Notes 7

Maryclare Griffin

3/9/2023

These notes are based on Chapters 1, 5, and 6 of KNNL.

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:

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

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 $\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \ldots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{n,p-1} \end{pmatrix}$ are known constants
- $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ is a random error term with mean $\boldsymbol{E}\left\{\boldsymbol{\epsilon}\right\} = \boldsymbol{0}$ and variance $\boldsymbol{\sigma}^2\left\{\boldsymbol{\epsilon}\right\} = \sigma^2\boldsymbol{I}_n$.

The least squares minimizing estimator $\boldsymbol{b}$ is given by:

$$\boldsymbol{b} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Recall that the variance of the least squares estimator is:

$$\sigma^2\left\{\boldsymbol{b}\right\} = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Without knowing σ^2, the variance of the least squares estimator $\boldsymbol{b}$ is unknown. In practice, we will **estimate** σ^2 in order to estimate the variance of the least squares estimator $\boldsymbol{b}$. Recall that σ^2 is the variance of the random errors, $\boldsymbol{\epsilon}$. We do not observe the random errors, but we can estimate them and, accordingly, σ^2. Given that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and given that $\hat{\boldsymbol{Y}}$ is an unbiased estimator of $\boldsymbol{X}\boldsymbol{\beta}$, a natural estimator of $\boldsymbol{\epsilon}$ is $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. We will refer to $\boldsymbol{e}$ as the **residuals**. Each values of the residuals $e_i = Y_i - \hat{Y}_i$ describes the deviation of the corresponding observed value Y_i from the fitted value $\hat{Y}_i$.

A naive estimator of σ^2 might be obtained by computing the sample variance of the residuals $\boldsymbol{e}$. Letting $\bar{e} = \frac{1}{n}\sum_{i=1}^n e_i$ be the sample mean of the residuals, the sample variance is

$$\frac{1}{n-1}\sum_{i=1}^n \left(e_i - \bar{e}\right)^2.$$

An important property of the residuals $\boldsymbol{e}$ allows us to simplify this expression. Under the linear model with an intercept, i.e. with $X_{i0} = 1$ for all $i = 1, \ldots, n$, $\bar{e} = 0$.

$$\frac{1}{n-1} \sum_{i=1}^{n} e_i^2$$

To assess if this is a good estimator of σ^2, we will take its expectation.

$$E\left\{ \frac{1}{n-1} \sum_{i=1}^{n} e_i^2 \right\} = \frac{1}{n-1} \sum_{i=1}^{n} E\left\{ e_i^2 \right\}.$$

To make this simpler, we're going to use some more linear algebra! We can recognize that

$$\begin{aligned} E\left\{ \frac{1}{n-1} \sum_{i=1}^{n} e_i^2 \right\} &= \frac{1}{n-1} E\left\{ \sum_{i=1}^{n} e_i^2 \right\} \\ &= \frac{1}{n-1} E\left\{ \boldsymbol{e}'\boldsymbol{e} \right\} \end{aligned}$$

We're going to see identity matrices of different dimensions up ahead, so we'll introduce some more notation. Let $\boldsymbol{I}_m$ refer to an $m \times m$ identity matrix. Each $\boldsymbol{e}$ can be written as a product of a matrix of fixed, constant elements and a vector of random variables, specifically $\boldsymbol{Y}$:

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b} \\ &= \boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= \left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\boldsymbol{Y} \\ &= \left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) \\ &= \left(\boldsymbol{X} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{\beta} + \left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\boldsymbol{\epsilon} \\ &= \left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\boldsymbol{\epsilon} \end{aligned}$$

For convenience, we will define $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$, so that we can parsimoniously write $\boldsymbol{e} = \left(\boldsymbol{I}_n - \boldsymbol{H}\right)\boldsymbol{\epsilon}$. Now let's plug this into our expression for $E\left\{ \frac{1}{n-1} \sum_{i=1}^{n} e_i^2 \right\}$.

$$\begin{aligned} E\left\{ \frac{1}{n-1} \sum_{i=1}^{n} e_i^2 \right\} &= \frac{1}{n-1} E\left\{ \left(\left(\boldsymbol{I}_n - \boldsymbol{H}\right)\boldsymbol{\epsilon}\right)'\left(\boldsymbol{I}_n - \boldsymbol{H}\right)\boldsymbol{\epsilon} \right\} \\ &= \frac{1}{n-1} E\left\{ \boldsymbol{\epsilon}'\left(\boldsymbol{I}_n - \boldsymbol{H}\right)'\left(\boldsymbol{I}_n - \boldsymbol{H}\right)\boldsymbol{\epsilon} \right\} \\ &= \frac{1}{n-1} E\left\{ \boldsymbol{\epsilon}'\left(\boldsymbol{I}_n - \boldsymbol{H}\right)\boldsymbol{\epsilon} \right\} \end{aligned}$$

The last step follows from expanding out $\left(\boldsymbol{I}_n - \boldsymbol{H}\right)'\left(\boldsymbol{I}_n - \boldsymbol{H}\right)$, plugging in $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$, simplifying, and rewriting in terms of $\boldsymbol{I}_n$ and $\boldsymbol{H}$.

> ***Note:*** Given a square $r \times r$ matrix $\boldsymbol{A}$ that is symmetric, with $A_{ij} = A_{ji}$ for all $i, j = 1, \ldots, r$, the transpose of $\boldsymbol{A}$ is $\boldsymbol{A}$ itself, i.e. $\boldsymbol{A}' = \boldsymbol{A}$.

We are not quite done, because $E\left\{\boldsymbol{\epsilon}'\left(\boldsymbol{I}_n - \boldsymbol{H}\right)\boldsymbol{\epsilon}\right\}$ is not a quantity we can readily take an expectation of. We only know how to compute expectations involving matrix products when the matrix product has a "sandwich" type form, with the matrix/matrices involving random elements in the middle (the "meat" of the "sandwich") and the matrix/matrices with fixed elements on the outside (the "filling" of the "sandwich") . To rewrite this expectation in a "sandwich" type form, we're going to need the trace of a matrix.

Note: Given a square $r \times r$ matrix $\boldsymbol{A}$, the trace of a matrix $\text{tr}(\boldsymbol{A})$ is the sum of its diagonal elements, $\text{tr}(\boldsymbol{A}) = \sum_{k=1}^{r} A_{kk}$.

Using this fact about the trace, we can recognize that the 1×1 matrix $\boldsymbol{\epsilon}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\epsilon}$ is equal to the trace of the 1×1 matrix $\text{tr}(\boldsymbol{\epsilon}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\epsilon})$. How is this helpful? We need another trace fact.

Note: Given an $r \times c$ matrix $\boldsymbol{A}$, a $c \times s$ matrix $\boldsymbol{B}$, and a $s \times r$ matrix $\boldsymbol{D}$, the trace of the product matrix $\text{tr}(\boldsymbol{ABC})$ has a cyclic property that allows the elements to be rearranged, specifically **cyclically permuted**. This means:

$$\text{tr}(\boldsymbol{ABC}) = \text{tr}(\boldsymbol{BCA}) = \text{tr}(\boldsymbol{CAB})$$

This allows us to go a step further, letting:

$$\begin{aligned} E\left\{\frac{1}{n-1}\sum_{i=1}^{n} e_i^2\right\} &= \frac{1}{n-1} E\left\{\text{tr}\left(\boldsymbol{\epsilon}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\epsilon}\right)\right\} \\ &= \frac{1}{n-1} E\left\{\text{tr}\left((\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right)\right\} \\ &= \frac{1}{n-1} \text{tr}\left((\boldsymbol{I}_n - \boldsymbol{H}) E\left\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right\}\right) \end{aligned}$$

The last step follows from recognizing that the trace of a matrix is just a special sum, so the expectation of the trace of a matrix $E\{\text{tr}(\boldsymbol{A})\}$ is the trace of the expectation, $E\{\text{tr}(\boldsymbol{A})\} = \text{tr}(E\{\boldsymbol{A}\})$.

Now we can really start getting somewhere! Using what we have assumed about the mean and variance of $\boldsymbol{\epsilon}$, we have:

$$\begin{aligned} E\left\{\frac{1}{n-1}\sum_{i=1}^{n} e_i^2\right\} &= \frac{1}{n-1} \text{tr}\left((\boldsymbol{I}_n - \boldsymbol{H})\left(\sigma^2 \boldsymbol{I}_n\right)\right) \\ &= \frac{\sigma^2}{n-1} \text{tr}\left((\boldsymbol{I}_n - \boldsymbol{H})\right) \\ &= \frac{\sigma^2}{n-1} \left(\text{tr}\left(\boldsymbol{I}_n\right) - \text{tr}\left(\boldsymbol{H}\right)\right) \\ &= \frac{\sigma^2}{n-1} \left(n - \text{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\right) \\ &= \frac{\sigma^2}{n-1} \left(n - \text{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\right)\right) \\ &= \frac{\sigma^2}{n-1} \left(n - \text{tr}\left(\boldsymbol{I}_p\right)\right) \\ &= \frac{\sigma^2}{n-1} \left(n - p\right) \\ &= \sigma^2 \left(\frac{n-p}{n-1}\right) \end{aligned}$$

Putting this all together, we have that the expectation of the sample variance of the residuals is $\frac{1}{n-1}\sum_{i=1}^{n} e_i^2 = \sigma^2\left(\frac{n-p}{n-1}\right)$.

This is only **unbiased** for σ^2 when $p = 1$, which is when we do not have any predictors. This suggests that we define the following **unbiased** estimator for σ^2, that is unbiased for any $p \geq 0$

$$\frac{1}{n-p}\sum_{i=1}^{n} e_i^2.$$

We will refer to this estimator going forward as s^2 or MSE, i.e.

$$s^2 = MSE = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2.$$

Remember - we went down this path because the variance of the least squares estimator $\boldsymbol{b}$ involved σ^2, which is unknown. Accordingly, we can now define the estimated variance of the least squares estimator,

$$s^2\{\boldsymbol{b}\} = s^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Putting this all together, this means that we can compute the least squares estimator $\boldsymbol{b}$ given data, know that it is an unbiased for $\boldsymbol{\beta}$, and obtain an estimated variance of $\boldsymbol{b}$ about its mean $\boldsymbol{\beta}$ by computing $s^2\{\boldsymbol{b}\}$.

Example: Let's return to the 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. Previously, we're constructed an design matrix and compute b_0, b_1, and b_2 by hand. Now we're going to go a step further, and compute s^2, the estimated variance of $\boldsymbol{b}$, fitted values $\hat{\boldsymbol{Y}}$, and the estimated variance of the fitted values $\hat{\boldsymbol{Y}}$.

```
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
n <- length(Y) # Record the number of observations
p <- 3 # Record the number of predictors + 1
X <- cbind(rep(1, n), X1, X2)
XtY <- t(X)%*%Y # Compute X'Y
XtX <- t(X)%*%X # Compute X'X
XtX.inv <- solve(XtX) # Invert XtX
b <- XtX.inv%*%XtY # Solve for b
b0 <- b[1] # Extract the intercept
b1 <- b[2] # Extract b1
b2 <- b[3] # Extract b2

Y.hat <- X%*%b # Compute fitted values
e <- Y - Y.hat # Compute residuals

s.sq <- sum(e^2)/(n - p) # Compute unbiased estimator of sigma^2
s.sq.b <- s.sq*XtX.inv # Compute estimated variance of b

s.sq.b0 <- s.sq.b[1, 1] # Extract the estimated variance of b_0
s.sq.b1 <- s.sq.b[2, 2] # Extract the estimated variance of b_1
s.sq.b2 <- s.sq.b[3, 3] # Extract the estimated variance of b_2
```

We obtain estimates $b_0 = -68.857$, $b_1 = 1.455$, and $b_2 = 9.366$, an estimate $s^2 = 121.163$ of the noise variance σ^2, and variance estimates $s^2\{b_0\} = 3602.035$, $s^2\{b_1\} = 0.045$, and $s^2\{b_2\} = 16.516$.

We could have also obtained these variance estimates from `lm` directly!

```
linmod <- lm(Y~X1+X2) # Fit linear model to our data
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
```

The **residual standard error** corresponds to $s = \sqrt{MSE}$, and the corresponding degrees of freedom gives $n - p$ for this dataset. The **standard errors** for each estimated regression coefficient correspond to the square roots of the estimated variances, i.e. the standard error of b_0 is $\sqrt{s^2\{b_0\}} = s\{b_0\}$. We can extract them directly from `lm` going forward, so we don't have to compute them by hand every time.

```
s <- summary(linmod)$s
s.b0 <- summary(linmod)$coef[1, 2]
s.b1 <- summary(linmod)$coef[2, 2]
s.b2 <- summary(linmod)$coef[3, 2]
```

Using similar logic, we can also compute the fitted values $\hat{\boldsymbol{Y}} = \boldsymbol{Xb}$, know that they are unbiased for $\boldsymbol{X\beta}$, and obtain an estimated variance of $\hat{\boldsymbol{Y}}$ about its mean $\boldsymbol{X\beta}$.

To obtain an estimated variance of $\hat{\boldsymbol{Y}}$, we first compute the variance of the fitted values:

$$\begin{aligned}
\sigma^2\left\{\hat{\boldsymbol{Y}}\right\} =& \sigma^2\left\{\hat{\boldsymbol{Y}}\right\} \\
=& \sigma^2\left\{\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}\right\} \\
=& \sigma^2\left\{\boldsymbol{HY}\right\} \\
=& \sigma^2\left\{\boldsymbol{H\epsilon}\right\} \\
=& \boldsymbol{H}\sigma^2\left\{\boldsymbol{\epsilon}\right\}\boldsymbol{H}' \\
=& \boldsymbol{H}\left(\sigma^2\boldsymbol{I}_n\right)\boldsymbol{H}' \\
=& \sigma^2\boldsymbol{HH}' \\
=& \sigma^2\boldsymbol{H}
\end{aligned}$$

It follows that we can define an estimated variance of $\hat{\boldsymbol{Y}}$ about its mean $\boldsymbol{X\beta}$ to be

$$s^2\left\{\hat{\boldsymbol{Y}}\right\} = s^2\boldsymbol{H}$$

Going back to the data, let's compute the estimated variance of the standard errors.

```
H <- X%*%XtX.inv%*%t(X)
s.sq.Y.hat <- s.sq*H
```

There are a lot of fitted values to summarize, so let's just zoom in on one - the fitted value for the second observation $\hat{Y}_2$ and its estimated variance $s^2\left\{\hat{Y}_2\right\}$

```
Y.hat.2 <- Y.hat[2]
s.sq.Y.hat.2 <- s.sq.Y.hat[2, 2]
```

We obtain a fitted value of $\hat{Y}_2 = 154.229$ and an estimated variance of $s^2\left\{\hat{Y}_2\right\} = 12.643$.

If we don't want to do this by hand, the `lm` function alone does not directly return the estimated variances of the fitted values. To obtain them, we need to apply the `predict` function to the saved output from `lm`.

```
pred <- predict(linmod, se.fit = TRUE) # Obtain fitted values and their standard errors
Y.hat <- pred$fit # Extract fitted values
s.Y.hat <- pred$se.fit # Extract standard errors
```

The standard errors returned by `predict` correspond to the square roots of the estimated variances we computed earlier, i.e. they correspond to $\sqrt{s^2\{\boldsymbol{Y}\}} = s\{\boldsymbol{Y}\}$.

To wrap up, we'll introduce one more concept! Often, it can be useful to summarize how much of the observed variability in the response can be attributed to the variability of the fitted values. We can think of this as summarizing what proportion of the variation in the data is explained by our linear model. We refer to this as R^2, and it is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}$$

The denominator $\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$ is often called the **total sum of squares**, and the numerator $\sum_{i=1}^{n} e_i^2$ is often called the **residual sum of squares** or SSR, for sum of squared residuals.

```
sse <- sum(e^2)
ssto <- sum((Y - mean(Y))^2)
r.sq <- 1 - sse/ssto
```

We obtain $R^2 = 0.917$, which tells us that approximately 92% of the variability in sales of portraits of children in the Dwaine data can be attributed to variability in the number of persons aged 16 or younger in a city and per capita disposable income.

R^2 is also included in the output of `lm`. We can find it above in the printed summary, and we can also extract it from the `lm` object.

```
r.sq <- summary(linmod)$r.sq
```