STAT 608 HW 6

Jack Cunningham (jgavc@tamu.edu)

11/15/24

1)

a)

$$\Sigma = \begin{bmatrix} 49 & 5 & 4 \\ 5 & 25 & 0 \\ 4 & 0 & 9 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & \frac{1}{7} & \frac{4}{21} \\ \frac{1}{7} & 1 & 0 \\ \frac{4}{21} & 0 & 1 \end{bmatrix}$$

b)

That formula can be generalized to:

$$Var(a^T X) = a^T \Sigma a$$

Where X is the matrix containing x_1, x_2, x_3, Σ is the covariance matrix and a is length three vector containing the coefficients on x_1, x_2, x_3.

```
Sigma = matrix(c(49,5,4,5,25,0,4,0,9),nrow = 3, ncol = 3)
Sigma
```

```
     [,1] [,2] [,3]
[1,]   49    5    4
[2,]    5   25    0
[3,]    4    0    9
```

1.

$a = (2, -2, 0)$

```
a_1 <- c(2,-2,0)
as.numeric(t(a_1)%*%Sigma%*%a_1)
```

```
[1] 256
```

2.

$a = (2, -1, 5)$

```
a_2 <- c(2,-1,5)
as.numeric(t(a_2)%*%Sigma%*%a_2)
```

```
[1] 506
```

3.

$a = (-1, 1, 2)$

```
a_3 <- c(-1,1,2)
as.numeric(t(a_3)%*%Sigma%*%a_3)
```

```
[1] 84
```

4.

$a = (-\beta_1, -\beta_2, 1)$

$$\text{Var}(a_1X_1 + a_2X_2 + a_3X_3) = 49\beta_1^2 + 25\beta_2^2 + 9 + 10\beta_1\beta_2 - 8\beta_1$$

Let's say we have the multiple linear model:

$$y = \beta_1X_1 + \beta_2X_2 + e$$

If we solve for e_i we have:

$$e = y - \beta_1X_1 - \beta_2X_2$$

This is the identical linear combination from before except this time $X_3 = y$. In this setting we naturally want to find the minimum of $\text{Var}(e)$, as it would give us the least square estimates of β_1, β_2.

c)

After taking derivatives with the respect of b_1 and b_2 and setting equal to zero we have the two normal equations:

$$2b_1\text{Var}(X_1) + 2b_2\text{Cov}(X_1, X_2) - 2\text{Cov}(X_1, X_3) = 0$$

$$2b_2\text{Var}(X_2) + 2b_1\text{Cov}(X_1, X_2) - 2\text{Cov}(X_2, X_3) = 0$$

Or in our case:

$$98b_1 + 10b_2 - 8 = 0$$

$$10b_1 + 50b_2 = 0$$

d)

$$2\begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \text{Cov}(X_1, X_3) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) & \text{Cov}(X_2, X_3) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$2\begin{bmatrix} 49 & 5 & 4 \\ 5 & 25 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Which has solution, $b_1 = \frac{1}{12}, b_2 = -\frac{-1}{60}$.

2)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

We have design matrix X:

$$\begin{bmatrix} 1_n & x_1 & x_2 \end{bmatrix}$$

a)

When we compute X^TX we have:

$$X^TX = \begin{bmatrix} 1_n \\ x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} 1_n & x_1 & x_2 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 \end{bmatrix}$$

Knowing that the mean of x_1 and x_2 is zero we can say $\sum_{i=1}^n x_{1i} = \sum_{i=1}^n x_{2i} = 0$.

Knowing that the length of x_1 and x_2 is one we can say $\sum_{i=1}^n x_{1i}^2 = \sum_{i=1}^n x_{2i}^2 = 1$.

We can rewrite $\sum_{i=1}^n x_{1i}x_{2i}$ as $\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$ since $\bar{x}_1 = \bar{x}_2 = 0$. That is the numerator of the sample correlation between x_1 and x_2.

The denominator of the sample correlation between x_1 and x_2 is $\sqrt{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2}$.

Using $\bar{x}_1 = \bar{x}_2 = 0$, we then have $\sqrt{\sum_{i=1}^n x_{1i}^2 \sum_{i=1}^n x_{2i}^2}$ both terms are equal to 1 since the length of x_1 and x_2 is one.

With that we have shown:

$$\sum_{i=1}^n x_{1i}x_{2i} = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2}} = \rho$$

So:

$$X^TX = \begin{bmatrix} n & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}$$

b)

We can find the inverse by using row operations:

$$\left[\begin{array}{ccc:ccc} n & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & \rho & 0 & 1 & 0 \\ 0 & \rho & 1 & 0 & 0 & 1 \end{array}\right]$$

We subtract ρ(row 2) from row 3:

$$\left[\begin{array}{ccc:ccc} n & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & \rho & 0 & 1 & 0 \\ 0 & 0 & 1-\rho^2 & 0 & -\rho & 1 \end{array}\right]$$

We divide row 3 by $(1-\rho^2)$:

$$\left[\begin{array}{ccc:ccc} n & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & \rho & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{array}\right]$$

We subtract ρ(row 3) from row 2:

$$\left[\begin{array}{ccc:ccc} n & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1+\frac{\rho^2}{1-\rho^2} & -\frac{\rho}{1-\rho^2} \\ 0 & 0 & 1 & 0 & -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{array}\right]$$

We divide row 1 by n:

$$\left[\begin{array}{ccc:ccc} 1 & 0 & 0 & 1/n & 0 & 0 \\ 0 & 1 & 0 & 0 & 1+\frac{\rho^2}{1-\rho^2} & -\frac{\rho}{1-\rho^2} \\ 0 & 0 & 1 & 0 & -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{array}\right]$$

So:

$$(X^TX)^{-1} = \begin{bmatrix} 1/n & 0 & 0 \\ 0 & \frac{1}{1-\rho^2} & -\frac{\rho}{1-\rho^2} \\ 0 & -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{bmatrix}$$

And with $c = \frac{1}{1-\rho^2}$ we have shown:

$$(X^TX)^{-1} = \begin{bmatrix} 1/n & 0 & 0 \\ 0 & c & -c\rho \\ 0 & -c\rho & c \end{bmatrix}$$

c)

We know:

$$Var(\beta) = \sigma^2(X^TX)^{-1}$$

Then:

$$Var(\hat{\beta}_1 - \hat{\beta}_2) = \sigma^2 \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1/n & 0 & 0 \\ 0 & c & -c\rho \\ 0 & -c\rho & c \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = 2\sigma^2(c + c\rho) = 2\sigma^2(\frac{\rho + 1}{1 - \rho^2}) = 2\sigma^2(\frac{1}{1 - \rho})$$

d)

The Variance of $\hat{\beta}_1$ is:

$$Var(\hat{\beta}_1) = \sigma^2 \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/n & 0 & 0 \\ 0 & c & -c\rho \\ 0 & -c\rho & c \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \sigma^2 c = \sigma^2 \frac{1}{1 - \rho^2}$$

The Variance of $\hat{\beta}_2$ is:

$$Var(\hat{\beta}_2) = \sigma^2 \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/n & 0 & 0 \\ 0 & c & -c\rho \\ 0 & -c\rho & c \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \sigma^2 c = \sigma^2 \frac{1}{1 - \rho^2}$$

So:

$$Var(\hat{\beta}_1) = Var(\hat{\beta}_2) = \sigma^2 \frac{1}{1 - \rho^2}$$

Then we solve:

$$\sigma^2 \frac{1}{1 - \rho^2} = 5\sigma^2$$

$$1 - \rho^2 = \frac{1}{5}$$

$$\rho = \sqrt{4/5}$$

When $\rho \geq \sqrt{4/5}$ $Var(\beta_1)$ and $Var(\beta_2)$ are greater than $5\sigma^2$.

e)

In part d we see that as the correlation between X_1 and X_2 increases the variance of their coefficients increases. This is the issue with multicollinearity, estimates of coefficients become unstable if correlation between the explanatory variables are high. It impacts our ability to perform inference and often lends towards counter intuitive results.

3)

a)

The R^2_{adj} criterion would select either X_1, X_2 or X_1, X_2, X_3 since each subset has $R^2_{\text{adj}} = 1$.

The AIC criterion would select X_1, X_2 as it has the lowest AIC of -316.2008.

THE BIC criterion would also select X_1, X_2 as it has lowest BIC of -317.3725.

b)

The forward selection method based on AIC would select X_3 as its model.

The backward selection method based on BIC would select X_3 as its model as well.

c)

The difference is due to the details of the forward selection procedure.

In forward selection the first step is to find the predictor that minimizes AIC, in this case that would be X_3. The next step finds another predictor to pair with X_3 , but there is no acceptable choice as both X_1 and X_2 do not reduce the AIC/BIC of the model.

The forward selection method does not have the ability to look past the current step, if say the full relationship in Y is described by X_1, X_2, X_3 we cannot look past X_3, X_2 and X_3, X_1 to get to that point.

On the other hand the all possible subsets method checks each combination of predictors, with p = 3 this is a reasonable approach but with a large amount of predictors this becomes unfeasible as we would have to build 2^p models.

d)

In this case I'd say it depends on the goal. If it is prediction I would use X_1, X_2, this combination seems to capture the full relationship in Y. An issue is the relationship between X_1 and X_2, they are essentially perfectly negatively correlated. This would make interpretation difficult as the model struggles with multicollinearity. So if the goal is interpretation I would use X_3, it explains most of the variance in Y and would be simpler to understand.

4)

a)

The VIF version for $Var(\hat{\beta}_j)$ is:

$$Var(\hat{\beta}_j) = \frac{1}{1 - r_{12}^2} \frac{\sigma^2}{(n-1)S_{xj}^2}$$

Where S_{xj}^2 is the squared sample standard deviation of x_j so:

$$S_{xj}^2 = \frac{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2}{n-1}$$

In our case we know that $\bar{x}_j = 0$ so:

$$S_{xj}^2 = \frac{\sum_{i=1}^{n} x_{ji}^2}{n-1}$$

We also know that x_j has length 1, so:

$$S_{xj}^2 = \frac{1}{n-1}$$

When we substitute that back into the VIF version of $Var(\hat{\beta}_j)$ we have:

$$Var(\hat{\beta}_j) = \sigma^2 \frac{1}{1 - r_{12}}$$

This is equivalent to what we find from $\sigma^2(X^T X)^{-1}$:

$$Var(\hat{\beta}_1) = Var(\hat{\beta}_2) = \sigma^2 \frac{1}{1 - r_{12}^2}$$

b)

With:

$$Var(\hat{\beta}_1) = Var(\hat{\beta}_2) = \sigma^2 \frac{1}{1 - r_{12}^2}$$

We can see that when X_1 and X_2 are either highly positively or highly negatively correlated the variance of our estimated coefficients $\hat{\beta}_1, \hat{\beta}_2$ will be large.

5.

In this case we have the design matrix X:

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 1 \end{bmatrix}$$

The VIF for x_1 is:

$$\text{VIF}_1 = \frac{1}{1 - r_{12}^2}$$

Where r_{12} is the sample correlation between x_1, x_2:

$$r_{12} = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}}$$

We have $\bar{x}_1 = 2$ and $\bar{x}_2 = 1/2$. Then we can find the vectors $(x_1 - \bar{x}_1)$ and $(x_2 - \bar{x}_2)$:

$$(x_1 - \bar{x}_1) = \begin{bmatrix} -1 \\ 0 \\ 1 \\ -1 \\ 0 \\ 1 \end{bmatrix}, (x_2 - \bar{x}_2) = \begin{bmatrix} -1/2 \\ -1/2 \\ -1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

We can then rewrite the numerator of sample correlation as:

$$\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = (x_1 - \bar{x}_1)^T(x_2 - \bar{x}_2)$$

Those two vectors are orthogonal, their dot product is equal to 0. So $r_{12} = 0$. Then VIF factor for x_1 is:

$$\text{VIF}_1 = \frac{1}{1 - r_{12}^2} = 1$$

b)

Now the design matrix X is:

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 4 & 1 \end{bmatrix}$$

```r
x_1 <- c(1,2,3,2,3,4)
x_2 <- c(0,0,0,1,1,1)
y <- c(1,2,3,4,5,6)
fit <- lm(y~x_1+x_2)
```

```r
library(car)
```

```
Warning: package 'car' was built under R version 4.3.3
```

```
Loading required package: carData
```

```
Warning: package 'carData' was built under R version 4.3.3
```

```r
vif(fit)[1]
```

```
Warning in summary.lm(object, ...): essentially perfect fit: summary may be
unreliable
```

```
  x_1
1.375
```

The VIF for x_1 is 1.375 now , this is larger than the VIF for x_1 in part a. This is because the correlation between x_1, x_2 is greater than 0:

```r
cor(x_1,x_2)
```

```
[1] 0.522233
```

And the formula for VIF is:

$$\text{VIF}_j = \frac{1}{1 - r_{12}^2}$$

As correlation strengthens VIF will increase, the smallest VIF possible is when x_1, x_2 are uncorrelated which is the case in part a.