STAT W4201 001, Homework 7

Brian Weinstein (bmw2148)

Mar 28, 2016

Code is attached here and also posted at https://github.com/BrianWeinstein/advanced-data-analysis. Where relevant, code snippets and output are are included in-line.

Problem 1: Ramsey 10.21

Let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

By deconstructing $(\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{X}^T\mathbf{Y}$ we find the set of normal equations:

$$(\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{X}^T\mathbf{Y}$$

$$\Rightarrow \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \right) \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} n & \sum X_{i1} & \sum X_{i2} & \cdots & \sum X_{ip} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} & \cdots & \sum X_{i1}X_{ip} \\ \sum X_{i2} & \sum X_{i2}X_{i1} & \sum X_{i2}^2 & \cdots & \sum X_{i2}X_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum X_{ip} & \sum X_{ip}X_{i1} & \sum X_{ip}X_{i2} & \cdots & \sum X_{ip}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \\ \vdots \\ \sum X_{ip}Y_i \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \beta_0 n + \beta_1 \sum X_{i1} + \beta_2 \sum X_{i2} + \cdots + \beta_p \sum X_{ip} \\ \beta_0 \sum X_{i1} + \beta_1 \sum X_{i1}^2 + \beta_2 \sum X_{i1}X_{i2} + \cdots + \beta_p \sum X_{i1}X_{ip} \\ \beta_0 \sum X_{i2} + \beta_1 \sum X_{i2}X_{i1} + \beta_2 \sum X_{i2}^2 + \cdots + \beta_p \sum X_{i2}X_{ip} \\ \vdots \\ \beta_0 \sum X_{ip} + \beta_1 \sum X_{ip}X_{i1} + \beta_2 \sum X_{ip}X_{i2} + \cdots + \beta_p \sum X_{ip}^2 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \\ \vdots \\ \sum X_{ip}Y_i \end{bmatrix},$$

where each $\sum$ indicates summation over all cases ($i = 1, 2, \ldots, n$).

As long as $\mathbf{X}^T\mathbf{X}$ is invertible, the least squares solution is $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. By the invertible matrix theorem, $\mathbf{X}^T\mathbf{X}$ is invertible whenever the columns of $\mathbf{X}$ are linearly independent.

Problem 2: Ramsey 10.22

Problem 2: 10.22

Problem 3: Ramsey 10.26

Does the effect of UVB exposure on the distribution of percentage inhibition differ at the surface and in the deep? How much difference is there? Analyze the data, and write a summary of statistical findings and a section of details documenting those findings.

A sample of the dataset is shown below, and a coded scatterplot of the `Inhibit` vs `UVB` is shown in Figure 1.

```
    Inhibit  UVB Surface
1       0.0 0.00       0
2       1.0 0.00       0
3       6.0 0.01       0
4       7.0 0.01       1
...
14     21.0 0.04       1
15     25.0 0.02       0
16     39.0 0.03       0
17     59.0 0.03       0
```

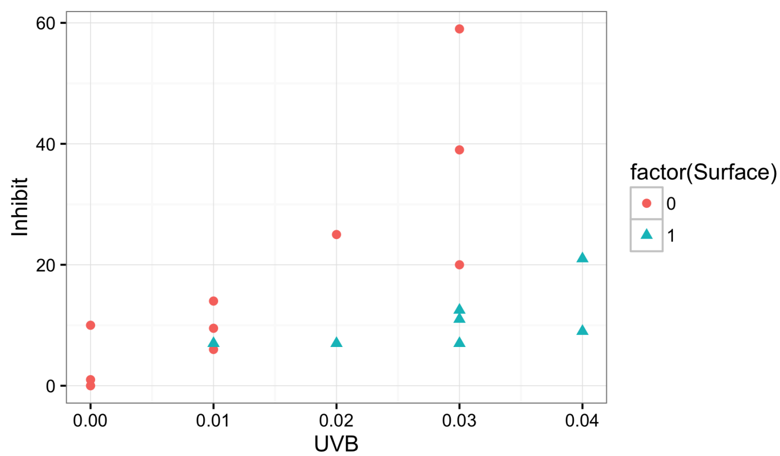

Figure 1: Coded scatterplot of `Inhibit` vs `UVB`. `Surface` = 1 indicates a measurement close to the surface, and `Surface` = 0 in deep water.

First, we fit a linear model on the full dataset, including an interaction term between `UVB` and `Surface`.

```
> lmFull <- lm(Inhibit ~ Surface*UVB, data=ozoneData)
> summary(lmFull)$coefficients
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    1.180556   4.292122  0.2750517 0.7875994449
Surface        1.277778  11.065902  0.1154698 0.9098372976
UVB         1226.388889 232.772995  5.2686047 0.0001518894
Surface:UVB -939.930556 409.839150 -2.2934133 0.0391340538
```

In this model, the p-value on the intercept term indicates that the data provides no evidence to suggest that the intercept is non-zero. Based on the dataset, we also know the intercept should be nearly 0 for both the surface-level and deep water measurements. We fit a second model, forcing the intercept term to 0.

```
> lmZeroInt <- lm(Inhibit ~ 0 + Surface*UVB, data=ozoneData)
> summary(lmZeroInt)$coefficients
               Estimate Std. Error    t value      Pr(>|t|)
Surface        2.458333   9.857139  0.2493962 0.80667620946
UVB         1275.000000 146.400333  8.7089966 0.00000050281
Surface:UVB -988.541667 357.358785 -2.7662442 0.01515365491
```

Although the model with 0-intercept has a slightly higher RMSE, the difference is negligible (8.537, vs 8.833 in the full model), and the significance of the coefficients is unchanged.

Based on the full model, the data provides moderate evidence that the effect of UVB exposure on the distribution of percentage inhibition differs at the surface and in the deep (two sided p-value of 0.03913, from a test that the `Surface:UVB` coefficient estimate is 0; estimated value -939.9, 95% CI from -1825 to -54.53). In deep water, for each 0.01 unit increase in UVB exposure, the percent inhibition increases by 12.26 percentage points. Near the surface, however, for each 0.01 unit increase in UVB exposure, the percent inhibition increases by only 2.861 percentage points.

The regression equations for this "separate lines" model are shown below

- Near the surface (`Surface` $= 1$): `Inhibit` $= 3.084 + 286.1 \times$ `UVB`,
- and in deep water (`Surface` $= 0$): `Inhibit` $= 1.806 + 1226 \times$ `UVB`.

Problem 4: Ramsey 11.8

(a) *Why does a case with large leverage have the potential to be influential? Why is it not necessarily influential?*

An observation with large leverage h_i has a residual with low variability

$$\mathrm{SD}(\mathrm{Residual}_i) = \sigma\sqrt{1-h_i}.$$

A high-leverage observation's explanatory variable values are abnormally high or low, as compared to the other observations in the dataset. As such, the high-leverage observation determines the value of the regression in the surrounding region.

This, however, does not mean that all high-leverage observations are influential. If the observation happens to fall close to the regression surface that is fit only to the other observations (i.e., excluding the high-leverage observation), then the high-leverage observation has little impact on the overall regression surface.

(b) *Draw a hypothetical scatterplot of Y versus a single X, where one observation has a high leverage but is not influential.*

A scatterplot with an observation with high-leverage, but low-influence is shown in Figure 2.

(c) *Draw a hypothetical scatterplot of Y versus a single X, where one observation has a high leverage and is influential.*

A scatterplot with an observation with high-leverage and high-influence is shown in Figure 3.

Problem 5: Ramsey 11.16

Problem 6: Ramsey 11.20

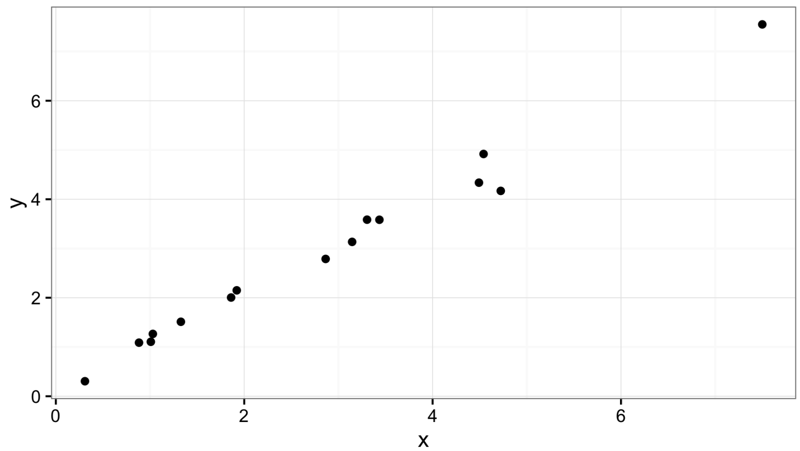

Figure 2: A scatterplot with an observation with high-leverage, but low-influence.

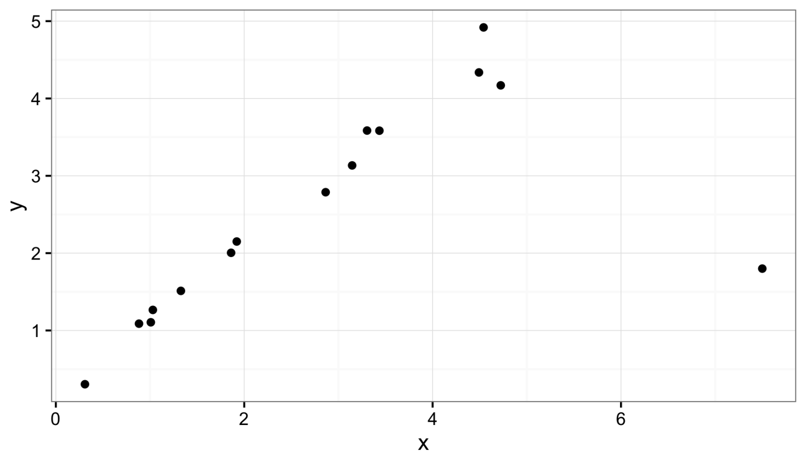

Figure 3: A scatterplot with an observation with high-leverage and high-influence.

Todo list