Notes 13

Maryclare Griffin

5/8/2023

These notes are based on Chapters 3, 7, and 8 of KNNL.

We will now differentiate between the assumed the **normal error linear regression model** for a dependent variable or response Y and independent variables, predictors, or covariates X_1, … X_{p-1} and the **true linear regression model**. Recall, the **normal error linear regression model** is defined as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

- The elements of $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ are parameters
- The elements of the $n \times p$ matrix $\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \dots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{n,p-1} \end{pmatrix}$ are known constants
- $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ is a random error term elements that are ϵ_i that are independent and normally distributed with mean $E\left\{\epsilon_i\right\} = 0$ and variance $\sigma^2\left\{\epsilon_i\right\} = \sigma^2$.

In real life, the **true linear regression model** may be different. We'll consider two possibilities:

- The true linear regression model is different from the assumed normal error linear regression model for a small number of observations, which we call **outliers**.
- The true linear regression model is different from the assumed normal error linear regression model for all observations.

An **outlier** is extreme observation which is not well explained by the assumed linear regression model. In small samples, when n is small relative to p, outliers can lead to poor estimates of the regression coefficients $\boldsymbol{b}$. They can be identified from the residuals; they correspond to points that have residuals that are especially large in magnitude. When it can be verified that an outlier corresponds to an erroneous measurement, e.g. an error in recording, a miscalculation, or a malfunctioning of equipment, it can be excluded and ignored.

> ***Example 1:*** Consider data from a study of the relation between plutonium activity (X) and the observed alpha count (Y).

```
load("~/Dropbox/Teaching/STAT525/Spring2023/bookdata/plutonium.RData")

X <- data$X
```

```
Y <- data$Y
n <- length(Y)

linmod <- lm(Y~X)
b0 <- linmod$coef[1]
b1 <- linmod$coef[2]

outlier <- which(X == 0 & Y > 0.1)
outlier
```

```
## [1] 24
```

```
linmod.new <- lm(Y~X, subset = which(!(data$X == 0 & data$Y > 0.1)))
b0.new <- linmod.new$coef[1]
b1.new <- linmod.new$coef[2]

plot(X, Y, pch = 16,
     xlab = "Plutonium Activity", ylab = "Alpha Count Rate")
abline(a = b0, b = b1, col = "blue")
abline(a = b0.new, b = b1.new, col = "red")
legend("bottomright", col = c("blue", "red"),
       legend = c("With Outlier", "Without Outlier"),
       lty = 1, title = "Estimated Regression Function",
       bty = "n")
```

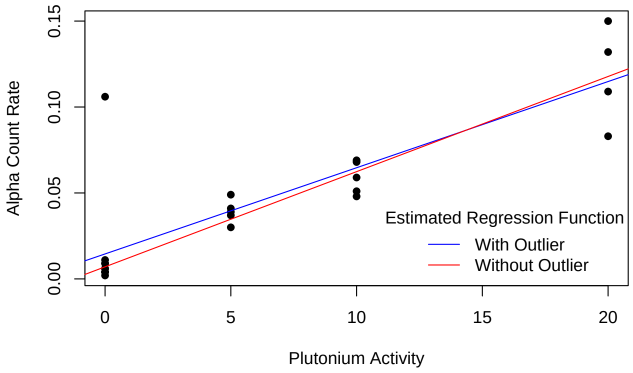

Figure 1: Example 1 (a)

We can clearly see an extreme value, specifically the value of $Y_i = 0.10$ when $X_i = 0$.

```
e <- linmod$residuals
plot(X[-outlier], e[-outlier], xlab = "Plutonium Activity", ylab = "Residuals",
     pch = 16, ylim = range(e))
points(X[outlier], e[outlier], pch = 18, col = "gray")
legend("topright", pch = 18, legend = "Outlier", col = "gray")
```

Especially when there are multiple covariates ($p > 1$), outliers can be easier to identify from a plot of the residuals. In this case, an examination of laboratory records revealed that the outlier corresponds to a situation where experimental conditions were not properly maintained, and can therefore be excluded.

When the true linear regression model is different from the assumed linear regression model for all observations,

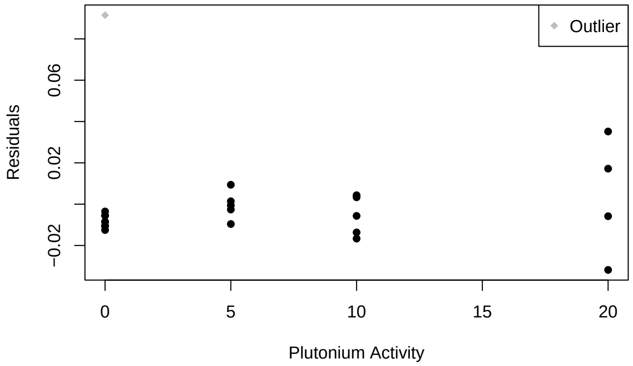

Figure 2: Example 1 (b)

it can be helpful to define the **true error linear regression model** is defined as:

$$\boldsymbol{Y} = f(\boldsymbol{X})\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where:

- The elements of $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ and $\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_q \end{pmatrix}$ are parameters
- The elements of the $n \times p$ matrix $\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \dots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{n,p-1} \end{pmatrix}$ and $n \times q$ matrix $\boldsymbol{Z} = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1q} \\ Z_{21} & Z_{22} & \dots & Z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \dots & Z_{nq} \end{pmatrix}$ are known constants
- $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ is a random error term.