Notes 4

Maryclare Griffin

2/14/2023

These notes are based on Chapters 1 and 6 of KNNL.

Recall from the previous notes, the linear regression model for a dependent variable or response Y and independent variables, predictors, or covariates $X_1, \ldots X_{p-1}$ is defined as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

where:

- $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters
- $X_{i1}, \ldots, X_{i,p-1}$ are known constants
- ϵ_i is a random error term with mean $E\{\epsilon_i\} = 0$ and variance $\sigma^2\{\epsilon_i\} = \sigma^2$; ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $\sigma\{\epsilon_i, \epsilon_j\} = 0$ for all i, j; $i \neq j$)
- $i = 1, \ldots, n$

Remember, we don't observe $\beta_0, \beta_1, \ldots, \beta_{p-1}$ in the real world. Instead, we **estimate** them by finding the values $b_0, b_1, \ldots, b_{p-1}$ that minimize the sum of squared deviations of the response values Y_i from the regression function $\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}$ with respect to $\beta_0, \beta_1, \ldots, \beta_{p-1}$. We call the sum of squared deviation the **sum of squares**, and denote it by Q:

$$Q = \sum_{i=1}^{n} \left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2.$$

We minimize the Q by taking derivatives of Q with respect to $\beta_0, \beta_1, \ldots, \beta_{p-1}$ and finding the values of $b_0, b_1, \ldots, b_{p-1}$ that set the derivatives equal to zero when substituted in for $\beta_0, \beta_1, \ldots, \beta_{p-1}$ in the derivatives.

Before we start taking derivatives, let's expand out Q to be a nicer function of $\beta_0, \beta_1, \ldots, \beta_{p-1}$.

$$\begin{aligned} Q &= \sum_{i=1}^{n} Y_i^2 - 2Y_i \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) + \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right)^2 \\ &= \sum_{i=1}^{n} Y_i^2 - 2Y_i \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) + \beta_0^2 + \left(\sum_{k=1}^{p-1} \beta_k^2 X_{ik}^2 \right) + 2 \left(\sum_{k=1}^{p-1} \beta_0 \beta_k X_{ik} \right) + \left(\sum_{k=1}^{p-1} \sum_{l=1, l \neq k}^{p-1} \beta_k \beta_l X_{ik} X_{il} \right) \end{aligned}$$

- $\frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^{n} -2Y_i + 2\beta_0 + 2\left(\sum_{k=1}^{p-1} \beta_k X_{ik} \right)$
- For $k > 0$, $\frac{\partial Q}{\partial \beta_k} = \sum_{i=1}^{n} -2Y_i X_{ik} + 2\beta_k X_{ik}^2 + 2\beta_0 X_{ik} + 2\sum_{l=1, l \neq k}^{p-1} \beta_l X_{il} X_{ik}$

We call this set of p equations, which set each derivative of Q equal to 0, the **normal equations**. Finding the values $b_0, b_1, \ldots, b_{p-1}$ of $\beta_0, \beta_1, \ldots, \beta_{p-1}$ that set these derivatives equal to zero doesn't look easy. Each equation depends on **all** of the regression coefficients, not just one. This is why we're going to need linear

algebra if we want to be able to write out a closed form solutions for $b_0, b_1, \ldots, b_{p-1}$ in terms of the data when we have more than one predictor and $p > 2$.

Fortunately, `R` and most statistical software packages have a function that can solve for $b_0, b_1, \ldots, b_{p-1}$ for us! In `R`, the function that finds the values $b_0, b_1, \ldots, b_{p-1}$ that minimizes the sum of squared errors Q is called `lm`. Examples of using `lm` are provided below.

Example 1: Consider data from a company that manufactures refrigeration equipment, called the Toluca company. They produce refrigerator parts in lots of different sizes, and the amount of time it takes to produce a lot of refrigerator parts depends on the number of parts in the lot and several other variable factors. Let X be the number of refrigerator plots in a lot, and let Y refer to the amount of time it takes to produce a size of lot X.

```
load("~/Dropbox/Teaching/STAT525/Spring2023/bookdata/toluca.RData")
X <- data$X # Extract the X values
Y <- data$Y # Extract the Y values
linmod <- lm(Y~X) # Obtain the linear regression coefficients
b0 <- linmod$coef[1]
b1 <- linmod$coef[2]
plot(X, Y, xlab = "Lot Size",
     ylab = "Hours",pch = 16) # Plot values of Y against values of X with axis labels
abline(a = b0, b = b1, col = "blue") # Add estimated regression function to plot
```

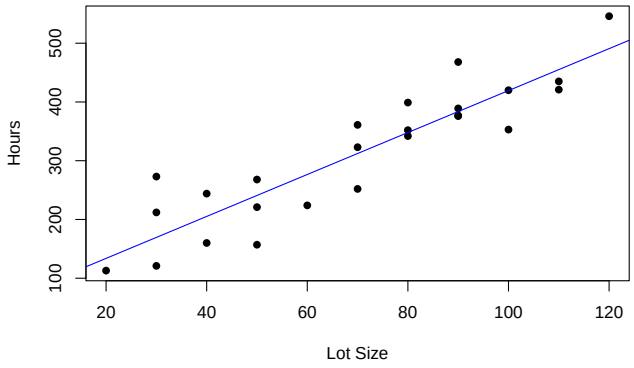

Figure 1: Example 1

We obtain an intercept estimate of $b_0 = 62.366$ and $b_1 = 3.57$.

We can intepret b_0 as the estimated mean number of hours it takes to produce a lot made up of 0 parts. This doesn't make much sense for this dataset or problem, because we did not observe lots of 0 parts.

We can intepret b_1 as indicating that a one unit increase in lot size in is associated with a 3.57 hour increase in time needed to produce the lot.

The data and estimated regression function are shown in Figure 1. Black, filled-in dots represent the observed lot sizes and corresponding hours needed, and the solid blue line represents the estimated regression function.

Example 2: Consider data on age (X) and plasma level of a polyamine (Y) for 25 healthy children.

```
load("~/Dropbox/Teaching/STAT525/Spring2023/bookdata/plasma.RData")
X <- data$X # Extract the X values
Y <- data$Y # Extract the Y values
linmod <- lm(Y~X) # Obtain the linear regression coefficients
b0 <- linmod$coef[1]
b1 <- linmod$coef[2]
plot(X, Y, xlab = "Age (years)",
     ylab = "Plasma Level", pch = 16)
abline(a = b0, b = b1, col = "blue") # Add estimated regression function to plot
```

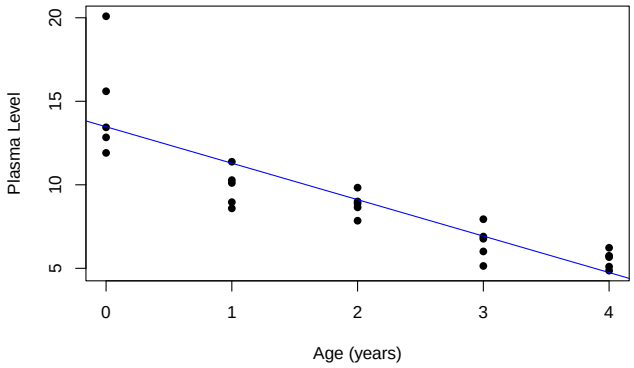

Figure 2: Example 2

We obtain an intercept estimate of $b_0 = 13.475$ and $b_1 = -2.182$.

We can intepret b_0 as the estimated mean plasma level for a 0 year old child. This does make sense for this dataset and problem, because we do observe the plasma levels of children at age 0.

We can intepret b_1 as indicating that a one unit increase in age is associated with a -2.182 decrease in plasma level.

The data and estimated regression function are shown in Figure 2. Black, filled-in dots represent the observed ages and plasma level, and the solid blue line represents the estimated regression function.

Example 3: Consider data from portrait studios in 21 cities run by Dwaine Studios, Inc. The studios specialize in portraits of children. Let X_1 be the number of persons aged 16 or younger in a city, let X_2 refer to per capita disposable income in a city, and let Y be the sales of portraits of children in that city from one of the 21 studies.

```
load("~/Dropbox/Teaching/STAT525/Spring2023/bookdata/dwaine.RData")
X1 <- data$X1 # Extract the first predictor
X2 <- data$X2 # Extract the second predictor
Y <- data$Y # Extract the response
linmod <- lm(Y~X1+X2) # Obtain the linear regression coefficients
b0 <- linmod$coef[1]
b1 <- linmod$coef[2]
b2 <- linmod$coef[3]
```

We obtain a estimates $b_0 = -68.857$, $b_1 = 1.455$, and $b_2 = 9.366$.

We can intepret b_0 as the estimated mean photo sales in a city with 0 persons aged 16 or younger and 0 per capita disposable income. This does make sense for this dataset and problem, because we do observe the plasma levels of children at age 0.

We can intepret b_1 as indicating that a additional person under 16 is associated with a 1.455 increase in sales of portraits of children, holding per capita disposable income constant.

We can intepret b_2 as indicating that one additional unit of per capita disposable income is associated with a 9.366 increase in sales of portraits of children, holding the number of persons under 16 constant.

We refer to $b_0, b_1, \ldots, b_{p-1}$ as the **estimated regression coefficients**. Sometimes we will state that they are **point estimates** of $\beta_0, \beta_1, \ldots, \beta_{p-1}$. The term **point estimate** is a general way of referring to an estimate of an unknown quantity.

Having obtained the estimated regression coefficients $b_0, b_1, \ldots, b_{p-1}$, we can write down the corresponding **estimated regression function** or **estimated mean response** is

$$\hat{Y} = b_0 + \sum_{k=1}^{p-1} b_k X_k,$$

where $\hat{Y}$ (pronounced Y hat) is the value of the estimated regression function at level $X_1, \ldots, X_{p-1}$ of the predictors.

Now let's talk about solving for b_0 and b_1 in closed form. We will rarely need to do this in practice, but learning the closed form solutions is crucial for understanding properties of b_0 and b_1 and building intuition. When we have just one predictor, we can solve for b_0 and b_1 in closed form. We'll focus on that for now. When we have one predictor and $p = 2$, the sum of squares Q simplifies to:

$$Q = \sum_{i=1}^{n} Y_i^2 - 2Y_i \left(\beta_0 + \beta_1 X_i\right) + \beta_0^2 + \beta_1^2 X_i^2 + 2\beta_0\beta_1 X_i.$$

There are just two derivatives to consider, which simplify to:

- $\frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^{n} -2Y_i + 2\beta_0 + 2\beta_1 X_i$
- $\frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^{n} -2Y_i X_i + 2\beta_1 X_i^2 + 2\beta_0 X_i$

The corresponding **normal equations** are:

- $\sum_{i=1}^{n} -2Y_i + 2b_0 + 2b_1 X_i = 0$
- $\sum_{i=1}^{n} -2Y_i X_i + 2b_1 X_i^2 + 2b_0 X_i = 0$

Let's rearrange them, starting with the first one:

$$\begin{aligned}
\sum_{i=1}^{n} -2Y_i + 2b_0 + 2b_1 X_i = 0 \implies nb_0 &= \sum_{i=1}^{n} -Y_i - b_1 X_{i1} \\
b_0 &= \frac{1}{n}\sum_{i=1}^{n} Y_i - b_1 X_i \\
b_0 &= \bar{Y} - b_1 \bar{X}.
\end{aligned}$$

This tells us that if we know b_1, we can determine b_0. What about the second equation?

$$\begin{aligned}
\sum_{i=1}^{n} -2Y_i b_1 X_i + 2b_1 X_i^2 + 2b_0 b_1 X_i = 0 \implies b_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} Y_i X_i - b_0 X_i \\
b_1 &= \frac{\sum_{i=1}^{n} X_i \left(Y_i - b_0\right)}{\sum_{i=1}^{n} X_i^2}.
\end{aligned}$$

This tells us that if we know b_0, we can determine b_1. What if we don't know either? Let's plug our expression for b_0 into the equation $b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i - b_0 X_i$. We get:

$$\begin{aligned}
b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i - \left(\bar{Y} - b_1 \bar{X}\right) X_i \implies b_1 \left(\sum_{i=1}^{n} X_i \left(X_i - \bar{X}\right)\right) &= \sum_{i=1}^{n} X_i \left(Y_i - \bar{Y}\right) \\
b_1 &= \frac{\sum_{i=1}^{n} X_i \left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} X_i \left(X_i - \bar{X}\right)}
\end{aligned}$$

Aha! We can compute the least squares estimate b_1 of β_1 from the data! Then we can plug our least squares estimate b_1 into the equation $b_0 = \bar{Y} - b_1\bar{X}$ to obtain the least squares estimate b_0 of β_0.

Example 4: Let's revisit the data from Problem 3 of Homework 1, where we have he regular season three point shooting percentages of Jaylen Brown, Al Horford, Marcus Smart, Jayson Tatum, Payton Pritchard, Grant Williams, Sam Hauser, and Derrick White during the 2021-2022 season (X) and 2022-2023 season (Y). What are the least squares estimates of b_0 and b_1?

```
x <- c(35.8, 33.6, 33.1, 35.3, 41.2, 41.1, 43.2, 30.6)
y <- c(33.4, 41.7, 33.3, 35.5, 33.0, 41.4, 39.7, 37.7)

x.bar <- mean(x)
y.bar <- mean(y)

b1 <- sum(x*(y - y.bar))/sum(x*(x - x.bar))
b0 <- y.bar - b1*x.bar
```

We obtain an intercept estimate of $b_0 = 32.194$ and $b_1 = 0.13$.

We can intepret b_0 as the estimated mean three point shooting percentage in 2022-2023 of a Celtics player who had a shooting percentage of 0 in 2021-2022. This doesn't make much sense for this dataset or problem, because we did not observe any 2021-2022 three point shooting percentages close to 0 and it is implausible that a player would have a three point shooting percentage of 0 in any season.

We can intepret b_1 as indicating that a one percent increase in 2021-2022 three point shooting percentage is associated with a 0.13 percent increase in 2022-2023 three point shooting percentages.

The data and estimated regression function are shown in Figure 4. Blue, filled-in dots represent the observed three point shooting percentages, a red filled-in dot represents the average data point $(\bar{X}, \bar{Y})$, the dashed line represents a 45° line, and the solid line represents the estimated regression function.

```
plot(x, y, xlab = "2021-2022 3PT",
     ylab = "2022-2023 3PT",
     ylim = c(20, 50),
     xlim = c(20, 50), pch = 16, col = "blue")
abline(a = 0, b = 1, lty = 3)
points(x.bar, y.bar, pch = 16, col = "red")
curve(b0 + b1*x, from = 10, to = 50, add = TRUE)
```

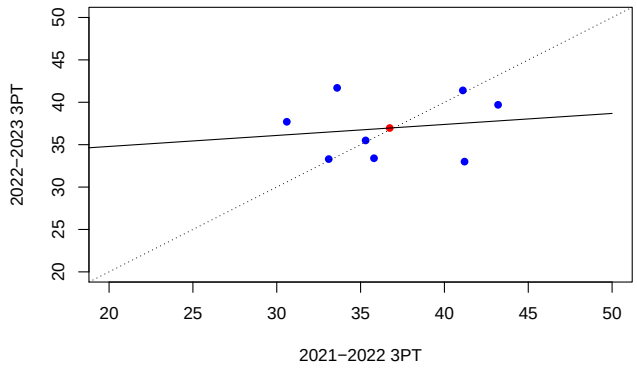

Figure 3: Example 4

Note: In practice, it is more common to see the equation

$$b_1 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}.$$

This gives the same result, because $\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) = \sum_{i=1}^{n} X_i \left(Y_i - \bar{Y}\right)$ and $\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} X_i \left(X_i - \bar{X}\right)$. Showing this is annoying, but a rite of passage. We'll do so below.

$$\begin{aligned}
\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) &= \sum_{i=1}^{n} X_i Y_i - \bar{X} Y_i - \bar{Y} X_i + \bar{X}\bar{Y} \\
&= n\bar{X}\bar{Y} - \bar{Y}\sum_{i=1}^{n} X_i - \bar{X}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} X_i Y_i \\
&= n\bar{X}\bar{Y} - 2n\bar{X}\bar{Y} + \sum_{i=1}^{n} X_i Y_i \\
&= -n\bar{X}\bar{Y} + \sum_{i=1}^{n} X_i Y_i \\
&= -\bar{Y}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} X_i Y_i \\
&= \sum_{i=1}^{n} X_i \left(Y_i - \bar{Y}\right) \\
\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 &= \sum_{i=1}^{n} X_i^2 - 2\bar{X} X_i + \bar{X}^2 \\
&= n\bar{X}^2 - 2\bar{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} X_i^2 \\
&= n\bar{X}^2 - 2n\bar{X}^2 + \sum_{i=1}^{n} X_i^2 \\
&= -n\bar{X}^2 + \sum_{i=1}^{n} X_i^2 \\
&= -\bar{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} X_i^2 \\
&= \sum_{i=1}^{n} X_i \left(X_i - \bar{X}\right)
\end{aligned}$$

Ta-da!