Homework 4

Tim Vigers

08 March 2019

1. Simulated outcome

a. Data-generating model

$$\hat{Y} = 0.132 + 0.617X_i + \epsilon_i$$
$$i = 1, 2, \ldots, 20$$
$$\epsilon \sim N(0, \sigma^2)$$

b.

Because indicator variables can only be 0 or 1, you can sort of think of them as a one unit change (even though "one unit change in gender" doens't make sense). Since the linear regression is generally expressing the change in outcome for a unit change in a covariate, treating a binary variable as continuous makes sense in the model.

c. ANOVA table

```
# Make table outline, define variables(p = number of covariates)
p <- 1
n <- 20
t <- 1.4727
f <- t^2
mse <- 0.876588
anova <- as.data.frame(matrix(nrow = 2,ncol = 5),
                       row.names = c("Model","Error"))
colnames(anova) <- c("Df","Sum Sq","Mean Sq","F value","Pr(>F)")
# Fill in model
anova["Model",1] <- p
anova["Model",2] <- f*mse*p
anova["Model",3] <- f*mse
anova["Model",4] <- f
anova["Model",5] <- 1-pf(t^2,1,18)
# Fill in residuals
anova["Error",1] <- n - 1 - p
anova["Error",2] <- mse*(n-p-1)
anova["Error",3] <- mse
anova["Error",4] <- NA
anova["Error",5] <- NA
```

```
# Print (round to 3 digits)
kable(anova,digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	1	1.901	1.901	2.169	0.158
Error	18	15.779	0.877	NA	NA

2. Group variable with 4 levels

a. Model 1

X is a full-rank model because the two columns are not linearly dependent on one another. This is because the first column is entirely 1s, and the second column will be a vector of 0s,1s,2s, and 3s, whose order depends on the data.

b. Model 2

Model 2 has an indicator variable for groups 1-3 (so group 0 is the reference group).

$$\begin{pmatrix} 1 & \mathrm{I}_{1\mathrm{group}=1} & \mathrm{I}_{1\mathrm{group}=2} & \mathrm{I}_{1\mathrm{group}=3} \\ 1 & \mathrm{I}_{2\mathrm{group}=1} & \mathrm{I}_{2\mathrm{group}=2} & \mathrm{I}_{2\mathrm{group}=3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{I}_{n\mathrm{group}=1} & \mathrm{I}_{n\mathrm{group}=2} & \mathrm{I}_{n\mathrm{group}=3} \end{pmatrix}$$

If there was an additional column for $\mathrm{I}_{\mathrm{group}=0}$ then the columns of the matrix would be linearly dependent (the first column of all 1s would be the sum of the other 4 columns). However, the X matrix for model 2 is reduced to 4 columns, so it has full rank.

c. Model 3

$X\gamma$ is similar to $X\alpha$ but with an additional column for the indicator variable representing group 0.

$$\begin{pmatrix} 1 & \mathrm{I}_{1\mathrm{group}=0} & \mathrm{I}_{1\mathrm{group}=1} & \mathrm{I}_{1\mathrm{group}=2} & \mathrm{I}_{1\mathrm{group}=3} \\ 1 & \mathrm{I}_{2\mathrm{group}=0} & \mathrm{I}_{2\mathrm{group}=1} & \mathrm{I}_{2\mathrm{group}=2} & \mathrm{I}_{2\mathrm{group}=3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{I}_{n\mathrm{group}=0} & \mathrm{I}_{n\mathrm{group}=1} & \mathrm{I}_{n\mathrm{group}=2} & \mathrm{I}_{n\mathrm{group}=3} \end{pmatrix}$$

Now the first column is the sum of the other columns, and therefore linearly dependent. This means that model 3 is not full rank.

d. Group-level means model

The means model is simply the one-way model without an intercept:

$$E(Y_i|\mathrm{group}_i) = \delta_1 I_{\mathrm{group}_i=0} + \delta_2 I_{\mathrm{group}_i=1} + \delta_3 I_{\mathrm{group}_i=2} + \delta_4 I_{\mathrm{group}_i=3}$$

The coefficients calculated for this model are just the coefficients from the one-way model added together. For example, in the myostatin data intercept = 3.77775, group c = 1.1335, time 24 = 2.2905, and the interaction term = -0.23975. The group means model coefficient for group c at time 24 is 6.962, which is equal to

3.77775+1.1335+2.2905-0.23975. So the models are giving you the same information, but in slightly different ways.

e. Unequally spaced group variable

When looking at a straight line fit of the outcome vs. group, the difference in estimates between 0 and 1 cigarettes per day is not the same as the difference in estimates between 20 and 100 cigarettes per day. Therefore, unless we know exactly how many cigarettes each subject smoked per day, group cannot be treated as a continuous variable. We were able to treat time as a continuous variable in the myostatin data because the timepoints were equally spaced, so the results could be expressed as the change in outcome per 24 hour change in time. With the unequal cigarette groups, the results cannot be expressed this way.