STAT 500 Homework 11

Vu Thi-Hong-Ha; NetID: 851924086

November 13, 2020

1 Question 1

(a)

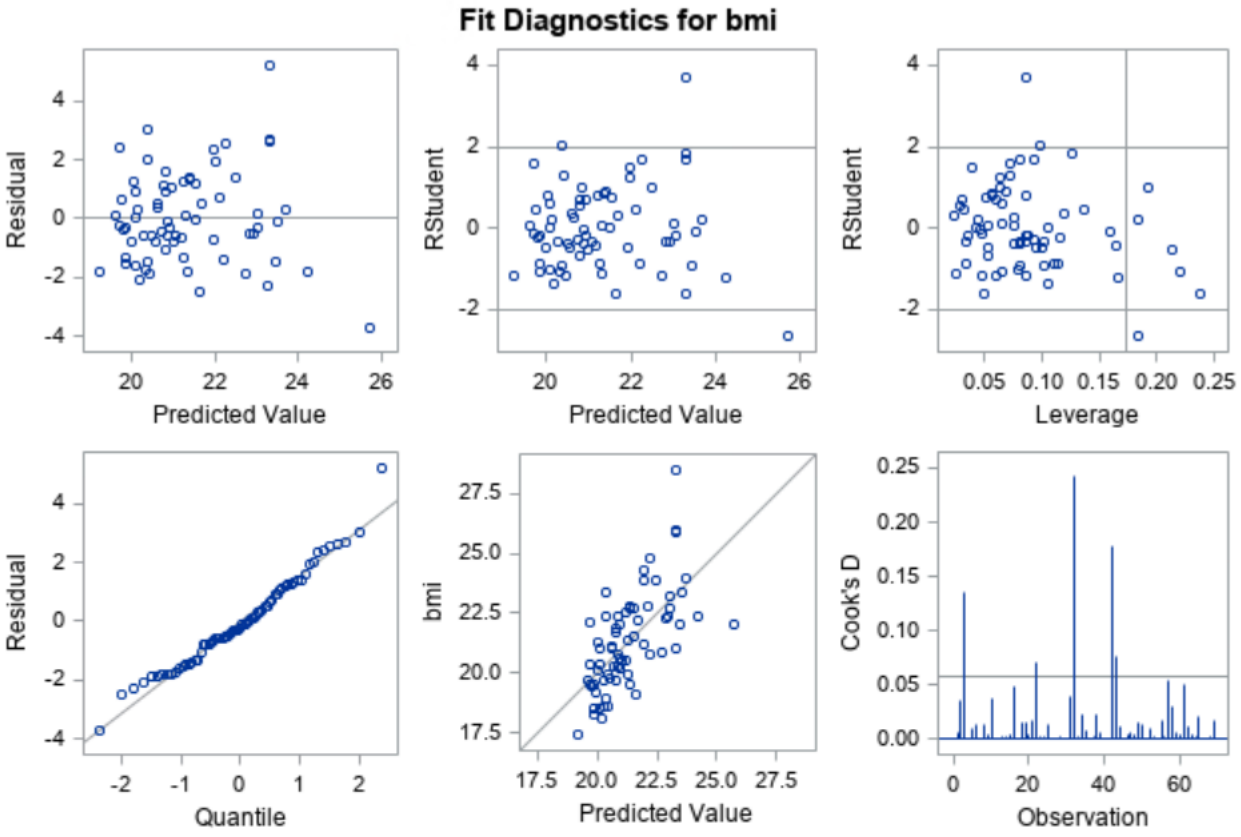
(i) The q-q plot indicates the data set may be heavy tailed. Shapiro-Wilk test gives test statistics = 0.869947 and p-value < 0.0001, so we reject the null that the data follows normal distribution at a significance level of 0.05.

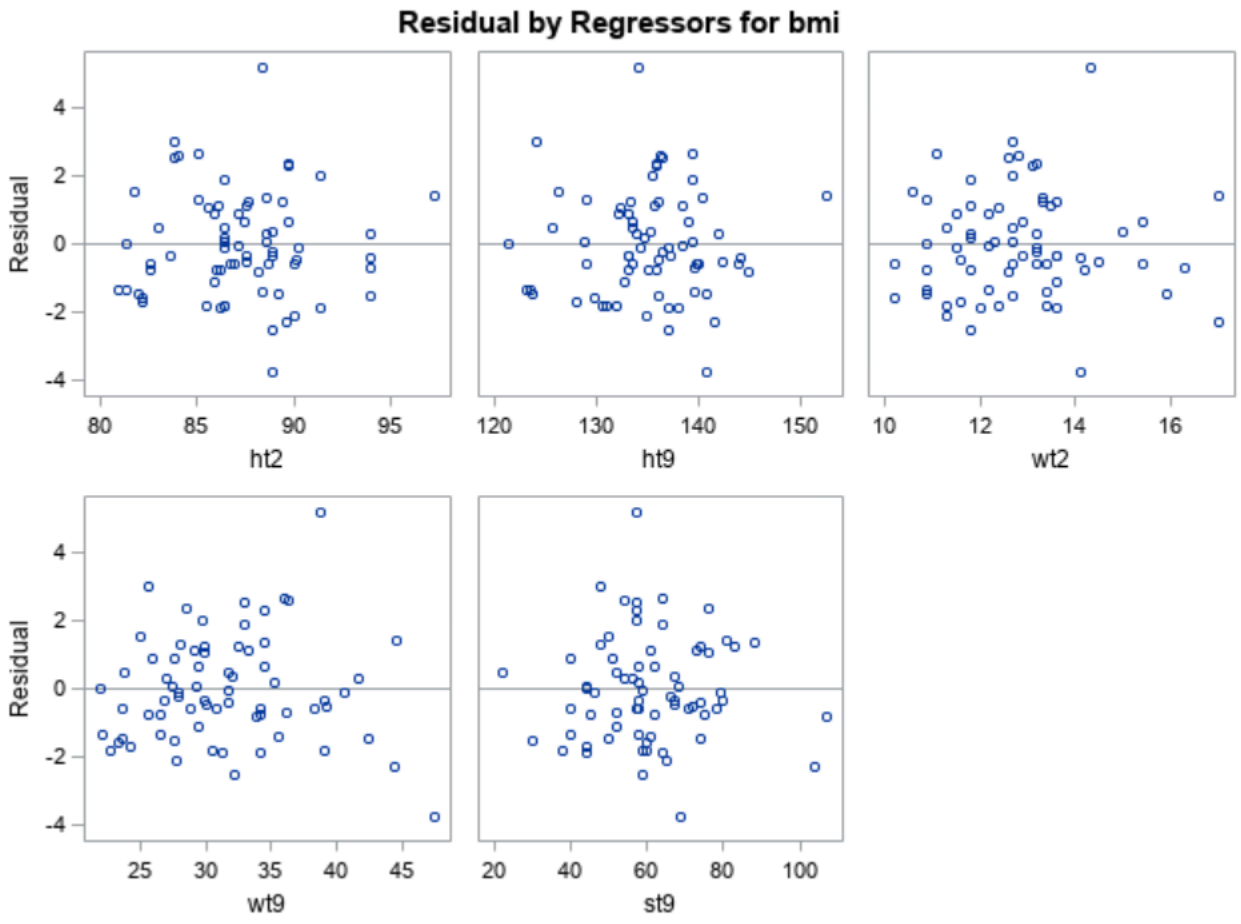
(ii) From the plot of residuals vs predicted values, we can see there is one potential outlier which has very high residual (¿ 10). The data points seem to spread out towards the right of the plot, indicating there might be a difference in variances, but if we remove the outlier, homogeneous variances seem to hold. Using the plot of studentized residual plot vs predicted values, there may be two more outliers, but they do not look too bad.

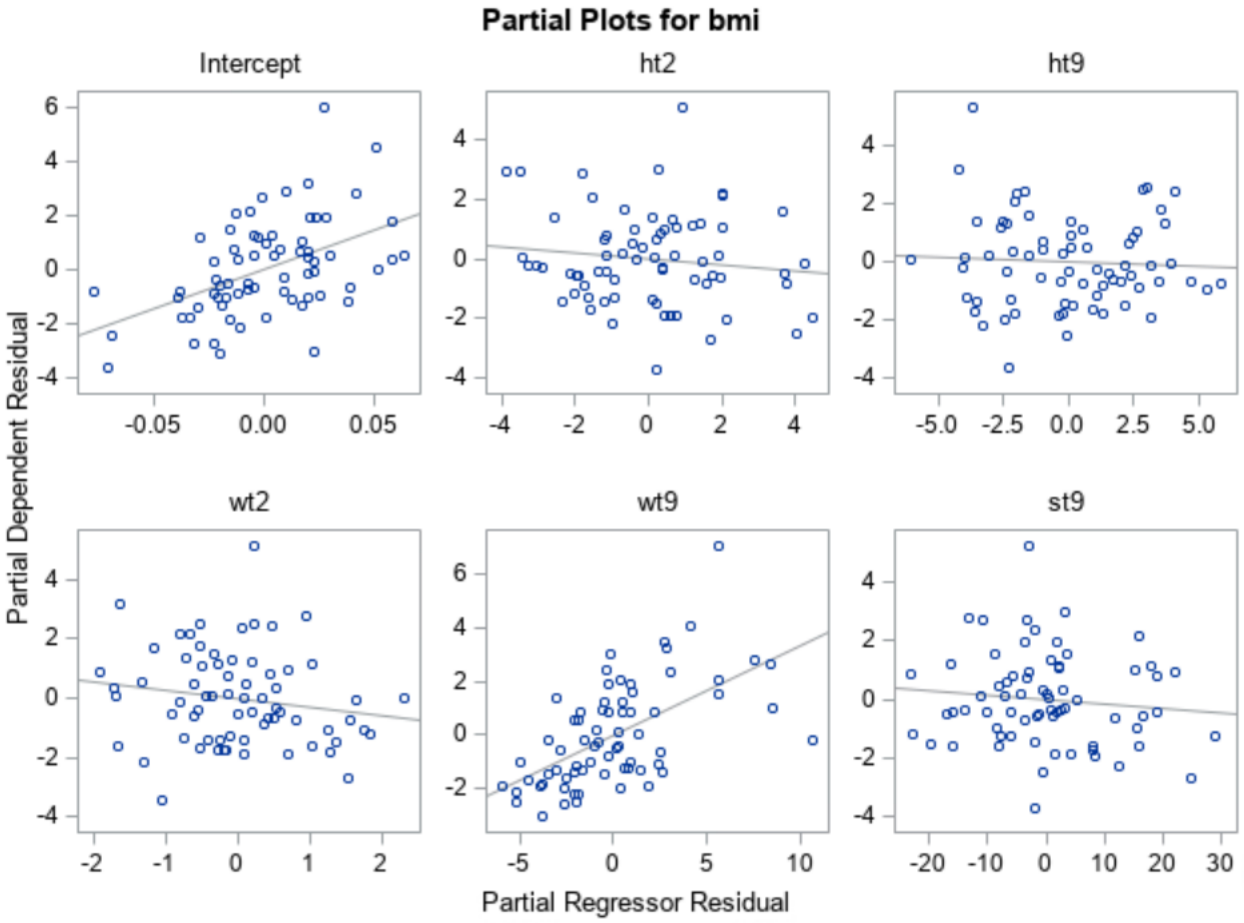
(iii) From the plots of residuals vs each explanatory variable, we can still see clearly the outlier. Otherwise, the data points seem to spread out randomly, indicating homogeneous variances.

(iv) From all the plots, we can still the the outlier. If we exclude the outlier, we see weak negative association between BMI and HT2, BMI and HT9, BMI and WT2, and BMI and ST9 after adjusting for the relationship between BMI and other variables. We also see a positive association between BMI and WT9.

(b)


Fit Diagnostics for bmi
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
bmi
Predicted Value
Cook's D
Observation


Residual by Regressors for bmi
Residual
ht2
ht9
wt2
wt9
st9



(i) The q-q plot looks good, the errors after removing the outlier seem to follow normal distribution. Shapiro-Wilk test gives test statistics = 0.976133 and p-value 0.2089, so we fail to reject the null that the data follows normal distribution at a significance level of 0.05.

(ii) From the plot of residuals vs predicted values, we can see there is some potential outliers that have high residual (> 4 and $\approx$ -4). The data points seem to spread randomly, indicating homogeneous variances seem to hold. Using the plot of studentized residual plot vs predicted values, there may be two outliers, as in the plot of residuals vs predicted values.

(iii) From the plots of residuals vs each explanatory variable, we can see some potential outliers. Otherwise, the data points seem to spread out randomly, indicating homogeneous variances.

(iv) From all the plots, we can still the the potential outliers, but they do not look too bad. We see weak negative association between BMI and HT2, BMI and HT9, BMI and WT2, and BMI and ST9 after adjusting for the relationship between BMI and other variables. We also see a strong positive association between BMI and WT9.

(c)

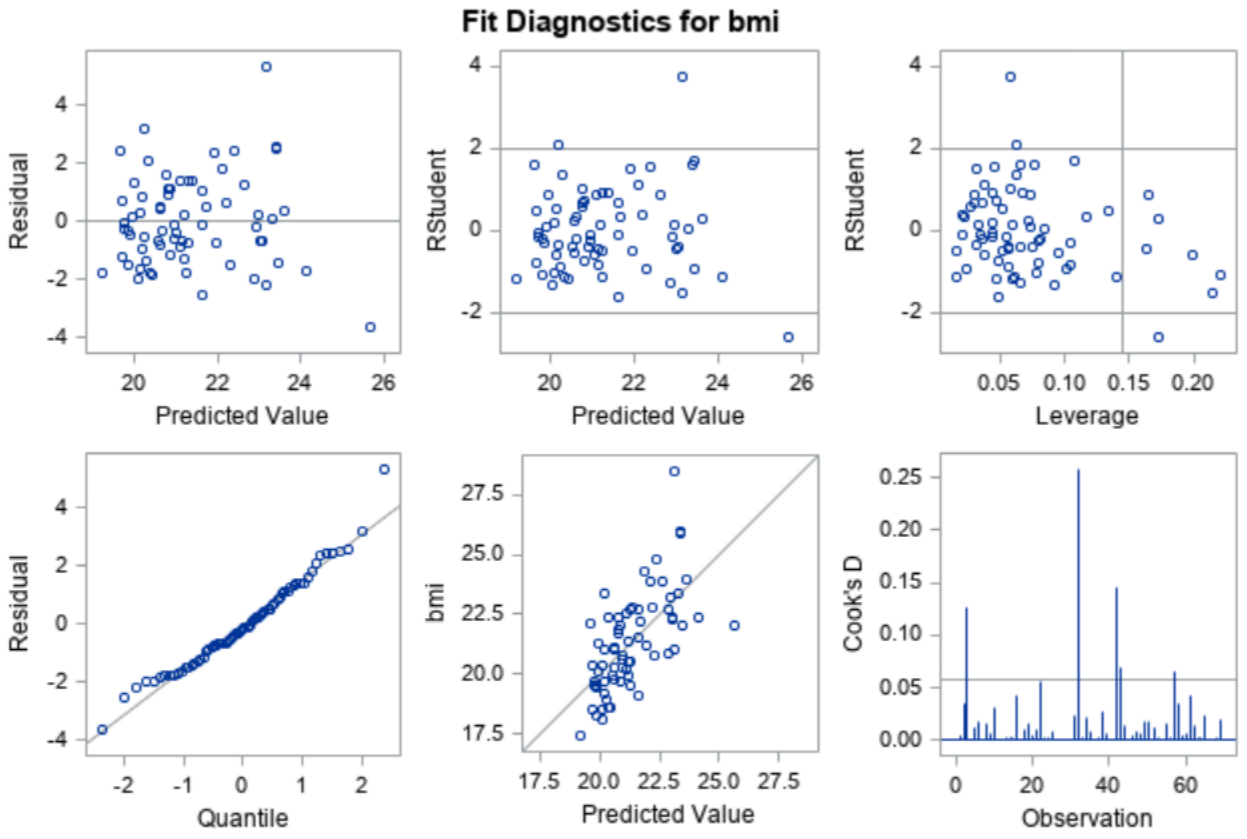
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	111.18570	55.59285	20.91	<.0001
Error	66	175.45719	2.65844		
Corrected Total	68	286.64290			

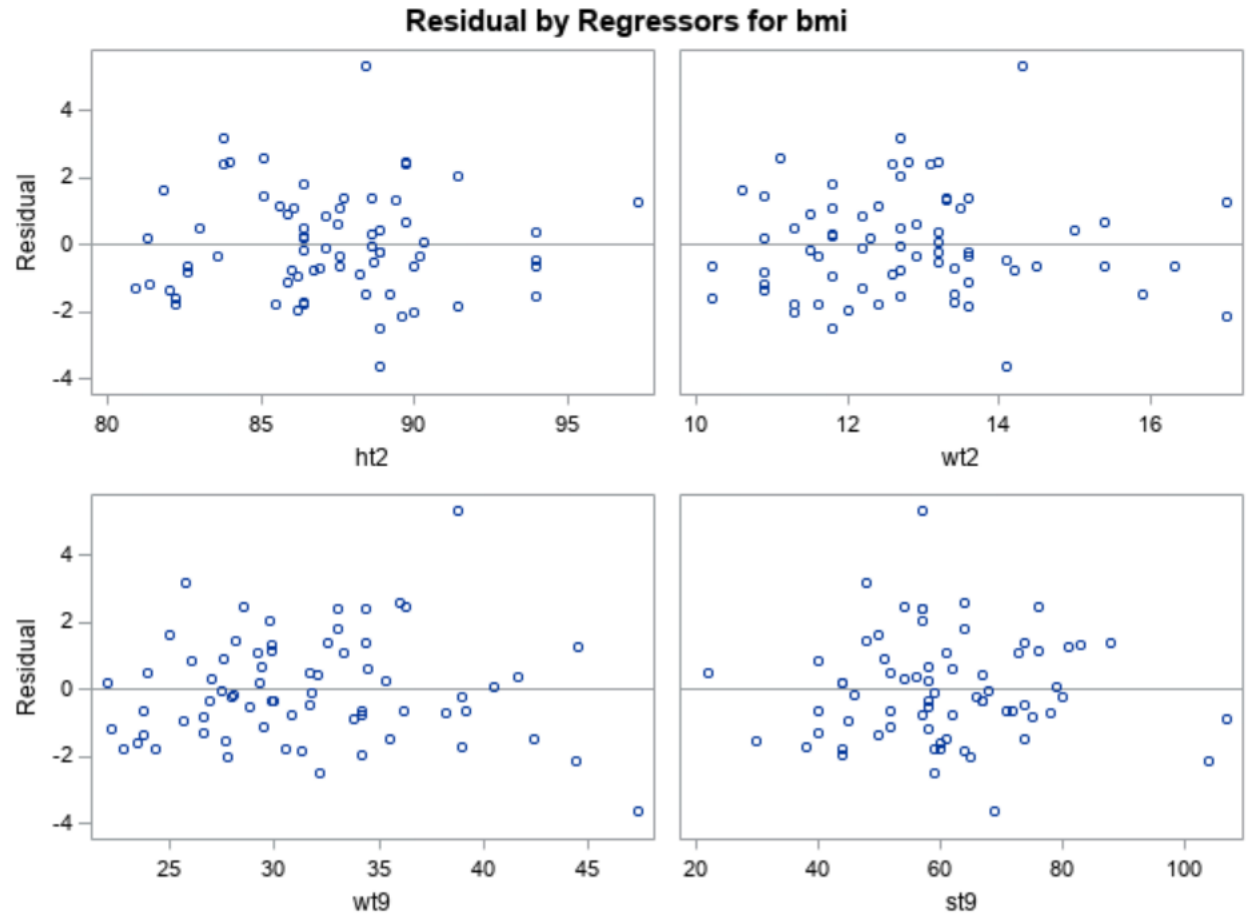
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	29.64834	5.53134	76.37786	28.73	<.0001
ht2	-0.19011	0.07010	19.55141	7.35	0.0085
wt9	0.26212	0.04092	109.06019	41.02	<.0001

The final model includes the intercept and two variables HT2 and WT9. The estimated intercept is 29.64834 with standard error = 5.53134. The estimated coefficient for HT2 is -0.19011 with standard error = 0.07010. The estimated coefficient for WT9 is 0.26212 with standard error = 0.04092.

(d) In my opinion, we should choose the model that includes HT2, WT2, WT9 and ST9. Firstly, this model has the largest R^2 in all 4-variable models. The R^2 is this model is also greater than those of 1, 2 and 3-variable models. It is smaller than the R^2 of 5 variable model, but the difference is negligible (only 0.2%). Having 4 variables gives us a simpler model than including every variable. The model that includes HT2, WT2, WT9 and ST9 also have reasonably small values of C_p, BIC and AIC.

(e)


Fit Diagnostics for bmi
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
bmi
Predicted Value
Cook's D
Observation


Residual by Regressors for bmi
Residual
ht2
wt2
Residual
wt9
st9

From the q-q plot, it looks like the data follows normal distribution. Shapiro-Wilk test gives test statistics = 0.973709 and p-value 0.1537, so we fail to reject the null that the data follows normal distribution at a significance level of 0.05.

From the residual vs predicted values plot, there are some potential outliers. The points scatter randomly and around the 0 line, indicating homogeneous variances.

(f)

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Variance Inflation
Intercept	1	27.55657	5.59833	4.92	<.0001	31334	62.69315	0
ht2	1	-0.13421	0.07738	-1.73	0.0877	2.12552	7.78362	1.76955
wt2	1	-0.26169	0.20564	-1.27	0.2078	13.76641	4.19034	2.49864
wt9	1	0.31747	0.04990	6.36	<.0001	100.90994	104.75292	2.15858
st9	1	-0.01938	0.01514	-1.28	0.2052	4.23849	4.23849	1.37439

All VIFs of the variables are smaller than 4, so there may not be any concerns about multicollinearity.

2 Question 2

(a) Model: $\hat{Y}_i = \beta_0 + \beta_1 * LivingArea + \beta_2 * Age + \beta_3 * BasementArea + \beta_4 * TotalRoom + \beta_5 * GarageSize$

- Testing for the significance of the overall model:

H_0 : all coefficients in the model are 0; H_a : at least one coefficient is non-zero. The test statistics F = 669.29 with p-value $<$ .0001. At a significance level of 0.05, we reject the null. We conclude that at least one coefficient is non-zero.

- Testing for the significance of the intercept:

$H_0 : \beta_0 = 0$; $H_a : \beta_0 \neq 0$. The test statistics t = 4.68 with p-value $<$ .0001. At a significance level of 0.05, we reject the null. We conclude that β_0 is non-zero.

- Testing for the significance of β_1:

$H_0 : \beta_1 = 0$; $H_a : \beta_1 \neq 0$. The test statistics t = 22.90 with p-value $<$ .0001. At a significance level of 0.05, we reject the null. We conclude that β_1 is non-zero.

- Testing for the significance of β_2:

$H_0 : \beta_2 = 0$; $H_a : \beta_2 \neq 0$. The test statistics t = -14.12 with p-value $<$ .0001. At a significance level of 0.05, we reject the null. We conclude that β_2 is non-zero.

- Testing for the significance of β_3:

$H_0 : \beta_3 = 0$; $H_a : \beta_3 \neq 0$. The test statistics t = 17.10 with p-value < .0001. At a significance level of 0.05, we reject the null. We conclude that β_3 is non-zero.

- Testing for the significance of β_4:

$H_0 : \beta_4 = 0$; $H_a : \beta_4 \neq 0$. The test statistics t = -4.85 with p-value < .0001. At a significance level of 0.05, we reject the null. We conclude that β_4 is non-zero.

- Testing for the significance of β_5:

$H_0 : \beta_5 = 0$; $H_a : \beta_5 \neq 0$. The test statistics t = 0.42 with p-value 0.6719. At a significance level of 0.05, we fail to reject the null. We conclude that β_5 is non-zero.

By stepwise model selection method, we should include the intercept, the variables LivingArea, Age, BasementArea and TotalRoom in the model. The estimated intercept is 32673 with standard error = 6566.47208. The estimated coefficient for LivingArea is 104.75703 with standard error = 4.45529. The estimated coefficient for Age is -746.48764 with standard error = 46.74967. The estimated coefficient for BasementArea is 58.12806 with standard error = 3.38917. The estimated coefficient for TotalRoom is -6481.53764 with standard error = 1330.68534.

(i) By all possible model method, and other tests carried out above, I decided to choose the model including LivingArea, Age, BasementArea, and TotalRoom as variables because this model has high R^2 than models with fewer variables or other 4-variable models. This model is also the only model that has $C_p < 6$. Including GarageSize clearly does not improve R^2 for the model.

(ii)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	32673	6566.47208	4.98	<.0001	19787	45559
LivingArea	1	104.75703	4.45529	23.51	<.0001	96.01417	113.49988
Age	1	-746.48764	46.74967	-15.97	<.0001	-838.22702	-654.74827
BasementArea	1	58.12806	3.38917	17.15	<.0001	51.47731	64.77881
TotalRoom	1	-6481.53764	1330.68534	-4.87	<.0001	-9092.81258	-3870.26269

(iii)

Analysis of Variance

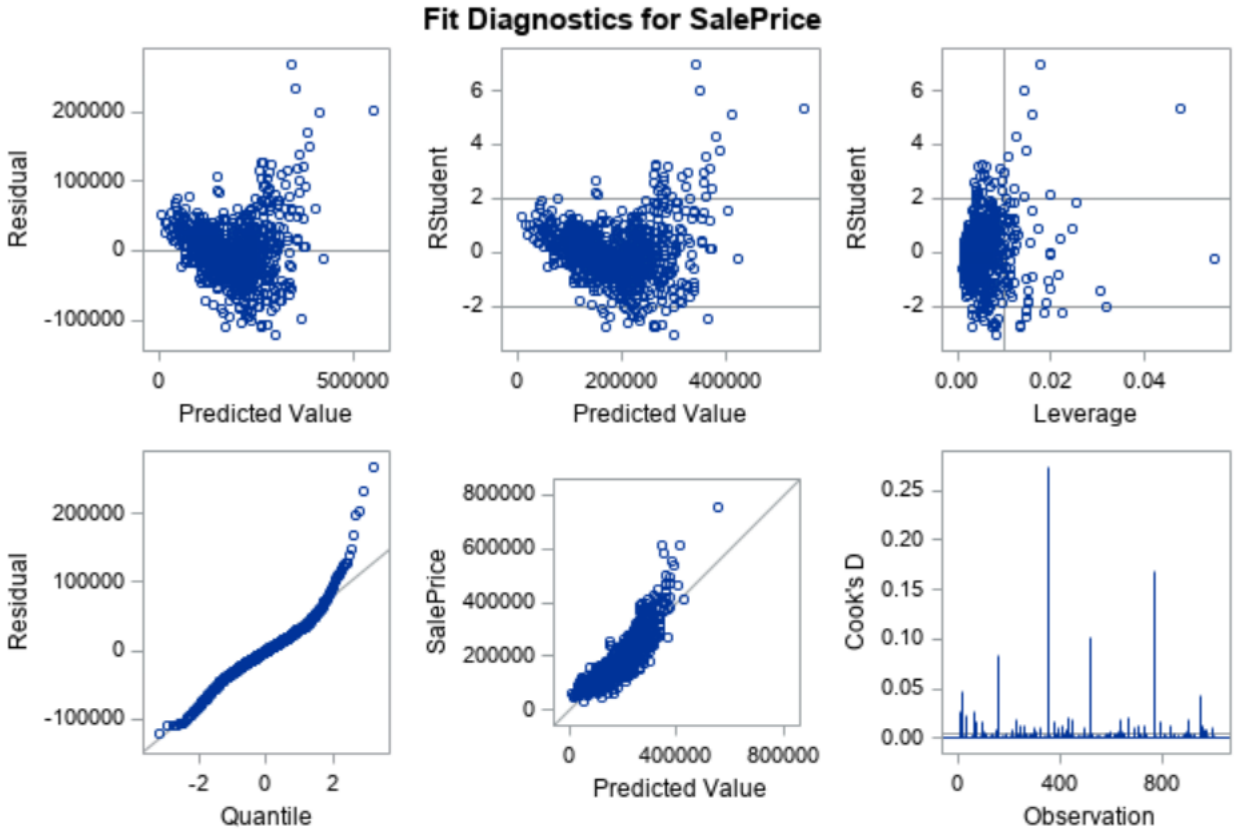
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5.277833E12	1.319458E12	837.26	<.0001
Error	994	1.566478E12	1575933256		
Corrected Total	998	6.844311E12			

(iv)

Root MSE	39698	R-Square	0.7711
Dependent Mean	183044	Adj R-Sq	0.7702
Coeff Var	21.68767		

$R^2 = 0.7711$, meaning that this model can explain 77.11% of the variances.

(v)



From the residual vs predicted value plot, we can see that there are many potential outliers, espcially those with residuals > 200000. The data points seem to spread out more to the right of the plot (megaphone shaped), indicating there may be differences in the variances. from studentized residual vs predicted values and leverage plots, we can see there are several outliers with residuals and leverage greater than the thresholds. From the q-q plot, the data may be heavy tailed, and there are potential outliers.

(b)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	9.037262E12	2.259316E12	1553.51	<.0001
Error	1919	2.790858E12	1454329388		
Corrected Total	1923	1.182812E13			

Root MSE	38136	R-Square	0.7640
Dependent Mean	179682	Adj R-Sq	0.7636
Coeff Var	21.22400		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	31813	4520.11936	7.04	<.0001	22948	40678
LivingArea	1	98.50489	3.31417	29.72	<.0001	92.00514	105.00464
Age	1	-776.22179	31.67872	-24.50	<.0001	-838.35012	-714.09345
BasementArea	1	60.94602	2.46440	24.73	<.0001	56.11284	65.77921
TotalRoom	1	-5303.86093	986.41063	-5.38	<.0001	-7238.41039	-3369.31146

$R^2 = 0.7640$, meaning that this model can explain 76.4% of the variances.

$MS_{error-evaluation} = 1454329388$, while $MS_{error-training} = 1575933256$. $MS_{error-evaluation} < MS_{error-training}$ because we have more observations in the evaluation set, so the error has more dfs.