.737	-0.372	0.773	-22.716	21.423

3<sup>rd</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	-675.7025	158.196	-4.271	0.146	-2685.776	1334.371
X1	1.8507	0.350	5.288	0.119	-2.596	6.298
X3	0.8177	0.343	2.385	0.253	-3.538	5.174

4<sup>th</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	-404.3874	688.141	-0.588	0.662	-9148.044	8339.269
X1	2.0364	1.706	1.194	0.444	-19.636	23.709
X3	-0.1392	1.470	-0.095	0.940	-18.812	18.534

5<sup>th</sup>

	coef	std err	t	P> t	[0.025	0.975]
const	1094.0232	1505.002	0.727	0.600	-1.8e+04	2.02e+04
X1	-9.0211	10.372	-0.870	0.544	-140.804	122.762
X3	7.1095	6.882	1.033	0.490	-80.331	94.550

## Question 7:

### Summary output:

Optimization terminated successfully.

Current function value: 0.213313

Iterations 9

Logit Regression Results						
Dep. Variable:	realY	No. Observations:	32950			
Model:	Logit	Df Residuals:	32934			
Method:	MLE	Df Model:	15			
Date:	Sat, 09 May 2020	Pseudo R-squ.:	0.3986			
Time:	06:03:58	Log-Likelihood:	-7028.7			
converged:	True	LL-Null:	-11688.			
Covariance Type:	nonrobust	LLR p-value:	0.000			
	coef	std err	z	P> z	[0.025	0.975]
const	-15.4255	15.704	-0.982	0.326	-46.204	15.353
duration	0.0046	8.15e-05	56.890	0.000	0.004	0.005
nr_employed	-0.0064	0.001	-5.548	0.000	-0.009	-0.004
poutcome_success	1.9147	0.094	20.271	0.000	1.730	2.100
emp_var_rate	-0.4837	0.074	-6.570	0.000	-0.628	-0.339
previous	0.0767	0.062	1.231	0.218	-0.045	0.199
poutcome_nonexistent	0.5958	0.104	5.756	0.000	0.393	0.799
contact_telephone	-0.3073	0.067	-4.572	0.000	-0.439	-0.176
month_mar	1.3251	0.121	10.947	0.000	1.088	1.562
month_oct	0.1288	0.115	1.119	0.263	-0.097	0.354
cons_price_idx	0.4652	0.110	4.222	0.000	0.249	0.681
month_sep	-0.1861	0.136	-1.373	0.170	-0.452	0.080
month_may	-0.8757	0.059	-14.935	0.000	-0.991	-0.761
default_no	0.3495	0.072	4.866	0.000	0.209	0.490
job_student	0.4418	0.107	4.124	0.000	0.232	0.652
job_retired	0.3845	0.086	4.458	0.000	0.215	0.554

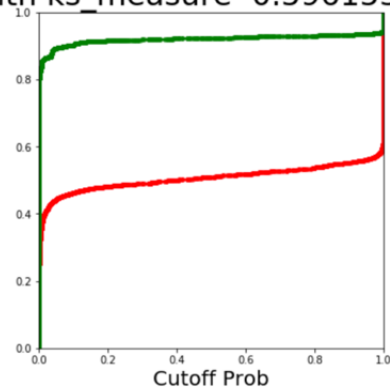
## Question 8:

### Confusion table:

	realY	0	1	PV
predictedY	0	26946.000000	1911.000000	
	1	2249.000000	1844.000000	
SS				

### KS chart:

KS\_chart with ks\_measure 0.5961532365705691



KS\_measure is achieved when the cutoff probability is 0.00028830886440440583



