Regression Analysis - STAT 482 - HW #6

Alex Towell (atowell@siue.edu)

Problem 1

Consider the simple linear regression model with inputs centered so that $\sum x_i = 0$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n,$$

where

$$\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Part (a)

Write the model using matrix notation.

See page 5 for remarks on notation.

The model may be written as

$$Y = X\beta + \epsilon,$$

where

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

is a $n \times 1$ response vector,

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

is a $2 \times n$ input (design) matrix,

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

is a 2×1 parameter vector, and

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

is a $n \times 1$ multivariate normal vector such that $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I)$.

Part (b)

Use matrix multiplication to derive a simplified expression for each of the following:

(i) b_0,b_1

(ii) $\text{var}(b_0)$, $\text{var}(b_1)$, $\text{cov}(b_0, b_1)$

(iii) $\text{var}(\hat{y}_h)$.

(i) b_0 and b_1

The least squares estimator for simpler linear regression is given by

$$(b_0\ b_1)' = (X'X)^{-1}X'Y.$$

To simplify the RHS, first we compute

$$X'X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}.$$

By the constraint $\sum x_i = 0$, this simplifies to

$$X'X = \begin{bmatrix} n & 0 \\ 0 & \sum x_i^2 \end{bmatrix},$$

which has the inverse

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum x_i^2} \end{bmatrix}.$$

Next, we compute

$$X'Y = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}.$$

Finally, we compute the product of $(X'X)^{-1}$ and $X'Y$,

$$\begin{aligned}(b_0\ b_1)' = (X'X)^{-1}X'Y &= \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum x_i^2} \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\ &= \begin{bmatrix} \bar{y} \\ \frac{\sum x_i y_i}{\sum x_i^2} \end{bmatrix}.\end{aligned}$$

(ii) $\mathrm{var}(b_0)$, $\mathrm{var}(b_1)$, $\mathrm{cov}(b_0, b_1)$

Let A be a constant matrix and B be a random vector. By the property that $\mathrm{cov}(AB) = A\,\mathrm{cov}(B)A'$,

$$\begin{aligned}\mathrm{cov}(b) &= \mathrm{cov}((X'X)^{-1}X'Y) \\ &= (X'X)^{-1}X'\,\mathrm{var}(Y)((X'X)^{-1}X')' \\ &= \sigma^2(X'X)^{-1}X'((X'X)^{-1}X')'.\end{aligned}$$

By the properties that $(AB)' = B'A'$, $C' = C$ if C is symmetric, and $D^{-1}D = DD^{-1} = I$ if D is non-singular,

$$\begin{aligned}\mathrm{cov}(b) &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \\ &= \sigma^2 \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum x_i^2} \end{bmatrix}.\end{aligned}$$

By definition,

$$\mathrm{var}(b_0) = \mathrm{cov}(b)_{11} = \frac{\sigma^2}{n},$$

$$\mathrm{var}(b_1) = \mathrm{cov}(b)_{22} = \frac{\sigma^2}{\sum x_i^2},$$

and

$$\mathrm{cov}(b_0, b_1) = \mathrm{cov}(b)_{12} = 0.$$

(iii) $\text{var}(\hat{Y}_h)$

Given that $\hat{Y}_h = b_0 + b_1 x_h$, we may rewrite this as $\hat{Y}_h = \mathbf{x_h'} b$ where $\mathbf{x_h} = [1\ x_h]'$ and $b = [b_0\ b_1]'$. Thus,

$$\begin{aligned}
\text{var}(\hat{Y}_h) &= \text{cov}(\mathbf{x_h'} b) \\
&= \mathbf{x_h'}\, \text{cov}(b) \mathbf{x_h} \\
&= \begin{bmatrix} 1 & x_h \end{bmatrix} \sigma^2 \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum x_i^2} \end{bmatrix} \begin{bmatrix} 1 \\ x_h \end{bmatrix} \\
&= \sigma^2 \begin{bmatrix} 1 & x_h \end{bmatrix} \begin{bmatrix} \frac{1}{n} \\ \frac{x_h}{\sum x_i^2} \end{bmatrix} \\
&= \sigma^2 \left[\frac{1}{n} + \frac{x_h^2}{\sum x_i^2} \right].
\end{aligned}$$

Problem 2

Refer to data from Exercise 6.15

The goal is to study the relationship between patient satisfaction (y) and patient's age (x_1), severity of illness (x_2), and anxiety level (x_3).

> Part (a)
>
> Provide an interpretation of a regression coefficient in a multiple regression model.

Let

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Then,

$$\frac{\partial E(Y|x_1, \ldots, x_3)}{\partial x_i} = \beta_i,$$

which means we may interpret the beta coefficients as representing partial effects.

In particular, it represents the difference in the mean response (satisfaction level) from a one unit increase in x_i, with all other input levels held fixed.

> Part (b)
>
> Compute b, the estimated regression coefficients for the patient satisfaction data.

```
data = read.table('CH06PR15.txt')
names(data) = c("age","severity","anxiety","satisfaction")
head(data)
```

age	severity	anxiety	satisfaction
48	50	51	2.3
57	36	46	2.3
66	40	48	2.2
70	41	44	1.8
89	28	43	1.8
36	49	54	2.9

```r
mod = lm(satisfaction ~ age + severity + anxiety, data=data)
summary(mod)
```

```
##
## Call:
## lm(formula = satisfaction ~ age + severity + anxiety, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33589 -0.13333 -0.03347  0.12599  0.52022
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.053245   0.613791   1.716  0.09354 .
## age         -0.005861   0.003089  -1.897  0.06468 .
## severity     0.001928   0.005787   0.333  0.74065
## anxiety      0.030148   0.009257   3.257  0.00223 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2098 on 42 degrees of freedom
## Multiple R-squared:  0.5415, Adjusted R-squared:  0.5088
## F-statistic: 16.54 on 3 and 42 DF,  p-value: 3.043e-07
```

We see that $b = (1.053, -0.006, 0.002, 0.03)'$. Our fitted model is thus given by

$$\hat{E}(Y|x) = x'b = 1.053 - .006x_1 + .002x_2 + .03x_3$$

where $x = [1\ x_1\ x_2\ x_3]'$.

Part (c)

Provide a comment on the direction of the input effects, stated in the context of the problem.

Analyzing b, we see that age has a negative effect on satisfaction and severity of illness and anxiety have a positive effect on satisfaction.

Part (d)

Compute $\widehat{\text{cov}}(b)$, the estimated covariance matrix for the regression coefficients.

```r
knitr::kable(vcov(mod), digits=5, caption='Covariance matrix')
```

Table 2: Covariance matrix

	(Intercept)	age	severity	anxiety
(Intercept)	0.37674	-0.00149	-0.00152	-0.00447
age	-0.00149	0.00001	0.00001	0.00001
severity	-0.00152	0.00001	0.00003	-0.00001
anxiety	-0.00447	0.00001	-0.00001	0.00009

For instance, we see that $\text{var}(b_0) = 0.37674$ and $\text{cov}(b_0, b_1) = -0.00149$.

Notation

To denote the (i, j)-th element of a matrix A, we may use the notation A_{ij}. To denote the i-th row, we may use $A_{i:}$. To denote the j-th column of a matrix, we may use $A_{:j}$. If A is a column vector, then $A_i \coloneqq A_{i1}$.

If we explicitly denote the elements of a matrix with, say, $A \coloneqq [a_{ij}]$, then a_{12} denotes the $(1, 2)$-th element of A.

The transpose of an object A is denoted by A'. The inverse of A is denoted by A^{-1}. If $*$ is a binary operator, then $*$ applied to the ordered pair (A, B) may be denoted by $A * B$, e.g., $+(A, B) \coloneqq A + B$. The product of A and B is denoted by AB.

I do not distinguish the type of a mathematical object by typography, except when there is already an established convention, like using $\mathbb{R}$ to denote the set of real numbers. Otherwise, instead of using $\mathbf{A}$ of type $\mathbb{R}^{n \times p}$ to denote a matrix, I may just use A. I also do not hold to the convention of only using uppercase letters to denote matrices, e.g., a, θ, or Θ may denote any kind of object, matrix, vector, set, number, tuple, and so on.

The pay-off comes by not having to think about how to consistently convey type, and instead depend on the context to determine the type of an object. When we introduce random variations of each of these types, combined with their corresponding estimators, using a consistent notation that allows one to infer the type of the object with just the typography becomes a losing proposition.

In many ways, I prefer the way type information is provided by programming languages, where recursive types may be easily specified using descriptive terms. In a mathematical paper, often my "programming" language of choice is English.