Homework Five

Statistics 151a (Linear Models)

Due on 18 November 2015

06 November, 2015

1. In the Bodyfat dataset, consider the linear model

$$\text{BODYFAT} = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{WEIGHT} + \beta_3 \text{HEIGHT} + \beta_4 \text{THIGH} + e$$

In R, plot the following graphs (**9 points = one for each graph**)

a) Residuals against fitted values.

b) Standardized Residuals against fitted values.

c) Residuals against Standardized Residuals.

d) Predicted residuals against fitted values.

e) Residuals against predicted residuals.

f) Residuals against leverage.

g) Predicted residuals against Standardized Predicted Residuals.

h) Standardized residuals against Standardized Predicted residuals.

i) Cooks Distance against the ID number of the subjects.

Comment on these plots. Based on these plots, assess whether there are any outliers in the dataset; are there any influential observations. (**2 points**)

For each subject, calculate the p-value for testing whether the ith subject is an outlier based on the standardized predicted residual. Plot these p-values against the ID number of the subjects. How may of these p-values are less than 0.05? Does it make sense to rule all such subjects as outliers? (**4 points**)

Based on the analysis, does it make sense to fit the linear model with any of the subjects removed? If not, why not? If so, which ones; and in this case, report the summary for the linear model with the subjects removed. (**3 points**)

2. Again for the body fat dataset, consider the following R code:

```
g = lm(BODYFAT ~ AGE + WEIGHT + HEIGHT + THIGH, data = bodyfat)
par(mfrow = c(2, 2))
plot(g)
```

When I run this code, R gives me four plots. Describe each of these plots and explain how to interpret them. (**8 points**)

3. Consider the linear model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$ where the errors $e_1, \ldots, e_n$ are i.i.d normal with mean zero and variance σ^2.

   a) Let $\hat{y}_i$ denote the ith fitted value and let $\hat{y}_{i(i)}$ denote the predicted response value for the ith subject without including the ith subject in the regression (in the notation used in class, $\hat{y}_{i(i)} = x_i^T \hat{\beta}_{[i]}$. Write the difference $\hat{y}_i - \hat{y}_{i(i)}$ in terms of the ith residual and the ith leverage. (**2 points**)

   b) Calculate the distribution of $\hat{y}_i - \hat{y}_{i(i)}$. (**3 points**)

   c) Can you obtain an unbiased estimator for σ^2 that is independent of $\hat{y}_i - \hat{y}_{i(i)}$? If yes, specify such an unbiased estimator. If no, explain why. (**3 points**).

4. I got the following linear model output for a dataset consisting of a response variable and three explanatory variables:

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9282 -1.3174  0.0059  1.3238  4.4560

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.056057   1.963805  -3.593 0.000481 ***
x1           3.058592   0.089442  34.196  < 2e-16 ***
x2          -5.763410   0.168072 -34.291  < 2e-16 ***
x3           0.000571   0.165153   0.003 0.997247
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.928 on 116 degrees of freedom
```

```
Multiple R-squared:  0.9546,Adjusted R-squared:  0.9535
F-statistic:   814 on 3 and 116 DF,  p-value: < 2.2e-16
```

The three plots in Figure 1 give the three partial regression plots for this regression.

a) Can you identify which plot corresponds to which variable? Provide reasoning. (**4 points**).

b) Consider the data in the first partial regression plot. Suppose I fit a linear model to the y-variable based on the x-variable. What is the value of the RSS for this regression? (**2 points**)

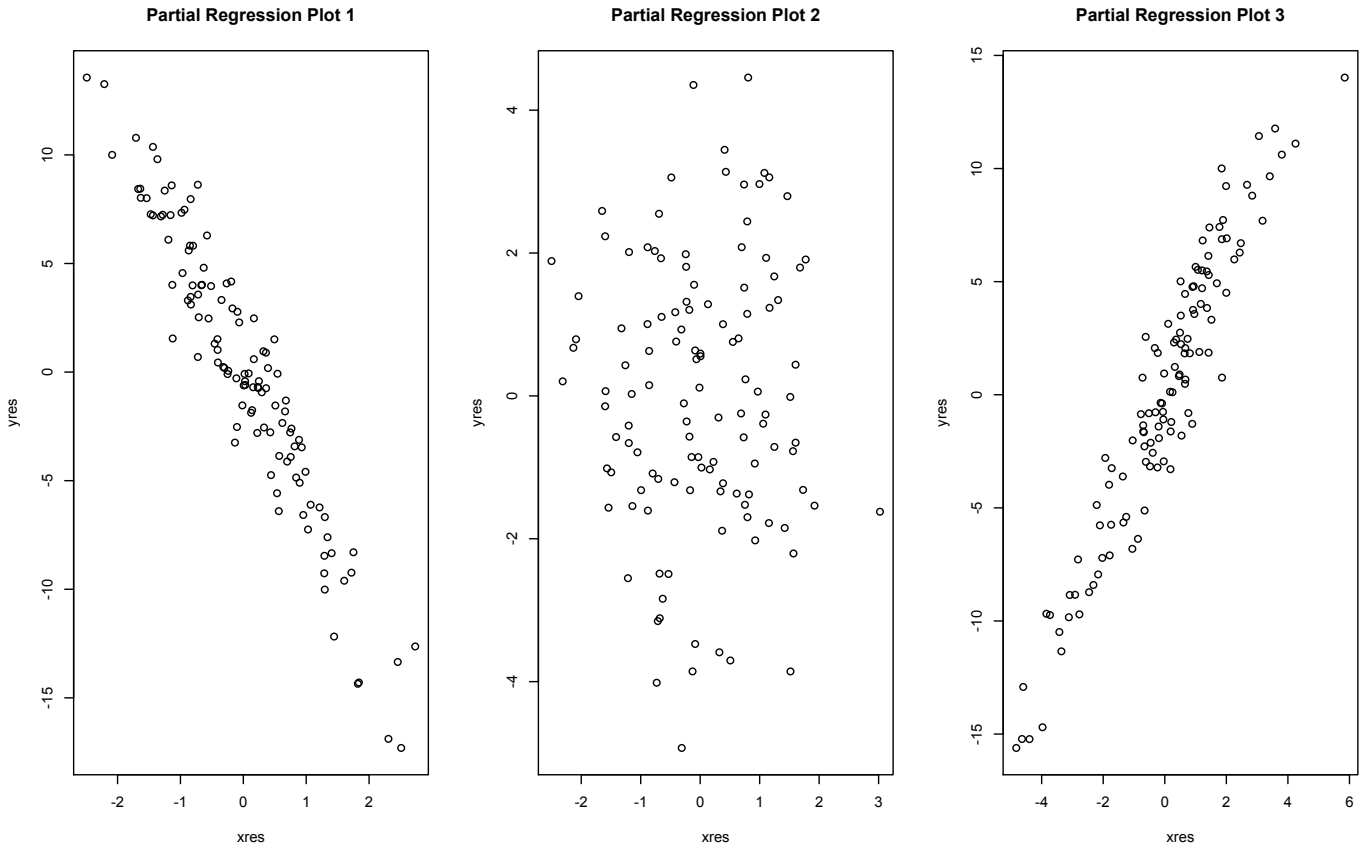

Figure 1: The three partial regression plots for the linear model shown above. One of these three plots corresponds to the first explanatory variable; one corresponds to the second explanatory variable and one to the third explanatory variable

.

5. Consider the bodyfat dataset and consider fitting a linear model for the response variable BODYFAT in terms of the explanatory variables AGE, WEIGHT, HEIGHT, ADIPOSITY, NECK, CHEST, ABDOMEN, HIP, THIGH, KNEE, ANKLE, BICEPS, FOREARM and WRIST.

   a) Using each of the following methods, perform variable selection to select a subset of the explanatory variables for modeling the response:

i. Backward elimination using the individual p-values.

ii. Forward Selection using p-values.

iii. Adjusted R^2.

iv. AIC

v. BIC

vi. Mallow's C_p.

b) Let $M_1, \ldots, M_6$ denote the six models selected by each of the six variable selection methods of the previous part. Select one of these models by cross-validation.

c) Let M be the model selected in the previous part. Fit this model to the data. Perform regression diagnostics. Comment on the validity of the assumptions of the linear model. Identify influential observations and outliers. Delete them if necessary and re-fit the model.