Notes 10

Maryclare Griffin

4/4/2023

These notes are based on Chapters 2 and 6 of KNNL.

We will continue to assume the **normal error linear regression model** for a dependent variable or response Y and independent variables, predictors, or covariates $X_1, \dots X_{p-1}$ 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 $\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \dots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{11} & X_{12} & \dots & X_{1,p-1} \end{pmatrix}$ 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 $\sigma^2\left\{\epsilon_i\right\} = \sigma^2$.

Under the **normal error linear regression model**, we have shown that the studentized statistic

$$\frac{b_k - \beta_k}{s\left\{b_k\right\}} \sim t\left(n-p\right),$$

where $t\left(n-p\right)$ refers to a t distribution with $n-p$ degrees of freedom.

This allows us to formally test a null hypothesis of the form H_0: $\beta_k = c$ versus an alternative hypothesis of the form H_a: $\beta_k \neq c$, for some pre-specified value c. In the previous set of notes, we did this in an informal way for $c = 0$ by visually comparing $\frac{b_k - c}{s\{b_k\}}$ to the density of a t distribution with $n-p$ degrees of freedom, and concluding that the null H_0 was unlikely to be true.

To formally test this null hypothesis, we will find an interval that contains $\frac{b_k - c}{s\{b_k\}}$ with probability $1-\alpha$ when the null H_0 is true, and conclude the alternative H_a if $\frac{b_k - c}{s\{b_k\}}$ is outside of that interval. We will call α the **level** of the test or the **Type I error**. The **level** of the test, α, describes the probability of concluding the alternative H_a when the null H_0 is true. Remember, if the null H_0 is true, then $\frac{b_k - c}{s\{b_k\}}$ has a t distribution with $n-p$ degrees of freedom. Let $t\left(\alpha/2; n-p\right)$ refer to the $\alpha/2$ quantile of a t distribution with $n-p$ degrees of freedom and let $t\left(1-\alpha/2; n-p\right)$ refer to the $1-\alpha/2$ quantile of a t distribution with $n-p$ degrees of freedom. The interval $\left[t\left(\alpha/2; \nu\right), t\left(1-\alpha/2; \nu\right)\right]$ will contain $\frac{b_k - c}{s\{b_k\}}$ with probability $1-\alpha$ when the null H_0 is true.

Note: Let $t(\nu)$ be a random variable distributed according to a t distribution with ν degrees of freedom. The α quantile of a t distribution with ν degrees of freedom is denoted by $t(\alpha;\nu)$, and defined as satisfying:

$$P(t(\nu) \leq t(\alpha;\nu)) = \alpha.$$

Under the normal errors linear regression model, the decision rule based on a the test statistic $\frac{b_k - c}{s\{b_k\}}$ for a level $1-\alpha$ test of the null hypothesis H_0: $\beta_k = c$ versus the alternative hypothesis H_a: $\beta_k \neq c$ is:

- If $t(\alpha/2; n-p) \leq \frac{b_k - c}{s\{b_k\}} \leq t(1-\alpha/2; n-p)$, conclude the null H_0
- If $\frac{b_k - c}{s\{b_k\}} < t(\alpha/2; n-p)$ or $\frac{b_k - c}{s\{b_k\}} > t(1-\alpha/2; n-p)$, conclude the alternative H_a

Note: We can think of a test of the null hypothesis H_0: $\beta_k = 0$ versus the alternative hypothesis H_a: $\beta_k \neq 0$ as a test of the null hypothesis that there is no linear statistical association between the response $\boldsymbol{Y}$ and the predictor $\boldsymbol{X}_k$ given the remaining predictors are included in the model versus the alternative hypothesis that there is a linear association between the response $\boldsymbol{Y}$ and the predictor $\boldsymbol{X}_k$ given the remaining predictors are included in the model.

We can make this simpler using a nice property of the t distribution.

Note: The t distribution with ν degrees of freedom is symmetrical about 0. As a result, $-t(\alpha/2; n-p) = t(1-\alpha/2; n-p)$.

Under the normal errors linear regression model, we can alternatively say that the decision rule based on a the test statistic $\frac{b_k - c}{s\{b_k\}}$ for a level $1-\alpha$ test of the null hypothesis H_0: $\beta_k = c$ versus the alternative hypothesis H_a: $\beta_k \neq c$ is:

- If $\left|\frac{b_k - c}{s\{b_k\}}\right| \leq t(1-\alpha/2; n-p)$, conclude the null H_0
- If $\left|\frac{b_k - c}{s\{b_k\}}\right| > t(1-\alpha/2; n-p)$, conclude the alternative H_a

Example 1: Again, 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. Suppose a cost analyst in the Toluca Company is interested in testing whether or not there is a linear association between work hours and lot size, i.e. the null hypothesis H_0: $\beta_1 = 0$ at level $\alpha = 0.05$.

```r
load("~/Dropbox/Teaching/STAT525/Spring2023/bookdata/toluca.RData")
n <- nrow(data) # Extract number of observations
Y <- data$Y # Extract response
X <- data$X # Extract predictor
linmod <- lm(Y~X) # Fit linear model
b1 <- linmod$coef[2]
s.b1 <- summary(linmod)$coef[2, 2]
alpha <- 0.05
tquantile <- qt(1 - alpha/2, n - 2)
```

We obtain $b_1 = 3.57$ and $s\{b_1\} = 0.347$. Accordingly, the test statistic is $b_1/s\{b_1\} = 10.29$. We compare this to the 0.975 quantile of a t distribution with 23 degrees of freedom, $t(0.975; 23) = 2.069$. Because the test statistic $b_1/s\{b_1\}$ exceeds $t(0.975; 23)$, we conclude H_a: $\beta_1 \neq 0$, i.e. we conclude that there is evidence of a linear association between work hours and lot size at level $\alpha = 0.05$.

When we are performing a test, it can also be helpful to compute the corresponding ***p*-value**, which is the probability of observing a test statistic that is more extreme than the observed value if the null H_0 is true.

When we are performing a level $1-\alpha$ test of the null hypothesis H_0: $\beta_k = c$ versus the alternative hypothesis H_a: $\beta_k \neq c$, the p-value is

$$
\begin{aligned}
P\left(t\left(n-p\right) < -\left|\frac{b_k - c}{s\left\{b_k\right\}}\right| \text{ or } t\left(n-p\right) > \left|\frac{b_k - c}{s\left\{b_k\right\}}\right|\right) &= P\left(t\left(n-p\right) < -\left|\frac{b_k - c}{s\left\{b_k\right\}}\right|\right) + P\left(t\left(n-p\right) > \left|\frac{b_k - c}{s\left\{b_k\right\}}\right|\right) \\
&= 2P\left(t\left(n-p\right) < -\left|\frac{b_k - c}{s\left\{b_k\right\}}\right|\right).
\end{aligned}
$$

The last line is a simpification that follows from the symmetry of a t distribution with ν degrees of freedom about 0.

> ***Example 2:*** Consider the same data. What is the p-value of the test of whether or not there is a linear association between work hours and lot size, i.e. the p-value of the test of the the null hypothesis H_0: $\beta_1 = 0$?

```
pvalue <- 2*pt(-abs(b1/s.b1), n - 2)
```

> We obtain a p-value of $4.4488276 \times 10^{-10}$.
>
> ***Note:*** We **never** say that a p-value is 0. When a p-value is extremely small, we either provide the value as we do above, write $p < 10^{-3}$, or write $p = 0+$.

We can also conduct **one-sided tests** of the form H_0: $\beta_k = 0$ versus the alternative H_a: $\beta_k > 0$ or H_0: $\beta_k = 0$ versus the alternative H_a: $\beta_k < 0$. These are rarely used in practice, so we will not discuss them here.

The last thing we will discuss is obtaining a $100 \times (1-\alpha)\%$ confidence interval for β_k. Because we know that $\frac{b_k - \beta_k}{s\{b_k\}}$ follows a t distribution with $n-p$ degrees of freedom, the following holds for all probabilities α:

$$
P\left(t\left(\alpha/2; n-p\right) \leq \frac{b_k - \beta_k}{s\left\{b_k\right\}} \leq t\left(\alpha/2; n-p\right)\right) = 1-\alpha.
$$

Let's rearrange the terms, to see if we can get an inequality for β_k.

$$
\begin{aligned}
P\left(t\left(\alpha/2; n-p\right) \leq \frac{b_k - \beta_k}{s\left\{b_k\right\}} \leq t\left(1-\alpha/2; n-p\right)\right) &= P\left(t\left(\alpha/2; n-p\right) s\left\{b_k\right\} \leq b_k - \beta_k \leq t\left(1-\alpha/2; n-p\right) s\left\{b_k\right\}\right) \\
&= P\left(t\left(\alpha/2; n-p\right) s\left\{b_k\right\} - b_k \leq -\beta_k \leq t\left(1-\alpha/2; n-p\right) s\left\{b_k\right\} - b_k\right) \\
&= P\left(b_k - t\left(1-\alpha/2; n-p\right) s\left\{b_k\right\} \leq \beta_k \leq b_k - t\left(\alpha/2; n-p\right) s\left\{b_k\right\}\right) \\
&= P\left(b_k + t\left(\alpha/2; n-p\right) s\left\{b_k\right\} \leq \beta_k \leq b_k - t\left(\alpha/2; n-p\right) s\left\{b_k\right\}\right)
\end{aligned}
$$

The last step follows again from symmetry of a t distribution with ν degrees of freedom about 0. We will often denote the limits of a $100 \times (1-\alpha)\%$ confidence interval for β_k as $b_k \pm t\left(\alpha/2; n-p\right) s\left\{b_k\right\}$.

> ***Example 3:*** Consider the same data. What is a 95% confidence interval for β_1?

```
lower <- b1 + s.b1*qt(alpha/2, n - 2)
upper <- b1 - s.b1*qt(alpha/2, n - 2)
```

> We obtain a 95% confidence interval of $(2.852, 4.288)$ for β_1.

To conclude, we'll work through one more examples.

> ***Example 4:*** 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. The portrait studio is interested in testing whether or not there is a linear association between the number of persons aged 16 or younger and the sales of portraits of children having accounted for per capita disposabe income, i.e. the null hypothesis H_0: $\beta_1 = 0$ at level $\alpha = 0.05$.

```r
load("~/Dropbox/Teaching/STAT525/Spring2023/bookdata/dwaine.RData")
n <- nrow(data)
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
summary(linmod)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4239  -6.2161   0.7449   9.4356  20.2151
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -68.8571    60.0170  -1.147   0.2663
## X1            1.4546     0.2118   6.868    2e-06 ***
## X2            9.3655     4.0640   2.305   0.0333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.01 on 18 degrees of freedom
## Multiple R-squared:  0.9167, Adjusted R-squared:  0.9075
## F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

From the printed regression results, we can see that we observe a p-value for a test of the null hypothesis H_0: $\beta_1 = 0$ that is less than $\alpha = 0.05$. Accordingly, we conclude H_a: $\beta_1 \neq 0$, i.e. we conclude that there is evidence of a linear association between the number of persons aged 16 or younger and the sales of portraits of children having accounted for per capita disposabe income at level $\alpha = 0.05$.