Assignment 8

Gilbert Watson

Friday, December 6th, 2013

Contents

1 10.6

Refer to Grocery retailer Problem 6.9:

a) Fit regression model (6.1) to the data using X_1 and X_2 only.

b) Prepare an added-variable plot for each of the predictor variables X_1 and X_2.

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

d) Obtain the fitted regression function in part(a) by separately regressing both Y and X_2 or X_1, and then regressing the residuals in an appropriate fashion.

1.1 Answer:

a) First, let's read in the data and estimate a simple model:

```
> library(xtable)
> grocery <- read.table(file="10.6.txt")
> names(grocery) <- c("labor.hours","cases","pct.labor","holiday")
> grocery.fit <- lm(labor.hours~cases+pct.labor,data=grocery)
> grocery.fitsum <- summary(grocery.fit)
> print(xtable(grocery.fitsum))
```

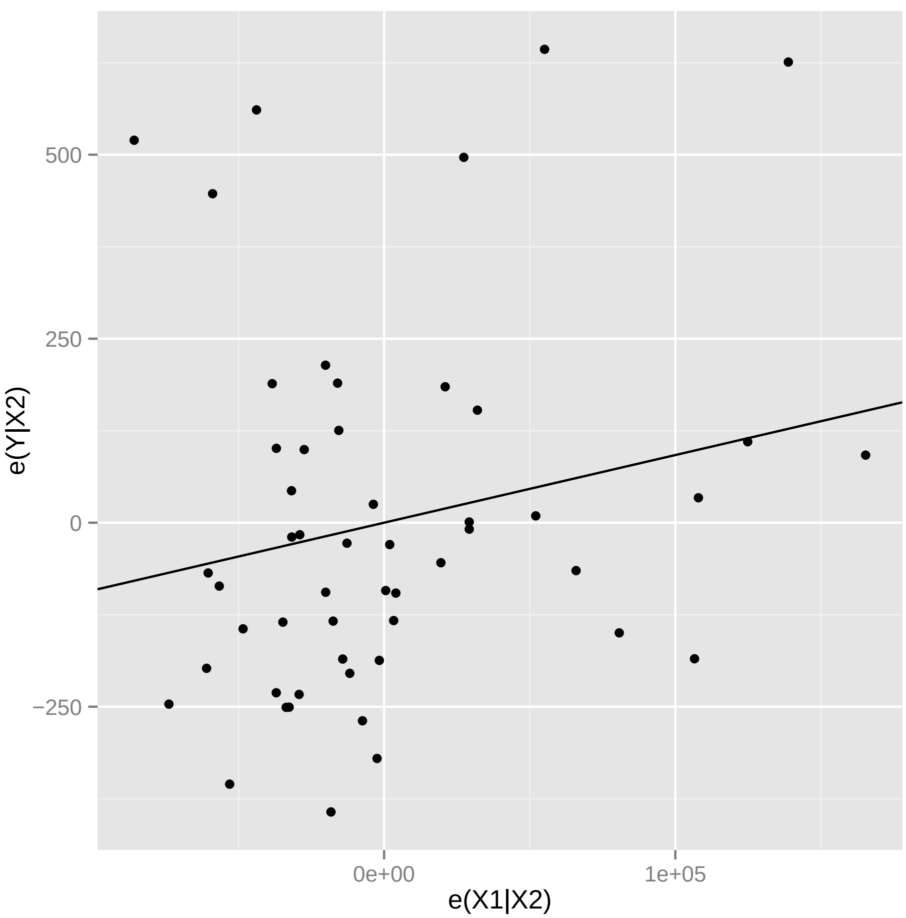
b) To make added variable plots we must first estimate several models:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3995.4787	337.7660	11.83	0.0000
cases	0.0009	0.0006	1.46	0.1517
pct.labor	12.1205	39.7656	0.30	0.7618

```
> yonx1 <- lm(labor.hours~cases,data=grocery)
> yonx2 <- lm(labor.hours~pct.labor,data=grocery)
> x1onx2 <- lm(pct.labor~cases,data=grocery)
> x2onx1 <- lm(cases~pct.labor,data=grocery)
```

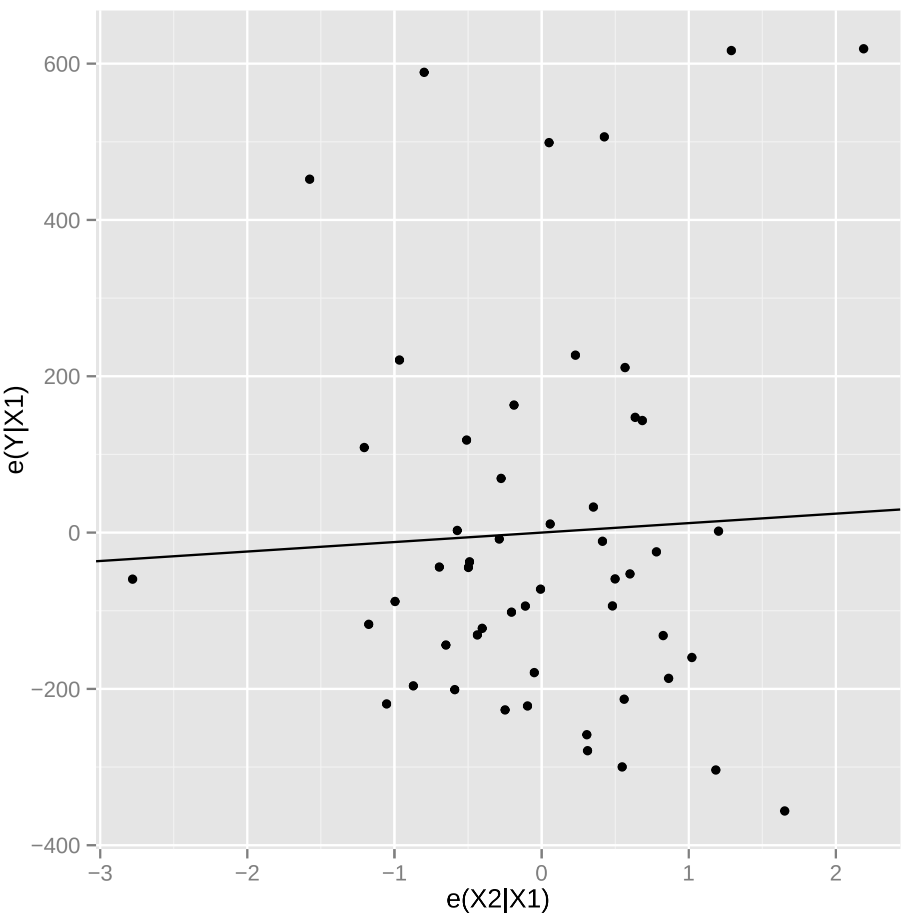
Now we can plot residuals against each other:

```
> library(ggplot2)
> a <- qplot(x=x2onx1$residuals,
+        y=yonx2$residuals,
+        xlab="e(X1|X2)",
+        ylab="e(Y|X2)")
> a + geom_abline(intercept=0,slope=grocery.fit$coefficients[2])
```




```
> b <- qplot(x=x1onx2$residuals,
+        y=yonx1$residuals,
```

```
+       xlab="e(X2|X1)",
+       ylab="e(Y|X1)")
> b + geom_abline(intercept=0,slope=grocery.fit$coefficients[3])
```



c) The added variable plots suggest that the model is not aided by the addition of X_2. While the slope of the line in the first plot is quite different than a horizontal line, indicating that X_1 should probably be added to a model already containing X_2, the line for the second is almost horizontal, indicating that X_2 should not be added to a model already containing X_1.

d) First we must regress the residuals from above through the origin:

```
> originreg <- lm(yonx2$residuals~x2onx1$residuals-1)
> originreg$coefficients

x2onx1$residuals
    0.0009191639
```

$$\epsilon(Y|\hat{X_2}) = 0.000919163915830026[\epsilon(X_1|X_2)]$$

$$[\hat{Y} - (4237.4687511802 + 17.036618933466X_2)] = 0.000919163915830026[X_1 - (263271.957683668 + 5348.44939715936X_2)]$$

$$\hat{Y} = 3995.47866762744 + 0.000919163915830026X_1 + 3995.47866762744X_2$$

2 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 $\alpha = 0.1$. State the decision rule and conclusion.

b) Obtain the diagonal elements of the hat matrix. Identify any 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 ocupancy rate is 0.05, and whose square footage is 350,000. Use (10.29) to determine whether this estimate will involve a hidden extrapolation.

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?

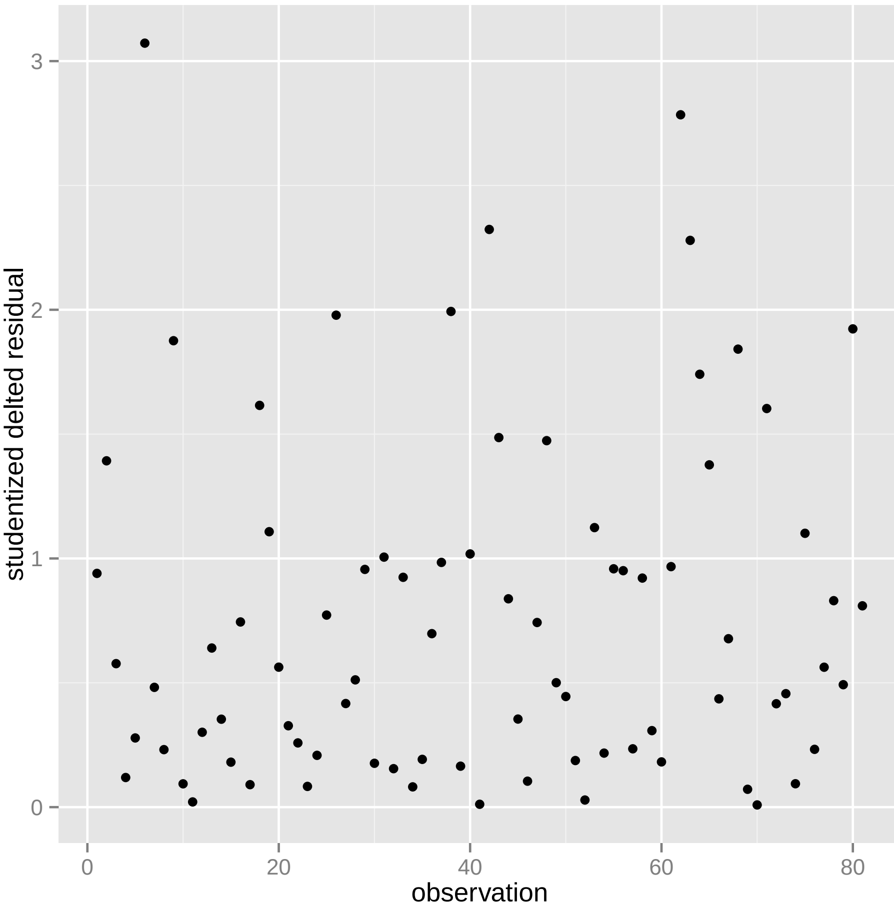
2.1 Answer:

a) First let's read in the data and estimate a model:

```
> commProps <- read.table(file="10.12.txt")
> names(commProps) <- c("rental.rates","age","expenses","vacancy","sqft")
> commProps.fit <- lm(rental.rates~age+expenses+vacancy+sqft,data=commProps)
```

Now let's compute the studentized delted residuals and identify potential outliers

```
> library(MASS)
> stud.del.res <- studres(commProps.fit)
> qplot(seq(1,length(stud.del.res),1),abs(stud.del.res),
+       xlab="observation",
+       ylab="studentized delted residual")
```



Now let's calculate the Bonferroni critical value for outliers:

```
> alpha <- 0.1
> critical.t <- qt(1-alpha/(2*length(stud.del.res)),
+                  length(stud.del.res)-
+                    length(commProps.fit$coefficients))
> paste0("The critical value is t = ",critical.t)

[1] "The critical value is t = 3.35660231603714"

> paste0("Any studentized residuals > ",critical.t," ?  ",
+        any(abs(stud.del.res) > critical.t))

[1] "Any studentized residuals > 3.35660231603714 ?  FALSE"
```

There are no outliers in the data by the bonferroni test. There aren't. I know what the next part of the question says and there aren't

b) Let's find the diagonals of the hat matrix, use it to find outliers, and print the observation number:

```
> diagonal <- lm.influence(commProps.fit)$hat
> threshold <- 2*(length(commProps.fit$coefficients)-1)/length(commProps$age)
```

```
> outliers <- as.vector(which(diagonal > threshold))
> paste(c("The outlier observations are:",outliers),collapse=" ")
```

```
[1] "The outlier observations are: 3 8 9 43 53 54 61 65 80"
```

c) First let's compute X_{new}:

```
> xnew <- c(10,12.00,0.05,350000)
> xold <- as.matrix(commProps.fit$model[,setdiff(names(commProps.fit$model),
+                                        c("rental.rates"))])
```

Then let's find $h_{new,new}$ and see if it is greater than the threshold value:

```
> hnewnew <- t(xnew)%*%solve(crossprod(xold))%*%xnew
> hnewnew
```

```
           [,1]
[1,] 0.05203178
```

$h_{new,new}$ is well within the range of the other diagonal entires of $\mathbf{H}$. It's prediction would not involve hidden extrapolation.

d) Let's find DFFITS, DFBETAS, and Cook's D for the following observations and make a table:

```
> library(xtable)
> weirdos <- c(61,8,3,53,6,62)
> diagnostics <- data.frame(DFFITS=dffits(commProps.fit)[weirdos],
+                           DFBETAS=dfbetas(commProps.fit)[weirdos],
+                           Cooks.D=cooks.distance(commProps.fit)[weirdos])
> diagnostics$`DFFITS > 1` <- abs(diagnostics$DFFITS) > 1
> diagnostics$`DFBETAS > 1` <- abs(diagnostics$DFBETAS) > 1
> diagnostics$`F Percentile` <- pf(diagnostics$Cooks.D,
+                                  4,
+                                  length(commProps$age)-4)
> print(xtable(diagnostics))
```

	DFFITS	DFBETAS	Cooks.D	DFFITS > 1	DFBETAS > 1	F Percentile
61	0.64	-0.06	0.08	FALSE	FALSE	0.01
8	0.12	-0.01	0.00	FALSE	FALSE	0.00
3	-0.28	-0.23	0.02	FALSE	FALSE	0.00
53	0.53	-0.02	0.05	FALSE	FALSE	0.01
6	-0.87	0.20	0.14	FALSE	FALSE	0.03
62	0.69	0.28	0.09	FALSE	FALSE	0.01

From our table it is clear that none of the appearant outliers are influential. DFFITS and DFBETAS are all less than 1 for these values, and Cook's D for these values is in a very low percentile of the associated F distribution.

3 10.20

Refer to Lung pressure Problems 9.13 and 9.14. The subset regression model containing first-order terms for X_1 and X_2 and the cross-product term X_1X_2 is to be evaluated in detail.

a) Obtain the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

b) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .05$. State the decision rule and conclusion.

c) Cases 3, 8, imd 15 are moderately far outlying with respect to their X values, and case 7 is relatively far outlying with respect to its Y value. Obtain DFFITS, DFBETAS, and Cook's distance values for these cases to assess their influence. What do you conclude?

3.1 Answer:

a) Let's first read in the data and estimate the model:

```
> Lung <- read.table(file="10.20.txt")
> names(Lung) <- c("arterial.pressure","emptying.rate","ejection.rate","blood.gas")
> Lung.fit <- lm(arterial.pressure~emptying.rate*ejection.rate,
+                 data=Lung)
```

Now let's obtain the variance inflation factors and $V\bar{I}F$:

```
> library(car)
> vif <- vif(Lung.fit)
> bar.vif <- mean(vif)
> vif

              emptying.rate               ejection.rate
                   5.431477                   11.639560
emptying.rate:ejection.rate
                  22.474469

> bar.vif

[1] 13.18184
```

Yes, multicollinearity is an issue. All VIF values are above 1 and the interaction term's value is far above the mean VIF.

b) Are there any studentized deleted residuals greater than the critical value?

```
> stud.del.res <- studres(Lung.fit)
> alpha <- 0.05
> critical.t <- qt(1-alpha/(2*length(stud.del.res)),
+                  length(stud.del.res)-
+                    length(Lung.fit$coefficients))
> paste0("The critical value is t = ",critical.t)

[1] "The critical value is t = 3.59890216225008"

> paste0("Any studentized residuals > ",critical.t," ?  ",
+        any(abs(stud.del.res) > critical.t))

[1] "Any studentized residuals > 3.59890216225008 ?  FALSE"
```

The decision rule is that if the studentized deleted residual is greater than the critical value, 3.59890216225008, then we reject the null hypothesis that the outcome value is not an outlier. No studentized deleted residuals are greater than this critical value in this case, so we conclude that none of the Y values are outliers.

c) Let's build a table to assess observation influence:

```
> weirdos <- c(3,8,15,7)
> diagnostics <- data.frame(DFFITS=dffits(Lung.fit)[weirdos],
+                           DFBETAS=dfbetas(Lung.fit)[weirdos],
+                           Cooks.D=cooks.distance(Lung.fit)[weirdos])
> diagnostics$`DFFITS > 1` <- abs(diagnostics$DFFITS) > 1
> diagnostics$`DFBETAS > 1` <- abs(diagnostics$DFBETAS) > 1
> diagnostics$`F Percentile` <- pf(diagnostics$Cooks.D,
+                                  3,
+                                  length(Lung$arterial.pressure)-3)
> print(xtable(diagnostics))
```

	DFFITS	DFBETAS	Cooks.D	DFFITS > 1	DFBETAS > 1	F Percentile
3	-0.68	-0.65	0.12	FALSE	FALSE	0.05
8	-4.78	-1.55	4.99	TRUE	TRUE	0.99
15	0.17	-0.02	0.01	FALSE	FALSE	0.00
7	1.75	1.45	0.46	TRUE	TRUE	0.29

We can conclude from this table that observations 8 and 7 are unduly influential to the model. They both have DFFITS and DFBETAS greater than 1 and Cook's D's that have a high percential ranking.

4 10.23

Show that (10.37) is algebraically equivalent to (1O.33a).

4.1 Answer:

5 System Information

```
> sessionInfo();

R version 3.0.1 (2013-05-16)
Platform: x86_64-apple-darwin10.8.0 (64-bit)

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] car_2.0-19      MASS_7.3-26     ggplot2_0.9.3.1 xtable_1.7-1

loaded via a namespace (and not attached):
```

```
 [1] colorspace_1.2-2   dichromat_2.0-0    digest_0.6.3       grid_3.0.1
 [5] gtable_0.1.2       labeling_0.2       munsell_0.4.2      nnet_7.3-6
 [9] plyr_1.8           proto_0.3-10       RColorBrewer_1.0-5 reshape2_1.2.2
[13] scales_0.2.3       stringr_0.6.2      tools_3.0.1
```