Problem Set 5 Solution

Math 430, Winter 2017

Problem 1

The data in website.csv are from an observational study of productivity of developers at a website development company. The eventual goal of the analysis is to "determine which variables have the greatest impact on the number of websites delivered." Consider the model

$$DELIVER_i = \beta_0 + \beta_1 BACKLOG_i + \beta_2 EXPERIENCE_i + \beta_3 PROCESS_i + \beta_4 YEAR_i + e_i$$

Fit this regression model, then consider the following questions. Please consider to be a "medium size" sample.

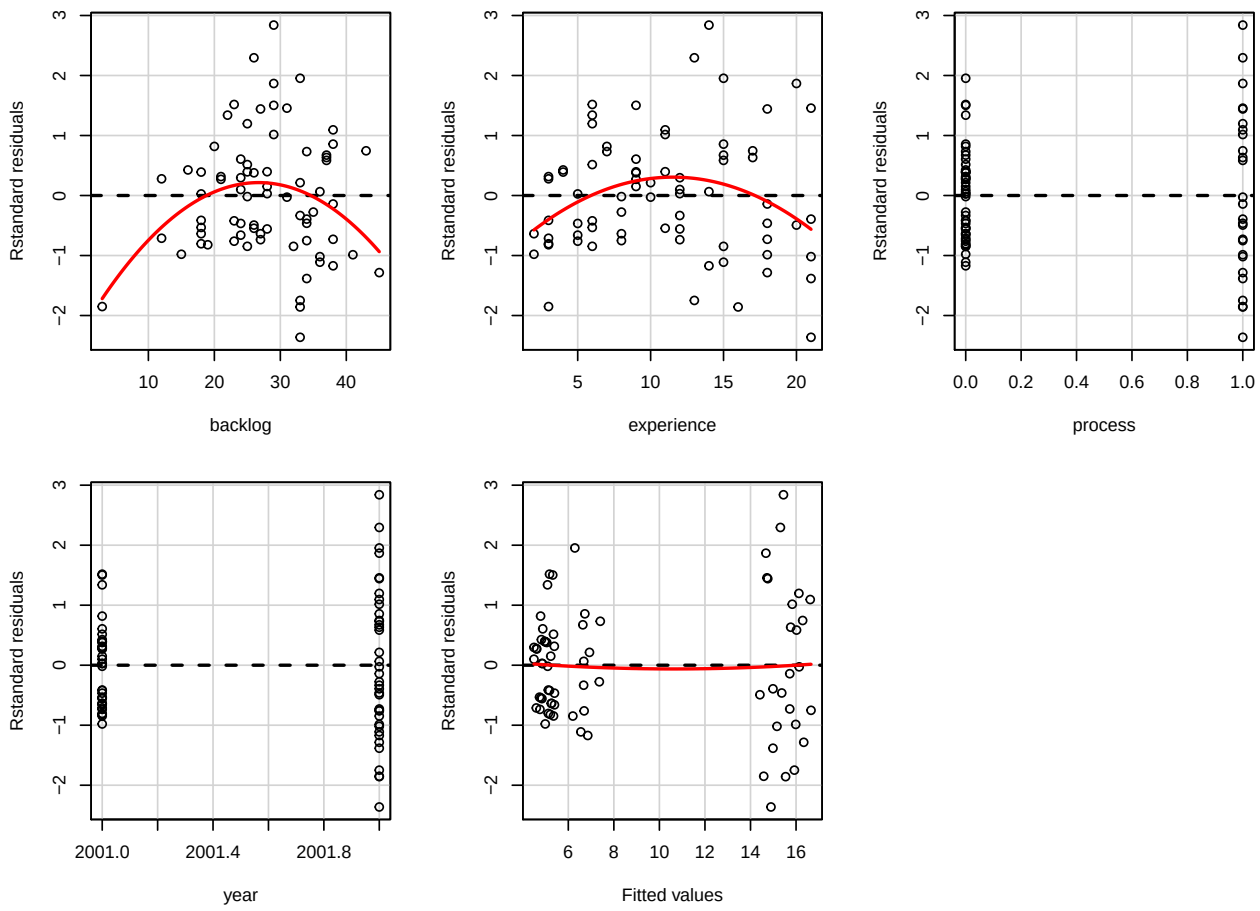
(a)

Consider each point. Do you have any concerns about regression outliers? If so, list the points (by id #) that are a concern and briefly explain why.

```
# Import the data
website <- read.csv("https://raw.githubusercontent.com/math430-lu/data/master/website.csv")

# Fit the model
web_mod <- lm(deliver ~ backlog + experience + process + year, data = website)

# Create standardized residual plot
library(car)
residualPlots(web_mod, layout = c(2, 3), type = "rstandard", tests = FALSE)
```




```
rstandard(web_mod)[abs(rstandard(web_mod)) > 2]
```

```
##        21        46        50
## -2.362750  2.295303  2.840130
```

The best answer: no problems with regression outliers.

- You should be looking at standardized residuals. Plotting these against the predicted values or the id number indicates some large values, but no values far from the general pattern.
- Here, there are 3 observations with standardized residuals larger than 2 (or < -2). These may be worth quickly investigating, especially if one is very large (or small), but you expect 73*0.05 = 3.65 points to fall outside the 0.025 and 0.975 normal quantiles, even if there are no outliers.

(b)

Do any points raise concerns about unusually large influence on the fitted values? If so, list the id #s that are of concern and briefly explain why.

```
inf_stats <- influence.measures(web_mod)

# print obs. identified as influential w.r.t DFFITS
which(abs(inf_stats$infmat[, "dffit"]) > 1)
```

```
## 65
## 65
```

```
# prints obs. identified as influential w.r.t. Cook's D
which(inf_stats$is.inf[, "cook.d"])
```

```
## named integer(0)
```

Yes, id 65 has a large DFFITS value (larger than 1), indicating that id 65 has a large influence on its own fitted value. The 0.5 quantile of the $F_{5,68}$ distribution is 0.879 (`qf(.5, 5, 68)`). Values of Cook's D less than this are no concern. There are no points with a Cook's D larger than 0.879.

- I told you to consider this a small-medium data set. If it were large, you would use a smaller critical value of DFFITS to identify influential points.
- If you told me about points with high leverage (h_{ii}) , that was fine but not necessary. Leverage is potential influence. The question asked about points with large influence.

(c)

Since the objective here is to examine regression coefficients, do any points have unusually large influence on the estimated regression coefficients? Again, if so, list the points (by id #) that are a concern and briefly explain why.

Yes, `id` 65 has a large influence on the slope for backlog because it has a DFBETA > 1. It is the only point with a DFBETA $<$ -1 or > 1.

```
# One way to pull off influential dfbetas
which(apply(inf_stats$is.inf[, 1:5], 1, any))
```

```
## 65
## 65
```

(d)

Do you have any concerns about multicollinearity? Explain why or why not.

No, the VIF values for each variable are all considerably less than 10.

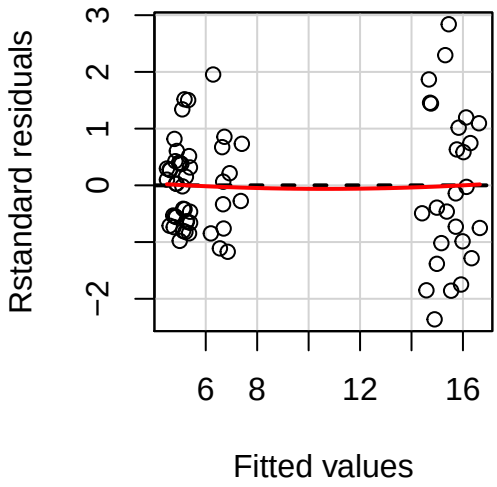
```
vif(web_mod)
```

```
##    backlog experience    process       year
##   2.984613   3.292191   2.478256   3.054964
```

(e)

Plot the standardized residuals vs. predicted values. Are there any issues that concern you? Explain why or why not.

```
residualPlot(web_mod, type = "rstandard")
```

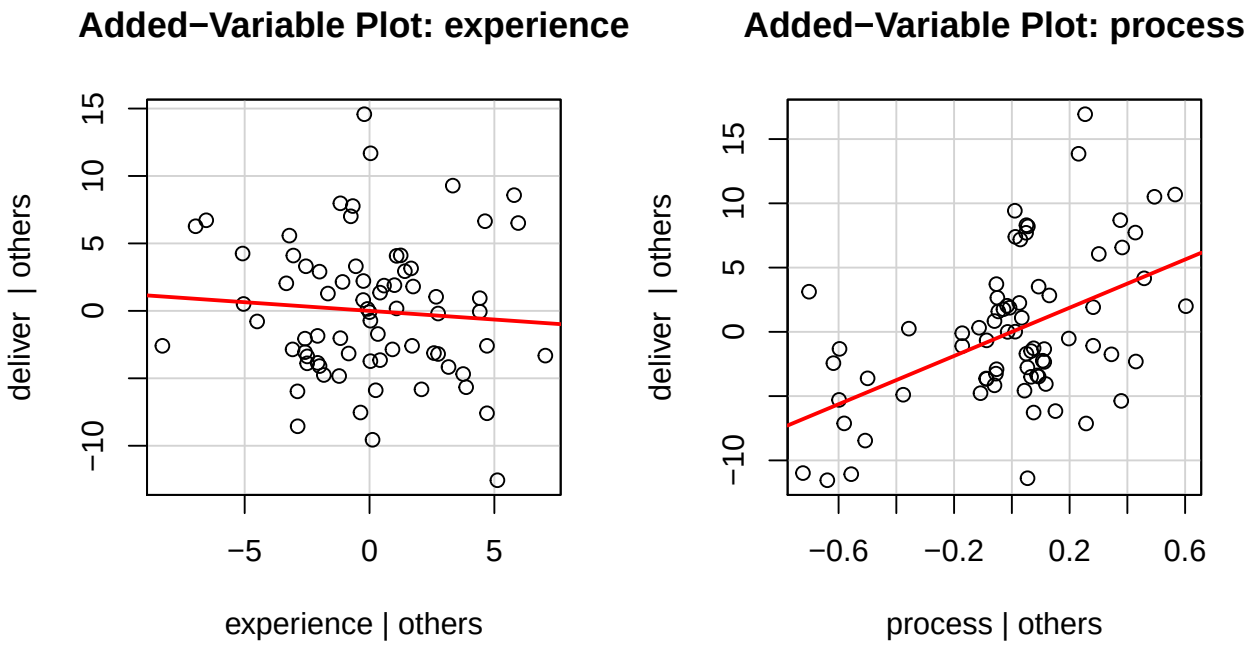


There is no sign of an unusual regression residual. The variability does seem to increase slightly with the mean, but the s.d.'s in the two groups are quite similar. I would probably assume equal variances. If I did transform, I would make very sure that the transformed values had no problems. The log transform is probably too strong here. The residual plot has two clusters of observations, which to the two PROCESS groups.

(f)

Look at the added variable plots for experience and for process. If the error variance is associated with experience, the experience added variable plot will show the same spreading pattern seen in plots of standardized residuals vs. predicted values. Does it seem that the error variance is related to experience? to process?

```
avPlot(web_mod, variable = "experience")
avPlot(web_mod, variable = "process")
```




```
## null device
##           1
```

It appears that the process added variable plot has the same spreading pattern seen in the plots of residuals vs. predicted values. This indicates that error variance is associated with process. The same pattern is seen less strongly with experience.

(g)

Use the Breusch-Pagan test to test whether there is an association between error variance and experience.

```
ncvTest(web_mod, ~ experience)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ experience
## Chisquare = 6.990111    Df = 1     p = 0.008196127
```

The p-value of the Breusch-Pagan test of an association between error (residual) variance and experience is 0.008; thus, there is evidence that the variance changes with experience.

(h)

Use the Breusch-Pagan test to test whether there is an association between error variance and process.

```
ncvTest(web_mod, ~ process)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ process
## Chisquare = 13.95918    Df = 1     p = 0.0001868234
```

The p-value of the Breusch-Pagan test of an association between error (residual) variance and process is 0.0002. There is a strong association between error variance and process.

Problem 2

The data in avp.csv were constructed to make some points about assessing the form of f(X) in a multiple regression. The data set has three variables, `x1`, `x2` and `x3`, that are to be used to predict the response `y`. All of the below questions are based on the model:

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + e_i$$

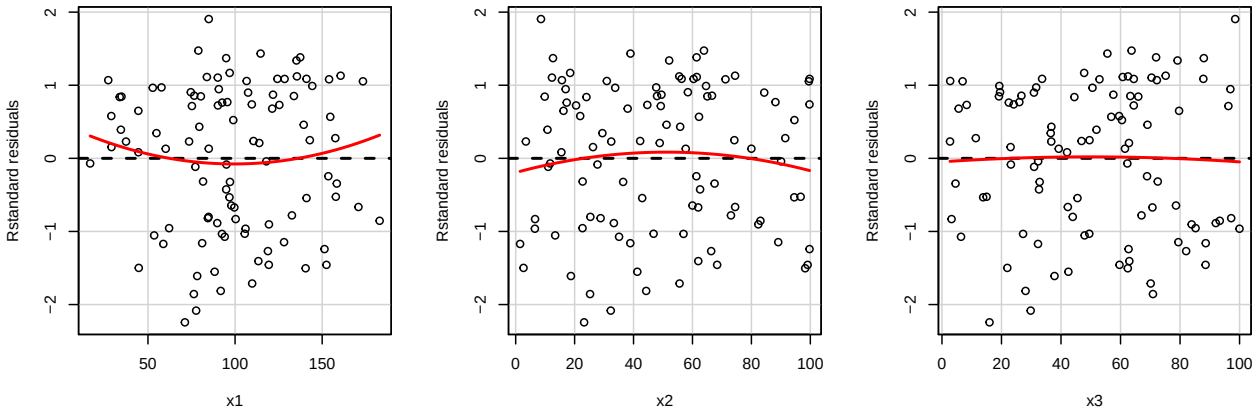
The investigators who collected these data are especially concerned about nonlinearity.

(a)

Fit the regression model and plot the standardized residuals against each of the predictor variables. Is there any indication of nonlinearity? Explain why or why not.

```
avp <- read.csv("https://github.com/math430-lu/data/raw/master/avp.csv")
avp_mod <- lm(y ~ ., data = avp)
```

```
residualPlot(avp_mod, type = "rstandard", variable = "x1")
residualPlot(avp_mod, type = "rstandard", variable = "x2")
residualPlot(avp_mod, type = "rstandard", variable = "x3")
```




```
## null device
##           1
```

No sign of lack of fit: all plots are a flat band of points.

(b)

Fit a full quadratic (i.e. including $x1^2$, $x2^2$, $x3^2$, $x1x2$, $x1x3$, and $x2x3$ in the model). Use the results from this model and the original regression to test for lack of fit (i.e. nonlinearity) using a partial F-test. Is there any evidence of lack of fit?

```
quad_mod <- lm(y ~ x1 * x2 * x3 + I(x1^2) + I(x2^2) + I(x3^2) - x1:x2:x3, data = avp)
anova(avp_mod, quad_mod)
```

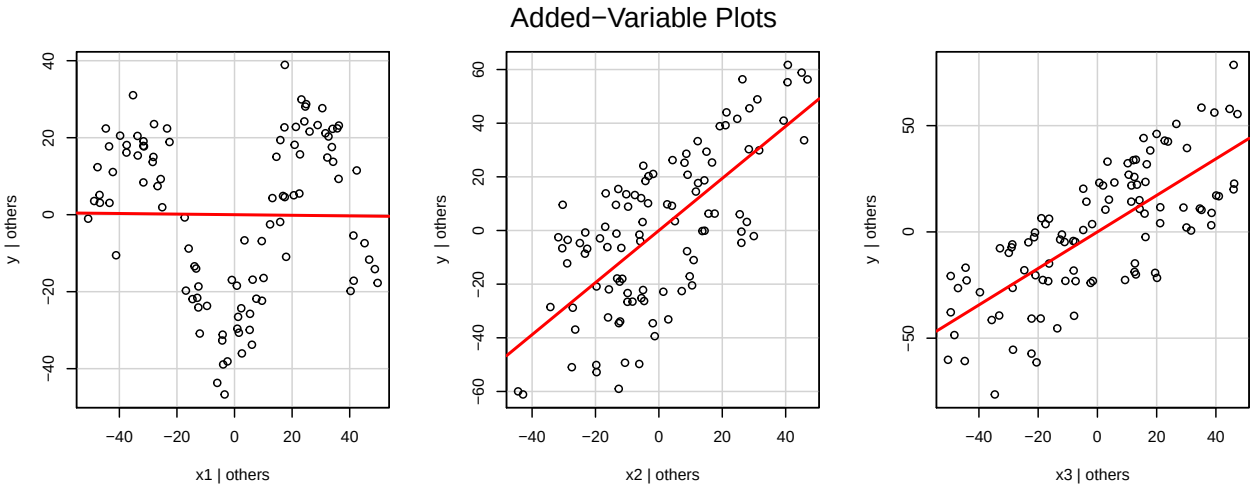
```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x1 * x2 * x3 + I(x1^2) + I(x2^2) + I(x3^2) - x1:x2:x3
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     96 43206
## 2     90 32898  6     10308 4.6999 0.0003329 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the partial F-test comparing the linear and quadratic models is 0.0008, providing strong evidence of lack of fit.

(c)

Examine added variable plots for the original model for each variable. Is there any indication of lack of fit? Explain your answer.

```
avPlots(avp_mod, layout = c(1,3))
```



There is lack of fit in X1. The dependence of Y on X1 is oscillatory. The dependence on X2 is reasonably linear and the dependence on X3 is reasonably linear.

Problem 3

For each of the following, give an appropriate regression model and carefully define the X variables in your model (including indicator variables).

Some examples are looking for estimates. For these, indicate how the desired quantities could be estimated from the regression parameters. You do not need to worry about standard errors or inference.

Other parts are looking for a test. For these, indicate how you would construct that test. Your answer could be "a t-test of (indicate a regression parameter or linear combination of regression parameters) = 0 (or other value)." It could be "an F test comparing (indicate a pair of models)" or it could be something else. I do not need formulas for the test statistics.

(a)

A study is comparing the energy content of constant-sized pieces of firewood from different tree species. If you are burning wood to heat a room or a house, a higher energy content is a good thing. One complication is that the energy released depends on the moisture content of the firewood, which is hard to standardize. You have studied three species (Red Oak, White Pine and Black Walnut). You believe that the relationship between energy content and moisture is linear with the same slope for each species. You want to estimate the difference in energy content at 10% moisture content between White Pine and Red Oak.

Variable	Definition
oak	1 if red oak, 0 otherwise
pine	1 if white pine, 0 otherwise
moisture	percent moisture

Model: $Energy = \beta_0 + \beta_1 oak + \beta_2 pine + \beta_3 moisture + e$

Desired quantity: $\beta_1 - \beta_2$

(b)

Consider the above again, except now you wish to test the null hypothesis that the three species have the same energy content at 10% moisture. Again, assume that all three species have the same slope.

X variables and model as in part (a).

Desired test: $\beta_1 = \beta_2 = 0$, full model as in (a); reduced model is $Energy = \beta_0 + \beta_3 moisture + e$

(c)

Consider the above again, except that now you assume that the three species have different slopes (for the association of moisture content on energy). You want to estimate the difference in energy content at 10% moisture between White Pine and Red Oak.

X variables and model as in part (a).

Model: $Energy = \beta_0 + \beta_1 oak + \beta_2 pine + \beta_3 moisture + \beta_4 oak * moisture + \beta_5 pine * moisture + e$

Desired quantity: $\beta_1 - \beta_2 + 10(\beta_4 - \beta_5)$

(d)

Assume that athletic performance for males and females in a certain sport can be described by quadratic functions of age. You are willing to assume that the curvature (β_2) is the same for both. You wish to test whether the age of maximum performance is the same for males and females. Hint: The intercepts (β_0) are probably not the same for males and females.

Define: male = 1 if male, 0 if female.

Model: $Y = \beta_0 + \beta_1 male + \beta_2 age + \beta_3 male * age + \beta_4 age^2 + e$

Test: t-test of $\beta_3 = 0$

(e)

Education researchers are studying whether watching television impacts the performance of graduate students. For each student in a class, they have Y, the exam score, and X, the number of hours spent watching television during the week prior to the exam. They assume that the relationship between Y and X is linear up to 20 hours. After X=20 hours, there is no relationship, i.e. the slope is 0 for X>20. They wish to estimate the slope for 0 to 20 hours and the expected difference between light television watching (3 hours) and heavy watching (25 hours).

There are a couple of ways to write this

Define: hours = # hours watching TV per week; $X_1 = 1$ if hours < 20, 0 otherwise

Model: $Y = \beta_0 + \beta_1 X_1 hours + \beta_1(1 - X_1)(20) + e$

Or, define $X = hours$ if $hours < 20$ and $X = 20$ if $hours \geq 20$

Model: $Y = \beta_0 + \beta_1 X + e$

Desired quantities: β_1 and $(20 - 3)\beta_1$.