Exam 1

Ayush Kumar

3/25/2021

1. Consider the model

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}$$

where $\vec{\epsilon} \sim MVN(\vec{0}, \sigma^2 D)$ with

$$D = \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix}$$

a. Notice that PDF for $\vec{\epsilon} = \frac{1}{\sqrt{(2\pi)^n|\sigma^2 D|}} exp(-\frac{1}{2\sigma^2}\vec{\epsilon}' D^{-1}\vec{\epsilon})$ or $L(\vec{\epsilon}) = \frac{1}{\sqrt{(2\pi)^n \sigma^{2n} d^n}} exp(-\frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})' D^{-1}(\vec{y} - X\vec{\beta})) = (2\pi\sigma^2 d)^{-n/2} exp(-\frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})' D^{-1}(\vec{y} - X\vec{\beta}))$. Hence the log-likelihood
$\mathcal{L}(\vec{\epsilon}) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{n}{2}log(d) - \frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})' D^{-1}(\vec{y} - X\vec{\beta})$
$\mathcal{L}(\vec{\epsilon}) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{n}{2}log(d) - \frac{1}{2\sigma^2}(\vec{y}' - \vec{\beta}' X') D^{-1}(\vec{y} - X\vec{\beta})$
$\mathcal{L}(\vec{\epsilon}) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{n}{2}log(d) - \frac{1}{2\sigma^2}(\vec{y}' D^{-1}\vec{y} - \vec{y}' D^{-1} X\beta - \beta' X' D^{-1}\vec{y} + \beta' X' D^{-1} X\beta)$
Now to maximize likelihood we set $\frac{\partial \mathcal{L}(\vec{\epsilon})}{\partial \beta} = 0$, to obtain $\vec{y}'(D^{-1})X = \beta' X' D^{-1} X$ or simply

$$\hat{\beta} = (X'(D^{-1})'X)^{-1}(X'(D^{-1})')\vec{y}$$

b. Now we estimate the $E(\hat{\beta})$, and $var(\hat{\beta})$: Notice that
$E(\hat{\beta}) = E((X'(D^{-1})'X)^{-1}(X'(D^{-1}))'\vec{y})$
$E(\hat{\beta}) = E(X^{-1}(D(X')^{-1}X'D^{-1})\vec{y})$
$E(\hat{\beta}) = E(X^{-1}\vec{y})$
$E(\hat{\beta}) = \beta$
Now see that
$var(\hat{\beta}) = E(X'(D^{-1})'X)^{-1}(X'(D^{-1})')\vec{y})$
$var(\hat{\beta}) = (X'(D^{-1})'X)^{-1}(X'(D^{-1})')\sigma^2 D(X'(D^{-1})')'((X'(D^{-1})'X)^{-1})'$
$var(\hat{\beta}) = \sigma^2(X'(D^{-1})'X)^{-1}$

c. Now we shall show that $\hat{\beta}$ is BLUE: Let $\tilde{\beta} = M\vec{y}$, with $var(\vec{y}) = \sigma^2 D$, now let $K = M - (X'(D^{-1})'X)^{-1}(X'(D^{-1})')$ thus, $M = K + (X'(D^{-1})'X)^{-1}(X'(D^{-1})')$. Hence as $\tilde{\beta}$ is unbiased, we have
$E(\tilde{\beta}) = E((K + (X'(D^{-1})'X)^{-1}(X'(D^{-1})'))\vec{y})$
$E(\tilde{\beta}) = E((K + (X'(D^{-1})'X)^{-1}(X'(D^{-1})'))X\beta)$
$E(\tilde{\beta}) = \beta + KX\beta$
Thus, we have $KX = 0 = X'K'$
Now notice that $var(\tilde{\beta}) = M\sigma^2 DM' = \sigma^2 MDM'$
Now, $MDM' = (K + (X'(D^{-1})'X)^{-1}(X'(D^{-1})'))D(K + (X'(D^{-1})'X)^{-1}(X'(D^{-1})'))'$
$MDM' = (KD + (X'(D^{-1})'X)^{-1}(X'(D^{-1})')D)(D^{-1}X(X'(D)^{-1}X)^{-1} + K')$
$MDM' = (KD + (X'(D^{-1})'X)^{-1}(X')(D^{-1}X(X'(D)^{-1}X)^{-1} + K')$
$MDM' = KDD^{-1}X(X'(D^{-1})X)^{-1} + KDK^{-1} + (X'(D^{-1})'X)^{-1}(X'D^{-1}X)(X'(D^{-1})X)^{-1} +$

$(X'(D^{-1})'X)^{-1}X'K'$
$MDM = KDK^{-1} + (X'(D^{-1})X)^{-1}$
Hence, $var(\tilde{\beta}) = \sigma^2(KDK^{-1} + (X'(D^{-1})X)^{-1})$, now notice in an argument similar to HW, and other-places each variance, then, is minimized by the corresponding with the row of being identically 0. Which, then gives us $\hat{\beta}$ being BLUE.

2. We conduct a hypothesis test, with the null hypothesis that machete, and wood-bow are equally effective, and the alternative hypothesis that machete is better than wood-bow.

```
dr <- read.table('http://math.uttyler.edu/nathan/data/digging-rates.data',header=TRUE)
mutate(dr,diff=machete-wood.bow)->dr
t.test(dr$machete,dr$subject,alternative='greater',paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  dr$machete and dr$subject
## t = 6.6, df = 4, p-value = 0.001
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  111.9   Inf
## sample estimates:
## mean of the differences
##                   165.2
```

Notice that the p-value is very close to 0 thus, we conclude that machete is indeed better than wood-bow.

3.

a. Here is the code, and summary table:

```
census<- read.table('http://math.uttyler.edu/nathan/data/census.data',header=TRUE)
m1<-lm(census$undercount~census$minority+census$highschool)
summary(m1)
```

```
##
## Call:
## lm(formula = census$undercount ~ census$minority + census$highschool)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.16  -1.18   0.04   1.04   4.04
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.7113     0.8691    3.12   0.0027 **
## census$minority     0.1237     0.0138    8.96  7.4e-13 ***
## census$highschool  -0.0950     0.0285   -3.34   0.0014 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.65 on 63 degrees of freedom
## Multiple R-squared:  0.57,   Adjusted R-squared:  0.556
## F-statistic: 41.7 on 2 and 63 DF,  p-value: 2.94e-12
```

the model is $Undercount_i = 2.7113 + 0.1237 minority_i - 0.0950 highschool_i$.

b. Based on our sample, assuming no change in high-school variable we expect the a unit increase in minority variable resulting in 0.124 unit increase in undercount variable, further assuming no change in minority we expect the a unit increase in high-school variable resulting in 0.0950 unit decrease in undercount variable. Given that both minority, and highschool variables are zero then undercount equals 2.7113.
c. The r^2 is 0.57, meaning the given model, can explain 57% variance in the undercount variable.

4. Consider data in *census.data*.

a. Here is the code:

```
m1<-lm(census$undercount~census$poverty+census$language+census$poverty*census$city+census$language*censu
m2<-lm(census$undercount~census$poverty+census$language+census$city)
anova(m2,m1)
```

```
## Analysis of Variance Table
##
## Model 1: census$undercount ~ census$poverty + census$language + census$city
## Model 2: census$undercount ~ census$poverty + census$language + census$poverty *
##     census$city + census$language * census$city
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     62 194
## 2     60 184  2        10 1.63    0.2
```

Notice that the Anova test had
H_o=The variables generated by census$poverty * census$city + census$language * census$city are jointly 0.
H_a=The variables generated by census$poverty * census$city + census$language * census$city are jointly not 0.
the p-value is 0.2, hence we fail to reject the null hypothesis, and conclude that there is no evidence that city makes any difference on the effect of poverty/language on undercount. The reason why we don't have any impact is because the variables poverty/language, and city are highly correlated thus, the variance explained by the additional term is already presnt in the slope of the variables.
b. Here is the summary table,

```
summary(m1)
```

```
##
## Call:
## lm(formula = census$undercount ~ census$poverty + census$language +
##     census$poverty * census$city + census$language * census$city)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.252 -1.080  0.064  0.789  5.551
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.46955    2.21927   -0.21    0.833
## census$poverty                0.23079    0.11417    2.02    0.048 *
## census$language               0.30211    0.12574    2.40    0.019 *
## census$citystate                  1.13422    2.39842    0.47    0.638
## census$poverty:census$citystate  -0.23605    0.13266   -1.78    0.080 .
## census$language:census$citystate  0.00821    0.21743    0.04    0.970
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.75 on 60 degrees of freedom
## Multiple R-squared:  0.536,  Adjusted R-squared:  0.497
## F-statistic: 13.9 on 5 and 60 DF,  p-value: 5.38e-09
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = census$undercount ~ census$poverty + census$language +
##     census$city)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.116 -1.073  0.163  0.936  5.252
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.6985     1.2520    2.16  0.03502 *
## census$poverty     0.0554     0.0587    0.94  0.34867
## census$language    0.2858     0.1031    2.77  0.00734 **
## census$citystate  -2.7382     0.6865   -3.99  0.00018 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.77 on 62 degrees of freedom
## Multiple R-squared:  0.511,  Adjusted R-squared:  0.487
## F-statistic: 21.6 on 3 and 62 DF,  p-value: 1.11e-09
```

we note that the p-value for poverty in m1 was 0.0441 while that in m2 was 0.34867, this shows that the variable city added to m2 subsumes the effect of poverty. This happens as city, poverty are highly correlated see below:

```
summary(lm(census$poverty~census$city))
```

```
##
## Call:
## lm(formula = census$poverty ~ census$city)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6.19  -3.09  -1.10   2.83  11.78
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        17.694      0.952    18.6  < 2e-16 ***
## census$citystate   -5.578      1.094    -5.1  3.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.81 on 64 degrees of freedom
## Multiple R-squared:  0.289,  Adjusted R-squared:  0.278
## F-statistic:   26 on 1 and 64 DF,  p-value: 3.26e-06
```

meaning when the city in m1 captures the effect of poverty on undercount. To check if poverty matters to undercount we consider the following:

```
m1<-lm(census$undercount~census$language+census$city)
m2<-lm(census$undercount~census$poverty+census$language+census$city)
anova(m2,m1)
```

```
## Analysis of Variance Table
##
## Model 1: census$undercount ~ census$poverty + census$language + census$city
## Model 2: census$undercount ~ census$language + census$city
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     62 194
## 2     63 197 -1     -2.79 0.89   0.35
```

Now we note that post addition of city the variable poverty is non-significant to explain the varience of undercount.