Statistics 151a, Spring 2015 (Linear Modelling - Theory and Applications) Homework Five

Due on April 27, 2015

14 April, 2015

1. Consider the linear model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$ for $i = 1, \ldots, n$ where $e \sim N(0, \sigma^2 I_n)$.
   (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. Write the difference $\hat{y}_i - \hat{y}_{i(i)}$ in terms of the ith standardized 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**)
2. I fit a linear model to the usual data $y_1, \ldots, y_n$ and x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let RSS denote the residual sum of squares and $\hat{e}$ denote the vector of residuals.

   I have been told that data on an explanatory variable has not been collected. More specifically, the right model here is apparently

   $$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \gamma z_i + e_i$$

   where $e_1, \ldots, e_n$ are uncorrelated mean zero errors with constant variance σ^2. Here $z_1, \ldots, z_n$ denote the values of a variable that has not been observed unfortunately.
   (a) Is $RSS/(n-p-1)$ an unbiased estimator of σ^2? If yes, explain with reason. If no, calculate the bias. (**4 points**)
   (b) Is the sum of the residuals $\hat{e}_i$ zero? Answer with reason. (**2 points**)
   (c) What is the expected value of $\hat{e}$? (**3 points**)
3. Consider the usual data on response $y_1, \ldots, y_n$ and explanatory variable data x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Suppose y_i can be modelled as a Poisson random variable with mean λ_i. Moreover suppose $y_1, \ldots, y_n$ can be assumed to be independent.
   (a) Write down the form of the canonical GLM. (**2 points**)
   (b) Write down the log-likelihood as a function of β. (**2 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
-4.4359 -1.3803  0.1258  1.4362  4.9530
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.94080    2.07498  -3.345  0.00111 **
x1           3.00762    0.07457  40.332  < 2e-16 ***
x2          -5.71932    0.17992 -31.789  < 2e-16 ***
x3           0.05069    0.17223   0.294  0.76904
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.827 on 116 degrees of freedom
Multiple R-squared: 0.9624,Adjusted R-squared: 0.9614
F-statistic: 989.2 on 3 and 116 DF,  p-value: < 2.2e-16
```

The three plots in Figure 1 give the three partial regression or added variable 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 added variable 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**)

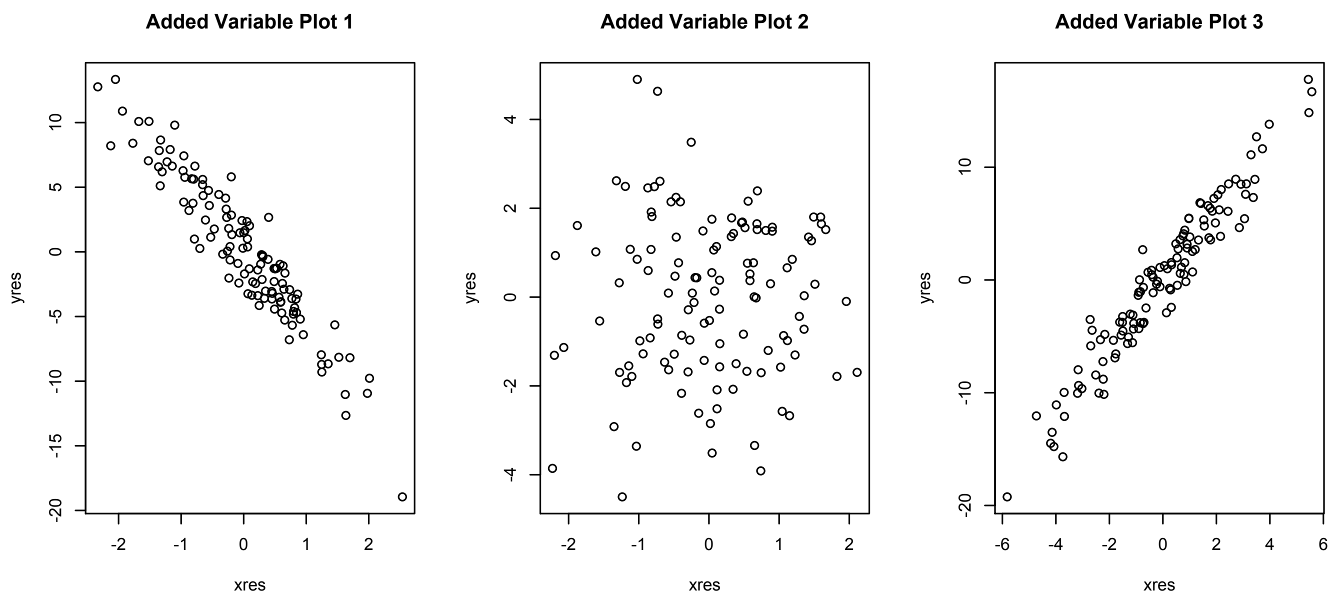

Figure 1: The three partial regression plots for the linear model shown above. One of these three plots corresponds to the first explanatory variable; one corresponds to the second explanatory variable and one to the third explanatory variable

.

5. Determine whether each of following statements is true or false. Provide reasons in each case.

(a) The magnitude of a predicted residual is never smaller than the magnitude of the corresponding residual. (**1 point**)

(b) Leverage of the ith subject depends on the value of y_i. (**1 point**)

(c) Model selection via AIC tends to produce smaller models than BIC. (**1 point**)

(d) The GCV is a computationally simpler model selection technique than PRESS. (**1 point**)

(e) The optimal model selected by Mallows' C_p criterion can have a C_p value that is more than $p+1$. (**1 point**)