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 β:

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

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 $A(X^T X)(X^T X)^{-1} = I_3$:

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

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

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:

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

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

$$E(\hat{\beta_{\text{WLS}}}) = (X^T W X)^{-1} X^T W X \beta = \beta$$

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)

β_0 is the expected response for those undergoing treatment A.

β_1 is the difference in response between treatment B and A on average.

β_2 is the difference in response between treatment C and A on average.

β_4 is the difference in response between treatment D and A on average.

b)

The coefficient of interest is β_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:

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

```
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:

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

Knowing that $Var(\hat{\beta}) = (X^T X)^{-1}\sigma^2$ we can find $SE(\hat{\beta}_1 + \hat{\beta}_0)$ with $\sqrt{[1,1,0,0]^T Var(\hat{\beta})[1,1,0,0]}$.

```
sigma_hat <- 19.45
X_T_X_inv <- matrix(c(0.02, -0.02, -0.02, -0.02, -0.02, 0.04, 0.02, 0.02,
                      -0.02, 0.02, 0.04, 0.02, -0.02, 0.02, 0.02, 0.04),
                      nrow = 4, ncol = 4)
index_vector <- c(1,1,0,0)
SE_beta_1_beta_0 <- as.numeric(sqrt(sigma_hat^2 * t(index_vector) %*%X_T_X_inv %*% index_v

beta_0_hat <- 37.5
conf_int_beta_0_1 <- beta_0_hat + beta_1_hat + c(-1, 1)*critical_value*SE_beta_1_beta_0

conf_int_beta_0_1
```

```
[1] 20.57534 31.42466
```

We can say with 95% confidence that the mean response in treatment group B lies between (20.5753389 , 31.4246611).

3.

Table 1: Analysis of Variance Table

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

```
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_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
```

4.

We are interested in the model:

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

a)

The added variable plot plots the residuals from:

$$Y = X\beta + e \quad \text{against} \quad X_2 = X\delta + e$$

In other terms:

$$\hat{e}_{Y.X} \text{ by } \hat{e}_{X_2.X}$$

Since $x_1 = 2.2x_2$ the residuals $\hat{e}_{X_2.X}$ will be the zero vector. So the added value plot will have points at varying heights with $x = 0$.

b)

In this case $Y = 3X_1$. So the first residual vector $\hat{e}_{Y.X}$ will be zeroes. Then the only points will be those on the x axis which are $(\hat{e}_{X_2.X}, 0)$. Since the correlation of predictors between the predictors are between 0 and 1, $\hat{e}_{X2.X}$ is guaranteed to not be a zero vector.

3)

a)

There are a few concerns about this model.

There is evidence a linear model does not fit the model well in the Residuals vs Fitted plot as it has a slight upward concave curve.

The normal QQ-plot shows standardized residuals have a longer right tail than what would be expected by the normal distribution.

There is an upward trend in the $\sqrt{\text{std. residual}}$ by Fitted Values plot.

There are also outlier values and bad leverage points worth investigating.

There are nonlinear relationships between predictors that we can see in figure 6.53.

This is not a valid model.

b)

The plot of residuals against fitted values tells us only that an invalid model has been fit to the data, not how the model has been misspecified.

c)

The leverage point rule we use is:

$$h_{ii} > 2(\frac{p+1}{n})$$

In this case with $n = 234, p = 6$ we classify a point as a leverage point when $h_{ii} > 0.0598$.

We are looking for bad leverage points, so both $h_{ii} > 0.0598$ and $|\text{std. res}| > 2$.

To start, there are two points flirting with this break-point at about leverage 0.08 and standardized residuals near , or a bit above, 2. Point 222 is close to the break-point of leverage and a large outlier, it is worth taking a look at. There is also point 223 which is firmly in the bad leverage point category per our definition.

d)

6.37 appears to be a valid model.

The relationship between predictors looks largely linear.

The curve in the residuals vs fitted plot is muted and points look like they randomly deviate around zero.

The normality assumption in residuals is better now, with slight deviation in the right tail.

The scale-location plot does not appear to have a particularly distinct pattern.

I would examine points 57 and 88 as they are bad leverage points before using this model though.

e)

In this case we are testing:

$$H_0 = 1/x_4 = \log(x_6) = 0, \quad H_a = \text{At least one } \neq 0$$

So we have statistic:

$$F = \frac{(RSS(\text{reduced}) - RSS(\text{full}))/(\text{df}_{\text{reduced}} - \text{df}_{\text{full}})}{RSS(\text{full})/(\text{df}_{\text{full}})}$$

In this case $\text{df}_{\text{full}} = n - p - 1 = n - 8$ and $\text{df}_{\text{reduced}} = n - p - k - 1 = n - 7 - 2 - 1 = n - 6$.

Both $RSS(\text{reduced}), RSS(\text{full})$ can be retrieved from the provided output in the problem.

```
RSS_reduced <- .1781^2*228
RSS_full <- .1724^2*226

df_reduced <- 234 - 6
df_full <- 234 - 8

numerator = (RSS_reduced - RSS_full)/(df_reduced - df_full)
denom = RSS_full/df_full

F_stat = numerator/denom
1 - pf(F_stat, df1 = df_reduced - df_full, df2 = df_full)
```

```
[1] 0.0002371326
```

With a p-value of 0.0002 we can strongly reject the null hypothesis that these variables are necessary. We can move forward with the reduced model.

f)

We can add dummy variables to indicate the manufacturer of the vehicle. If we had n manufacturers we would add n-1 dummy variables. This would let us analyze the differences based on the manufacturer of the vehicle.