Statistics HW3, Programming (Changheng Chen)

Problem 1: "NASDQ"

(A) Construct a full regression model.

See code.

(B) Write down the selected model.

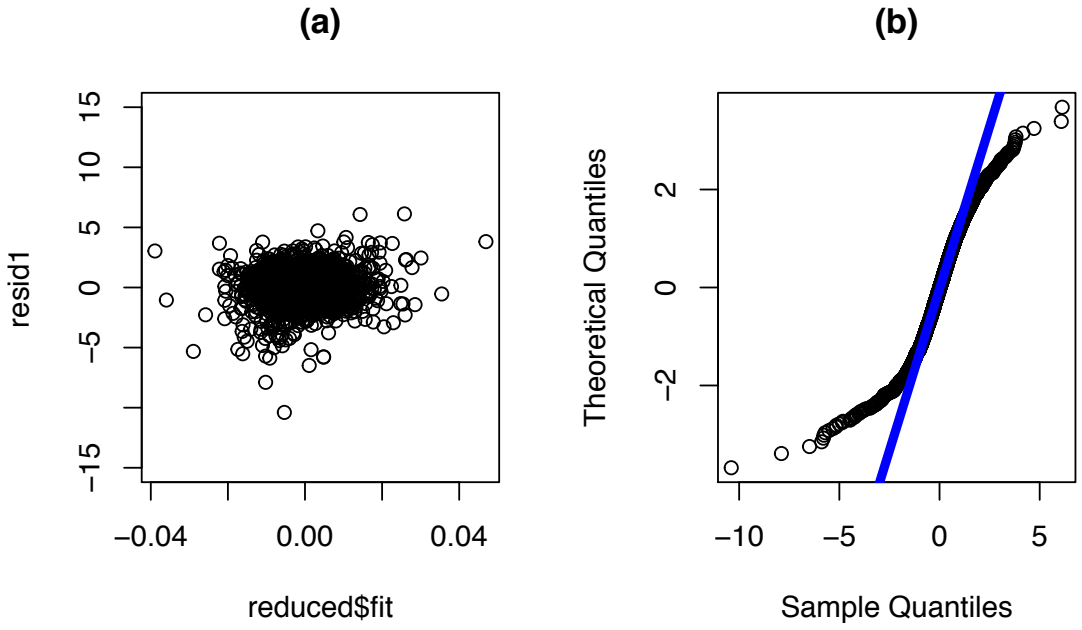
Nasdaq ~ AAPL + ALTR + AMAT + AMD + AMGN + AOC + APD + ASH +
AT + AXP + BA + BBY + BC + C + CCL + CEG + CFC + COST + DD +
DELL + DIS + ED + EDS + FRX + GE + GPS + HON + HPQ + HSY +
HUM + IBM + INTC + IPG + JNJ + JPM + KBH + KMI + KO + KR +
LEG + LEN + LM + LUV + MAS

(C) Compare the full and selected models. Summarize your comparison in an ANOVA table.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Selected	4241	0.035751				
Full model	4203	0.035353	38	0.0003975	1.2436	0.1457

The sum of squared residual for the selected model is only slightly larger than that of the full model. Thus, the selected model is good enough for NASDAQ log return, considering its significantly smaller number of predictors.

(D) For the selected regression model in (B), perform residual diagnostics.



The plot of residual versus fitted value demonstrated that the residual is distributed normally. The QQ plot shows that the sample has heavier tails than a normal distribution.

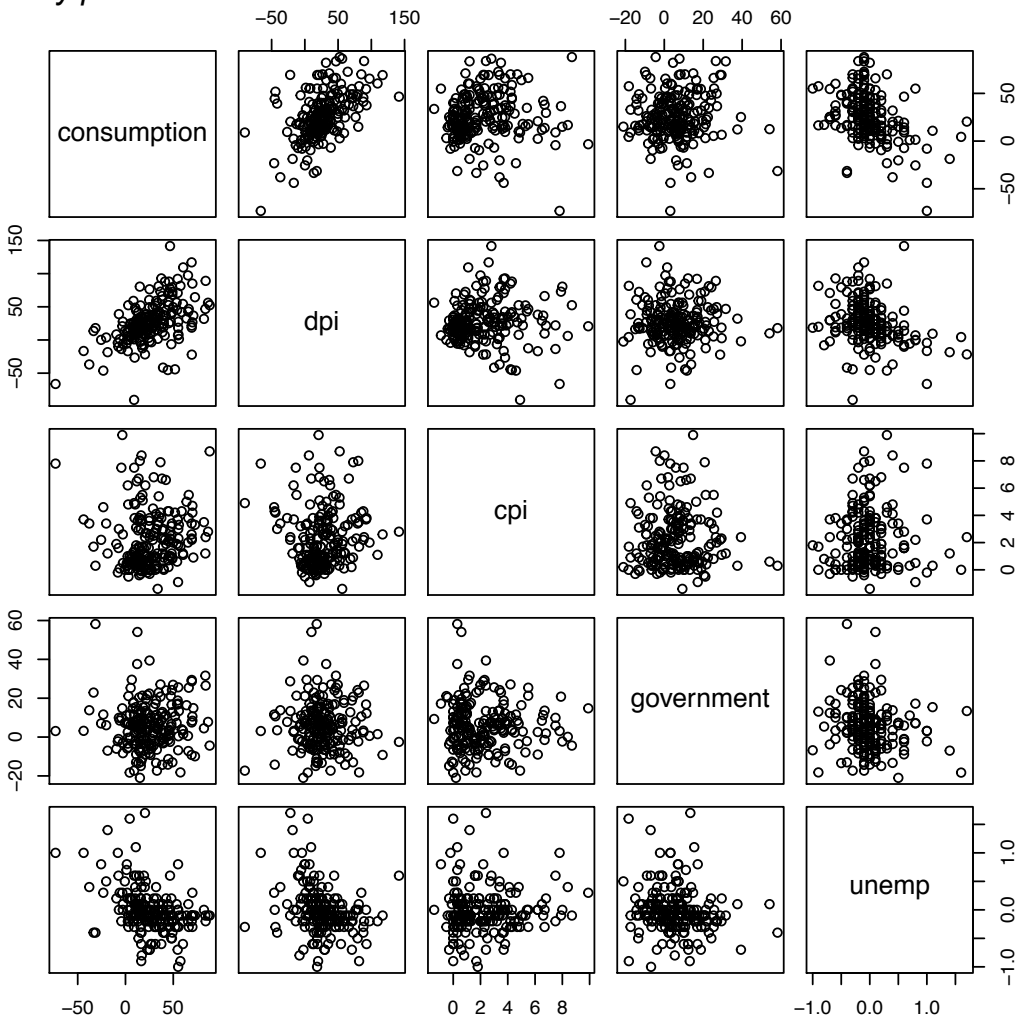
(E) If you can only use at most five stocks to track the daily NASDAQ log returns, describe your model selection procedure and your constructed model.

Using the "regusbsets" function in the "leaps" library with the exhaustive method as the selection method, we get

Nasdaq ~ ALTR + AMAT + GE + INTC + LM

Problem 2: Chapter 12 R-lab

1. Describe any interesting features, such as, outliers, seen in the scatterplot matrix. Keep in mind that the goal is to predict changes in consumption. Which variables seem best suited for that purpose? Do you think there will be collinearity problems?



By checking the scatter plots between “consumption” with other variables, we can notice that there are outliners in the consumption-cpi plot. The dpi, cpi and government seem to have linear relationships with consumption, and they are likely to be suitable for predicting the changes in consumption. There might be collinearity among the three variables.

2. From the summary, which variables seem useful for predicting changes in consumption?

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.752317	2.520168	5.854	1.97e-08 ***
dpi	0.353044	0.047982	7.358	4.87e-12 ***
cpi	0.726576	0.678754	1.070	0.286
government	-0.002158	0.118142	-0.018	0.985
unemp	-16.304368	3.855214	-4.229	3.58e-05 ***

Variables, dpi and unemployment rate, seem to be useful for predicting changes in consumption, since their corresponding Pr(>|t|) are significantly small.

3. For the purpose of variable selection, does the AOV table provide any useful information not already in the summary?

Response: consumption

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dpi	1	34258	34258	82.4294	< 2.2e-16 ***
cpi	1	253	253	0.6089	0.4361
government	1	171	171	0.4110	0.5222
unemp	1	7434	7434	17.8859	3.582e-05 ***
Residuals	198	82290	416		

The AOV table shows that the sum of squared residual (SSR) corresponding to unemp is smaller than that corresponding to dpi, meaning that dpi tends to explain a higher percentage of total sum of squares.

4. Which variables are removed from the model, and in what order?

First government, then cpi.

5. How much of an improvement in AIC was achieved by removing variables? Was the improvement huge? Is so, can you suggest why? If not, why not?

The AIC decreases from 1228.98 with all four variables as predictor to 1226.98 with “government” being removed, and it further decreases to 1226.15 with “cpi” being removed. The improvement is not huge; it may be due to multicollinearity among the variables.

6. Was there much collinearity in the original four-variable model? Was the collinearity reduced much by dropping two variables?

VIFs:

dpi	cpi	government	unemp
1.100321	1.005814	1.024822	1.127610

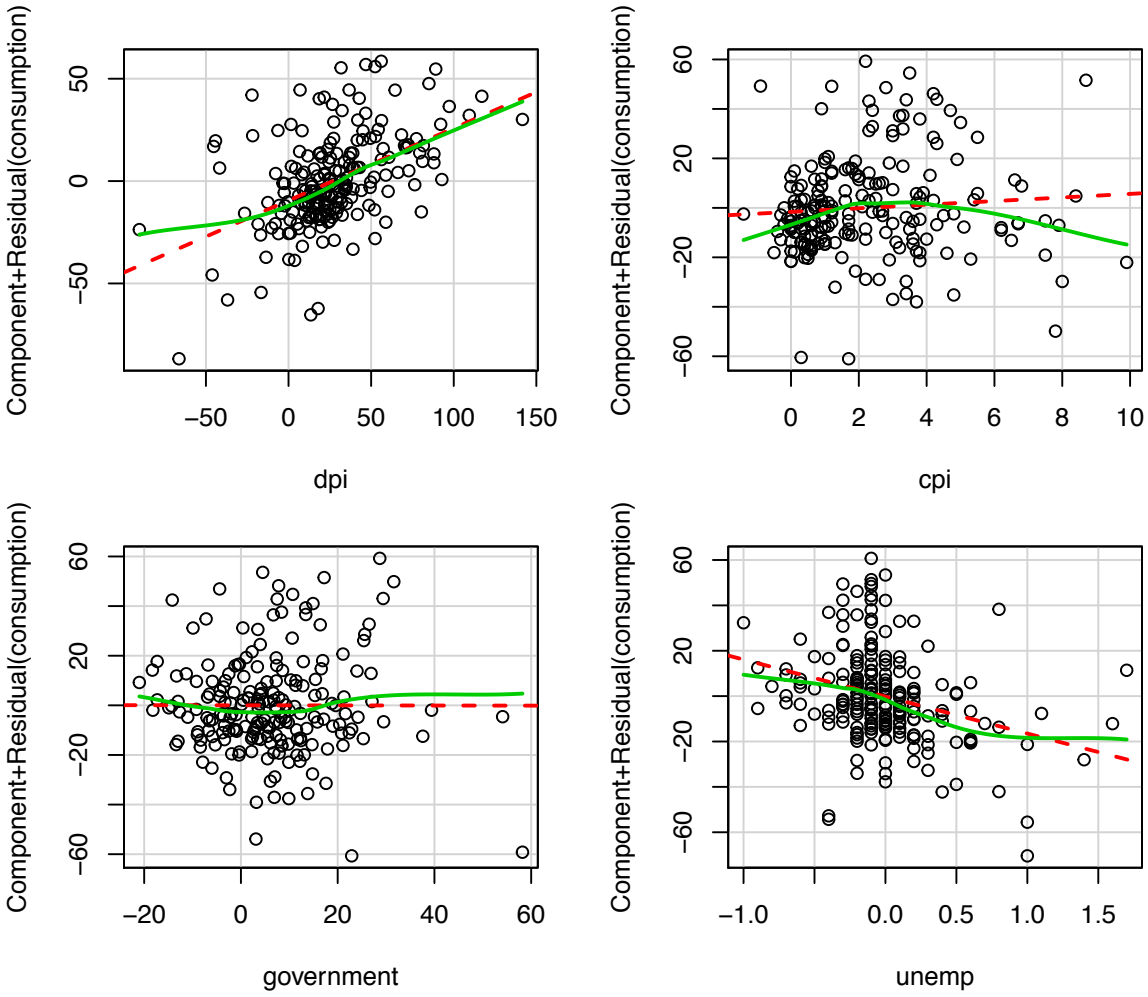
VIFs:

dpi	unemp
1.095699	1.095699

A high VIF value suggests a strong linear relationship between one variable and the rest. As shown above, all the VIF values are close to 1, indicating weak collinearity among the four variables. The VIF values for dpi and unemp decrease slightly in the model with dpi and unemp as predictors.

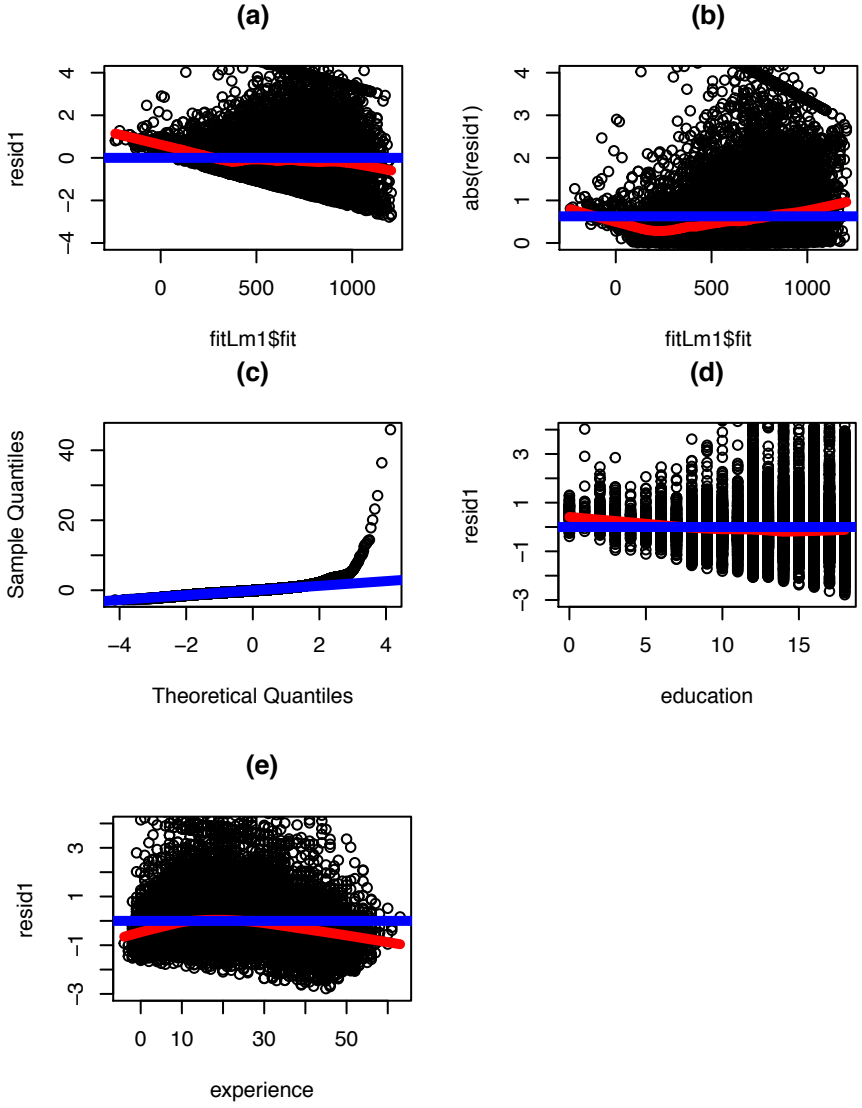
7. What conclusions can you draw from the partial residual plots?

The plots show that variables of dpi, government and unemployment rate have an explanatory power, as their partial residuals have linear trends. While the curve of cpi deviates from the lease squares line, suggesting that cpi has a nonlinear effect on consumption.



Problem 3: Problem 13.1

For each of the panels (a)–(e) in the figure you have just created, describe what is being plotted and any conclusions that should be drawn from the plot. Describe any problems and discuss how they might be remedied.



Plot (a) is residual versus fitted value. It shows that the variance of residual is large at the tail of large fitted values, indicating heteroskedasticiy. Log transformation or polynomial regression may help alleviate this problem.

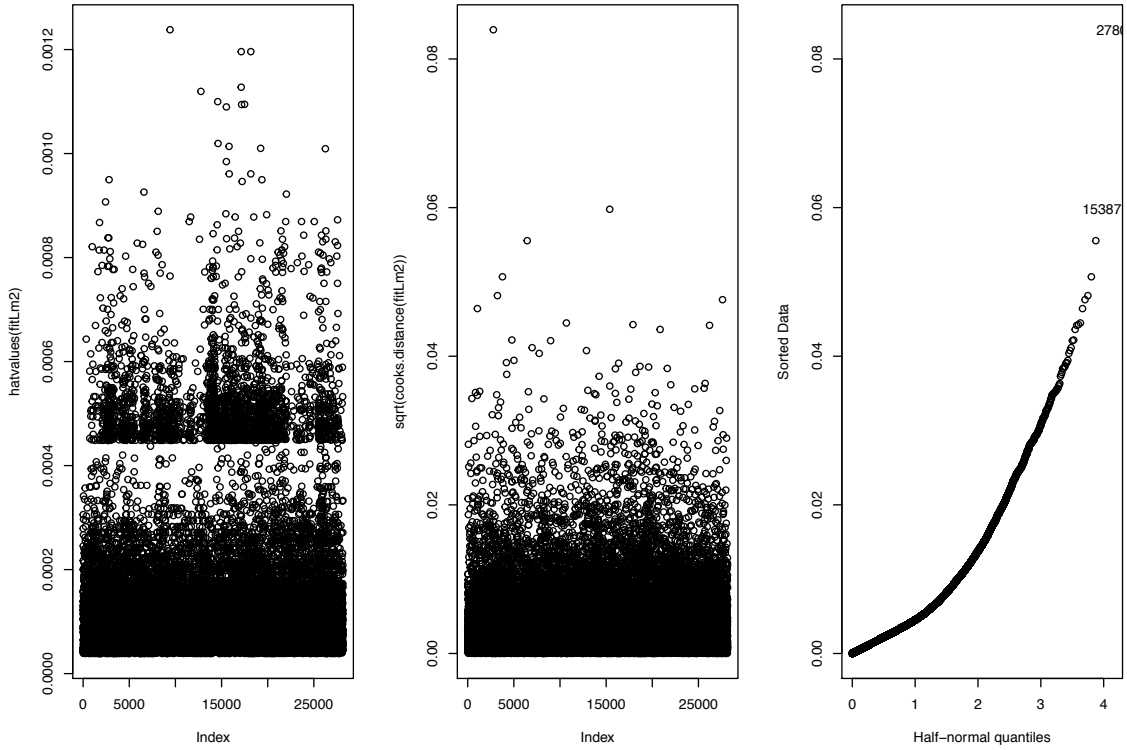
Plot (b) is absolute residual versus fitted value. Same as in (a), the absolute residuals are centered at the large fitted values.

Plot (c) is QQ plot. The data are right skewed as indicated by the convex curve.

Plot (d) is residual versus education. The residual shows large variances when education is at a high level, indicating heteroskedasticity. A transformation on education variable can help.

Plot (e) is residual versus experience. It seems to be normally distributed and has no obvious problem.

Problem 4: Problem 13.5



Point 2780 has low leverage but is a residual outlier. Point 15387 has high leverage and is not a residual outlier.