# STAT 608 HW 5

Jack Cunningham (jgavc@tamu.edu)

11/1/24

1)

The linear model is:

$$Y = X\beta + e$$

With design matrix X:

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

And response vector Y:

$$Y = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 \end{bmatrix}^T$$

And coefficient vector $\beta$:

$$\beta = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix}^T$$

b)

$$X^T X = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} y_1 + y_4 + y_5 + y_7 \\ y_2 + y_4 + y_6 + y_7 \\ y_3 + y_5 + y_6 + y_7 \end{bmatrix}$$

c)

We know that $(X^T X)(X^T X)^{-1} = I_3$ so $A(X^T X)(X^T X)^{-1} = I_3$:

$$A(X^T X)(X^T X)^{-1} = c \begin{bmatrix} 8 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 8 \end{bmatrix} = I_3$$

So $c = \frac{1}{8}$.

d)

$$\hat{\beta} = (X^T X)^{-1} X^T y = \frac{1}{8} \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix} \begin{bmatrix} y_1 + y_4 + y_5 + y_7 \\ y_2 + y_4 + y_6 + y_7 \\ y_3 + y_5 + y_6 + y_7 \end{bmatrix} = \frac{1}{8} \begin{bmatrix} 3y_1 - y_2 - y_3 + 2y_4 + 2y_5 - 2y_6 + y_7 \\ -y_1 + 3y_2 - y_3 + 2y_4 - 2y_5 + 2y_6 + y_7 \\ -y_1 - y_2 + 3y_3 - 2y_4 + 2y_5 + 2y_6 + y_7 \end{bmatrix}$$

e)

If $e_i = \sigma^2/n_i$, then $w_i = 1/n_i$. Where $n_i$ is the number of drugs in each measurement. Then weight matrix W is:

$$W = \text{diag}(1, 1, 1, 1/2, 1/2, 1/2, 1/3)$$

f)

The normal equation for weighted least squares is $\hat{\beta}_{\text{WLS}} = (X^T W X)^{-1} X^T W y$.

```
W = diag(x = c(1,1,1,1/2,1/2,1/2,1/3))
X = matrix(c(1,0,0,1,1,0,1,0,1,0,1,0,1,1,0,0,1,0,1,1,1),nrow = 7, ncol = 3)
XTWX = t(X) %*% W %*% X
XTWX_inv <- solve(XTWX)
XTWX_inv_X <- XTWX_inv %*% t(X)
colnames(XTWX_inv_X) <- c("y1", "y2", "y3", "y4", "y5", "y6","y7")
rownames(XTWX_inv_X) <- c("Beta 1", "Beta 2", "Beta 3")
XTWX_inv_X
```

```
                y1         y2         y3         y4         y5         y6   y7
Beta 1  0.5277778 -0.1388889 -0.1388889  0.3888889  0.3888889 -0.2777778 0.25
Beta 2 -0.1388889  0.5277778 -0.1388889  0.3888889 -0.2777778  0.3888889 0.25
Beta 3 -0.1388889 -0.1388889  0.5277778 -0.2777778  0.3888889  0.3888889 0.25
```

g)

We can show that the estimates are unbiased for least square estimates:

$$E(\hat{\beta}_{WLS}) = (X^TWX)^{-1}X^TWE(y)$$

$$E(y) = X\beta$$

$$E(\widehat{\beta_{\text{WLS}}}) = (X^TWX)^{-1}X^TWX\beta = \beta$$

The individual measurements are now heavily weighted more to the individual measurements. This makes sense because the variance of those estimates are smaller than the combined measurements. They are a more accurate observation of the individual time to recoveries for each drug, and should be weighted as such.

2)

The linear model used is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

Where:

$$x_{1i} = 1\{\text{Treatment B}\}, x_{12} = 1\{\text{Treatment C}\}, x_{3i} = 1\{\text{Treatment D}\}$$

a)

$\beta_0$ is the expected response for those undergoing treatment A.

$\beta_1$ is the difference in response between treatment B and A on average.

$\beta_2$ is the difference in response between treatment C and A on average.

$\beta_4$ is the difference in response between treatment D and A on average.

b)

The coefficient of interest is $\beta_1$, so we construct a 95% confidence interval for it:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-p-1} SE(\hat{\beta}_1)$$

Where $\hat{\beta}_1 = -11.5$ and $SE(\hat{\beta}_1) = 3.89$. Then:

$$-11.5 \pm t_{\alpha/2,n-p-1}(3.89)$$

```r
beta_1_hat <- -11.5
se_beta_1_hat <- 3.89
critical_value <- qt(1 - .05/2,200-3-1)
conf_int <- beta_1_hat+c(-1, 1)*critical_value*se_beta_1_hat
conf_int
```

```
[1] -19.171629  -3.828371
```

We can say with 95% confidence that the difference between treatment groups B and A lies within (-19.1716294 and -3.8283706).

c)

We can represent the mean response in treatment group B as $(\hat{\beta}_1 - \hat{\beta}_0)$. The the 95% confidence interval is:

$$(\hat{\beta}_1 - \hat{\beta}_0) \pm t_{\alpha/2,n-p-1} SE(\hat{\beta}_1 - \hat{\beta}_0)$$

```r
beta_0_hat <- 37.5
se_beta_0_hat <- 2.75
conf_int_beta_0 <- beta_0_hat + c(-1, 1)*critical_value*se_beta_0_hat
conf_int_beta_1_beta_0 <- conf_int - conf_int_beta_0
conf_int_beta_1_beta_0
```

```
[1] -51.24824 -46.75176
```

We can say with 95% confidence that the mean response in treatment group B lies between (-51.248241 , -46.751759 ).

3.

Table 1: Analysis of Variance Table

| Source of Variation | Degrees of Freedom (df) | Sum of squares (SS) | Mean square (MS) | F |
|---|---|---|---|---|
| Regression | $p$ | $\mathrm{SS}_{\mathrm{reg}}$ | $\mathrm{SS}_{\mathrm{reg}}/p$ | $F = \frac{\mathrm{SS}_{\mathrm{reg}}/p}{\mathrm{RSS}/(n-p-1)}$ |
| Residual | $n-p-1$ | RSS | $S^2 = \frac{\mathrm{RSS}}{(n-p-1)}$ | |
| Total | $n-1$ | SST = SYY | | |

```r
p = 2
n = 6
df_reg <- p
df_res <- n - p - 1
df_total <- n - 1

y <- c(3,2,4,6,7,1)
res <- c(.5, .25 , -0.5, 0.5, -1, 0.25)
y_bar <- mean(y)

SS_total <- sum((y - y_bar)^2)
RSS <- sum((res)^2)
SS_reg <- SS_total - RSS

MS_reg <- SS_reg/df_reg
MS_res <- RSS/df_res

F_stat <- MS_reg/MS_res
```

The Analysis of Variance table is below (rounded to 3 decimal points):

| Source of Variation | Degrees of Freedom (df) | Sum of squares (SS) | Mean square (MS) | F |
|---|---|---|---|---|
| Regression | 2 | 24.958 | 12.479 | 19.967 |
| Residual | 3 | 1.875 | 0.625 | |
| Total | 5 | 26.833 | | |

b)

$$R^2 = SS_{reg}/SS_{total} \quad R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SS_{total}/(n-1)}$$

```r
R_sq <- SS_reg/SS_total
R_sq_adj <- 1 - (RSS/df_res)/(SS_total/df_total)
c(R_Squared = R_sq, R_Squared_Adj = R_sq_adj)
```

```
 R_Squared R_Squared_Adj
 0.9301242     0.8835404
```