Leyli Garryyeva
Professor deCamp

Assignment 9

Ch.10-Building the Regression Model II: Diagnostics: 8, 12, 18, 24.

Problem 10.8: Refer to Commercial properties Problem 6.18c.

a. Prepare an added-variable plot for each of the predictor variables.

$R_{Y1|234} = 36.8\%$
$R_{Y2|134} = 20.8\%$
$R_{Y3|124} = 0.4\%\%$
$R_{Y4|124} = 30.1\%$

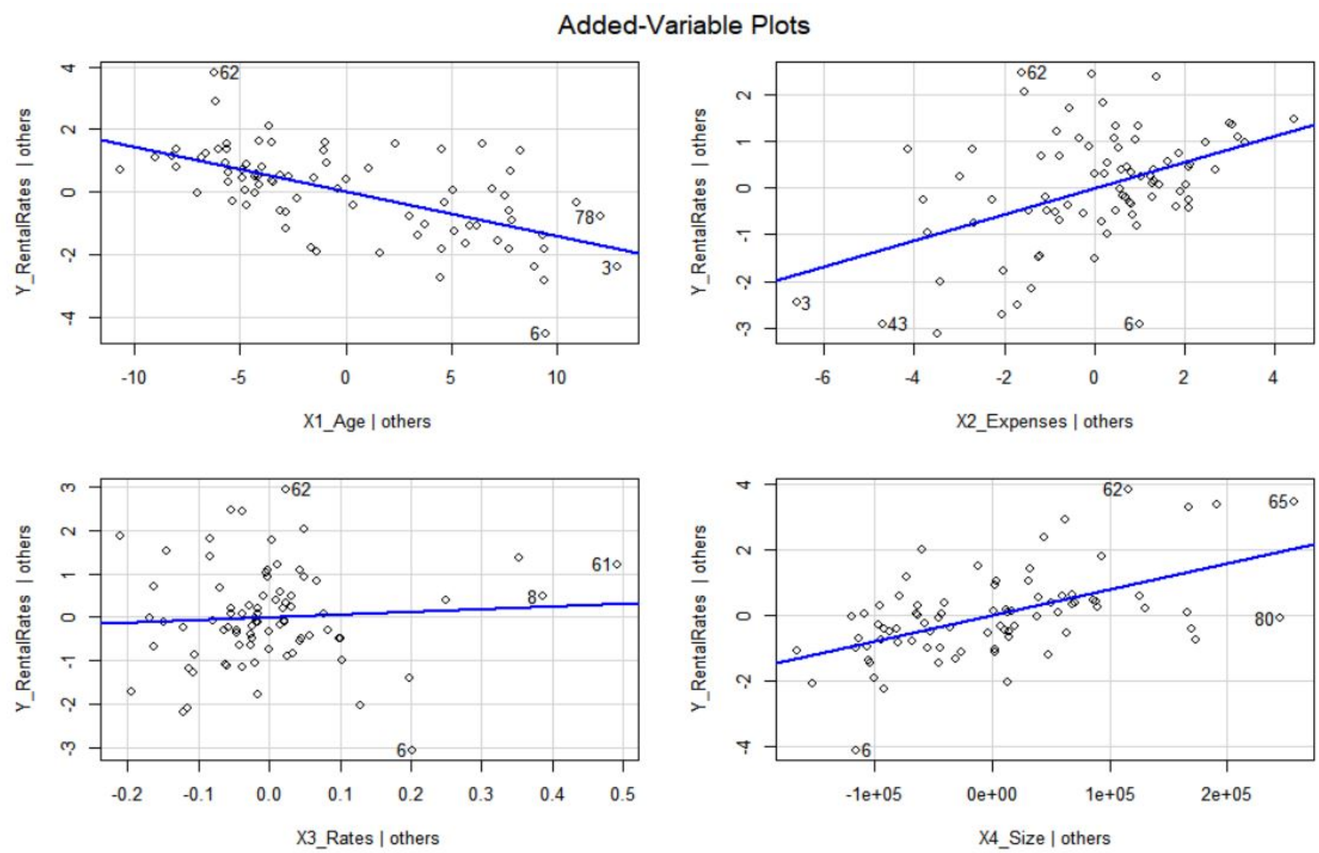

Figure 1: Added-Variable Plots

b. Do your plots in part (a) suggest that the regression relationships in the fitted regression function in Problem 6.18c are inappropriate for any of the predictor variables? Explain.

Figure 1 shows that adding X_3 `Vacancy Rates` contains extremely little additional information useful for predicting Y beyond that contained in X_1, X_2, and X_4. Then, adding X_3 to the regression model will not be helpful and the the fitted regression function in Problem 6.18c is not appropriate for this predictor variable.

On the other hand, the added-variable plots for X_1, X_2, and X_4 show linear bands with non-zero slopes and indicate that these predictor variables may be helpful when added to the regression model when other variables are held constant.

Problem 10.12: Refer to Commercial Properties Problem 6.18.

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = 0.01$. State the decision rule and conclusion.

Calculating the Boneferroni critical value:

$$t(1 - \alpha/2n; n - p - 1) = t(1 - 0.1/(2 * 81); 81 - 5 - 1) = t(0.9999; 75) = 4.049$$

Decision Rule: If $|t_i| \leq 4.049,$ conclude that a case i is an outlier

Conclusion: Based on the Figure 2 there are no points such that $|t_i| \leq 4.049$, hence, there are no outliers. Yet, we might want to investigate cases 6 and 62, because if the Bonferroni procedure required $\alpha \geq 0.05$ these two cases would have been potential outliers.

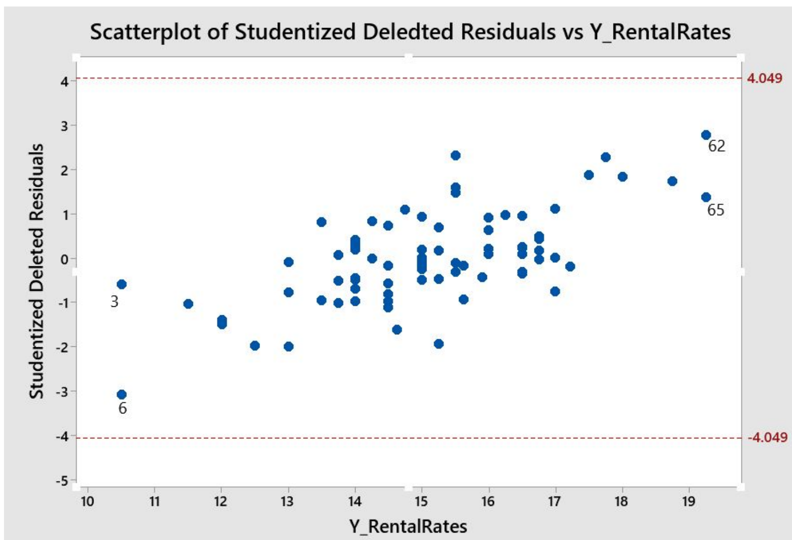

Figure 2: Studentized Deleted Resiudals Plot

b. Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

Using $2p/n = 2 * 5/81 = 0.1235$ as a reference line and refereing to the Figure 3, we identify cases 3, 8, 53, 61, 65 as outlying X observations.

c. The researcher wishes to estimate the rental rates of a property whose age is 10 years, whose operating expenses and taxes are 12.00, whose occupancy rate is 0.05, and whose square footage is 350,000. Use (10.29) to determine whether this estimate will involve a hidden extrapolation.

(10.29) $h_{\text{new,new}} = X'_{\text{new}}(X'X)^{-1}X_{\text{new}}$

$$X_{\text{new}} = \begin{bmatrix} 1 \\ 10 \\ 12 \\ 0.05 \\ 350000 \end{bmatrix}$$

$h_{\text{new,new}} = X'_{\text{new}}(X'X)^{-1}X_{\text{new}} = 0.0529$, there does not seem to be hidden extrapolation since $h_{\text{new,new}}$ is a reasonable leverage value and X values are within the data region.

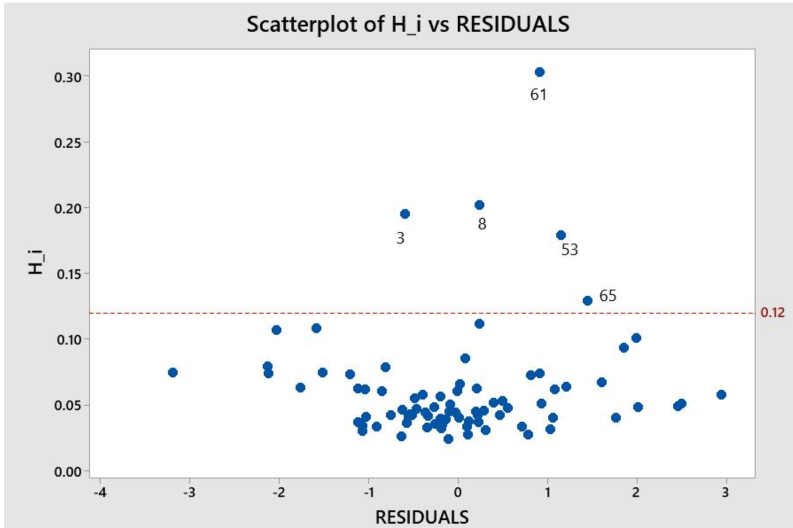

Figure 3: Diagonal Elements of the hat matrix Plot

d. Cases 61, 8, 3, and 53 appear to be outlying X observations, and cases 6 and 62 appear to be outlying Y observations. Obtain the DFFITS, DFBETAS, and Cook's distance values for each case to assess its influence. What do you conclude?

Consider case influential if DFFITS $> 2\sqrt{p/n} = 0.5$ *and if COOKS* $> F_{0.5,p,n-p} = F_{0.5,5,76} = 0.88$. *Based on the Table below, none of the cases are influential (although Cook's distance for the Case 6 is close to 0.88).*

Cases	3	8	53	61	6	62
DFFITS	0.0163	0.0027	0.0550	0.0817	0.1374	0.0875
COOKS	-0.2843	0.1164	0.5252	0.6387	-0.8735	0.6903

e. Calculate the average absolute percent difference in the fitted values with and without each of the cases. What does this measure indicate about the influence of each case?

All average absolute percent differences between $\hat{Y}$ and $\hat{Y}_{(i)}$ are below 1 %:

Case 3: 0.192 %
Case 6: 0.556 %
Case 8: 0.054%
Case 53: 0.235 %
Case 61: 0.3 %
Case 62: 0.417 %

f. Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

Based on the Figure 5, Case 6 seems to be most influential compared to other cases. After assessing the magnitude of the influence, we find that D_6 is the 41st percentile of this distribution. Then, despite the influence of case 6, this influence may not be large enough to call for consideration of remedial measures since it's below 0.88.

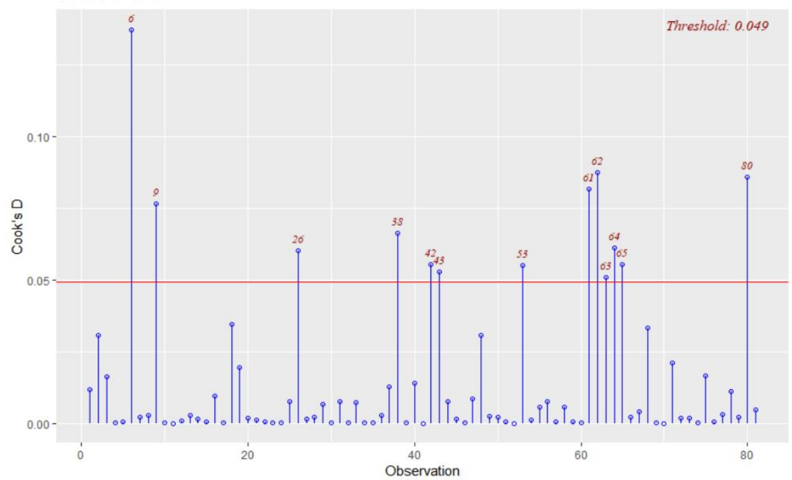

Figure 4: Cook's Distance index influence plot

Problem 10.18: Refer to Commercial properties Problem 6.18b.

a. What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among the predictor variables?

Figure 5: Scatter plot matrix

Based on the scatter plot matrix and the correlation matrix, `Age` *is positively correlated* `Expenses and Taxes` *and* `Size`*, while latter two are also positively correlated with each other. Also,* `Rates` *is negatively correlated with* `Age` *and* `Expenses and Taxes`*.*

Correlation: X1_Age, X2_Expenses, X3_Rates, X4_Size

Method

Correlation type	Pearson
Rows used	81

ρ: pairwise Pearson correlation

Correlations

	X1_Age	X2_Expenses	X3_Rates
X2_Expenses	0.389		
X3_Rates	-0.253	-0.380	
X4_Size	0.289	0.441	0.081

Pairwise Pearson Correlations

Sample 1	Sample 2	Correlation	95% CI for ρ	P-Value
X2_Expenses	X1_Age	0.389	(0.186, 0.560)	0.000
X3_Rates	X1_Age	-0.253	(-0.446, -0.036)	0.023
X4_Size	X1_Age	0.289	(0.075, 0.477)	0.009
X3_Rates	X2_Expenses	-0.380	(-0.552, -0.176)	0.000
X4_Size	X2_Expenses	0.441	(0.246, 0.601)	0.000
X4_Size	X3_Rates	0.081	(-0.140, 0.294)	0.474

Figure 6: Correlation matrix

b. Obtain the four variance inflation factors. Do they indicate that a serious multicollinearity problem exists here?

X	1	2	3	4
VIF_k	1.240348	1.648225	1.323552	1.412722

None of the VIF values are large and $V\bar{I}F_k = 1.406$ *, which indicates that the multicollinearity is not severe for the predictor variables.*

Problem 10.24

If $n = p$ and the **X** matrix is invertible, use (5.34) and (5.37) to show that the hat matrix **H** is given by the pxp identity matrix. In this case, what are $h_i i$ and $\hat{Y}_i$?

(5.34) $(AB)^{-1} = B^{-1}A^{-1}$
(5.37) $(A')^{-1} = (A^{-1})'$

$$H = X(X'X)^{-1}X' = X\left[X^{-1}(X')^{-1}\right]X' = \left[XX^{-1}\right]\left[(X')^{-1}X'\right] = I \cdot I = I_{n\times n} = I_{p\times p},$$

so $\hat{Y} = HY = I_n Y = Y$ and all points are on the regression line and $\sigma^2 = 0$.