610_HW3

Emma Schmidt

2022-10-18

Question 1

a) The first model is a simple linear regression model that explores the relationship between the DNA marker and cardiometabolic risk score.

$$y_i = \mu + \beta x_i + \epsilon_i$$

$$\mu : \text{is the cardiometabolic risk when the marker value is 0}$$

$$\beta : \text{is the change expected in in cardiametabolic risk per one of the marker}$$

$$\epsilon_i \sim N(0, \sigma^2)$$

b) To gather the estimates below I utilized a bootstrapping strategy: over 1000 iterations I sampled, ran the model, predicted the above and below values, and finally stored their difference. I averaged the results to get a point estimate, and used the quantile function to estimate the intervals.

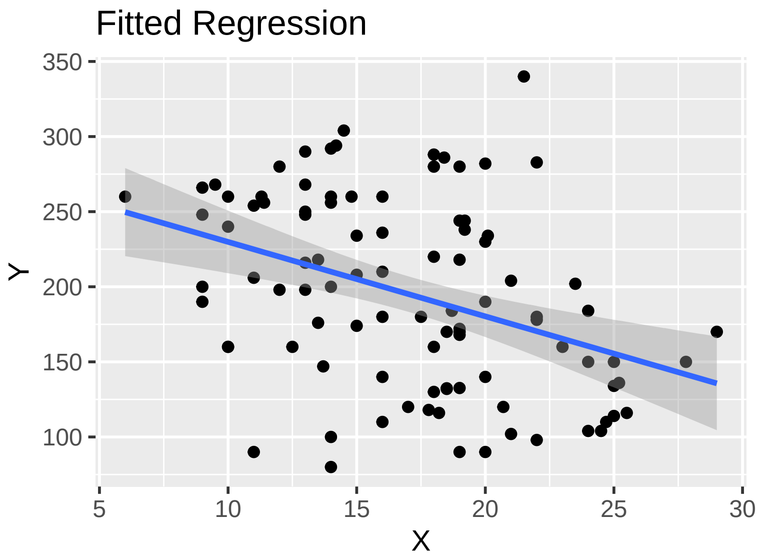
```
## [1] "Point Estimate:"

## [1] -48.24588

## [1] "Confidence Interval:"

##      2.5%     97.5% 
## -66.27921 -28.57587
```

c)



d)

According to the model, higher levels of the marker are associated with lower cardiometabolic risk. The evidence to support this claim comes from the model summary, where the intercept for the marker is -4.957 and is deemed statistically significant. However, a red flag is raised because the adjusted r-squared value is on .1392. This suggests that the marker does not explain very much of the variation in cardiometabolic risk and that this model may not be the right fit for the data.

Question 2

a) The second model is a random intercepts model that now that now takes into consideration hospital level differences in cardiometabolic risk. Specifically, this model includes a fixed effect on marker and random intercept on hospital.

$$y_{ij} = \beta x_{ij}(Marker_{ij}) + \mu_j(Hospital_j) + \epsilon_{ij}$$

β : is the change expected in in cardiametabolic risk per one unit of the marker

μ_j : is the random intercept for each hospital

$$\mu_j = \mu + \alpha_j$$

$$\mu_j \sim N(\mu, \tau^2)$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

b) The model indicates that the between hospital variation is 5729.80 and that the within hospital variation is 34.09. This means that heterogeniety between hospitals is a large contributor to the overall variance of the data, as it is much larger than the within hospital variation.

c) To gather the estimates below I again utilized a bootstrapping strategy: over 1000 iterations I sampled, ran the model, predicted the above and below values, and finally stored their difference. I averaged the results to get a point estimate, and used the quantile function to estimate the intervals.

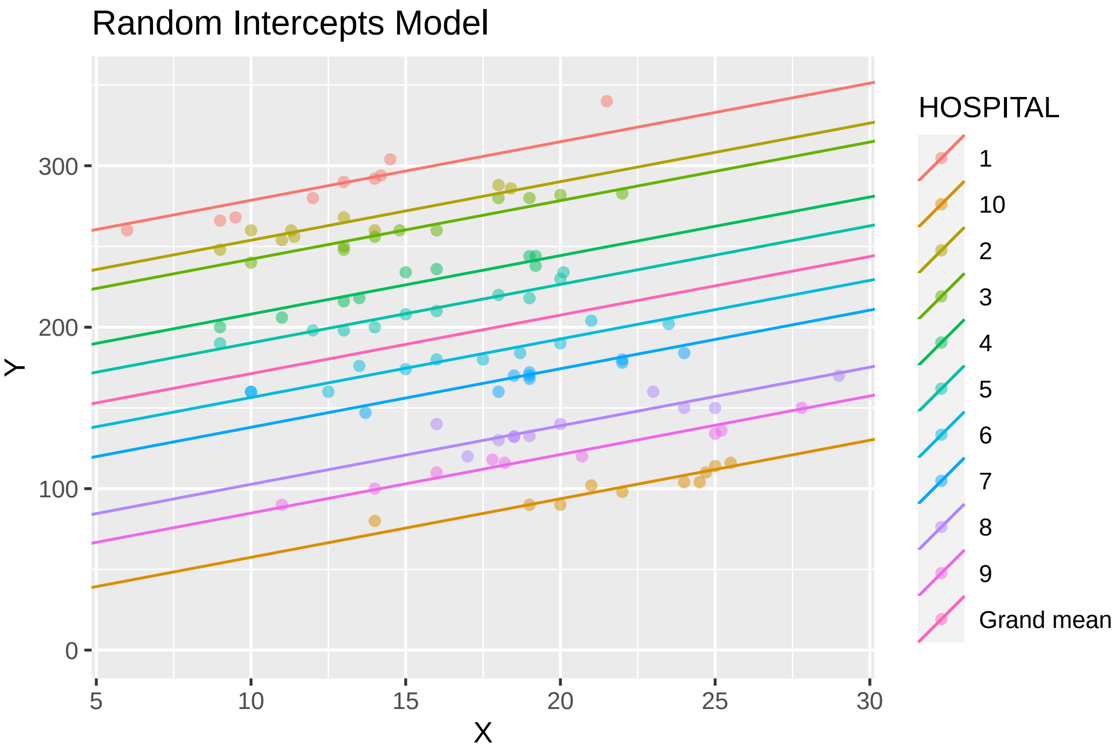
```
## [1] "Point Estimate:"

## [1] 35.34066

## [1] "Confidence Interval:"

##     2.5%    97.5%
## 29.81142 41.06209
```

d)



e) The new model contradicts the original findings that the marker was negatively correlated with cardiometabolic risk. Grouping by hospital reveals that the hospitals exhibit varying levels of cardiometabolic risk, and that the marker actually illustrates a positive correlation with cardiometabolic risk. In other words a higher DNA marker value is expected to increase cardiometabolic risk. Additionally, the severity of cardiometabolic risk varies between hospitals, for example we would expect a patient at hospital 9 to exhibit lower risk than a patient at hospital 1. These relationships can be seen in the plot above.

QUESTION 3

The population average slope that is most suitable to explain the data comes from the second model. In statistical terms, this is because there is a large presence of between hospital heterogeneity in the data that cannot be seen by exploring just the relationship of the marker and risk. In other words, different hospitals see varying degrees of cardiometabolic risk, and the first model does not consider these differences, and thus inaccurately portrays the relationship between the the DNA marker and cardiometabolic risk. The second model accounts for the differences amongst hospitals and paints a clearer picture of the data.

QUESTION 4

a) The third model I employed has a random slope and a random intercept. This model allows us to explore not only the varying levels of cardiometabolic risk between hospitals, but also the allows us to explore the relationship between the DNA marker and risk across hospitals

$$y_{ij} = \mu_j(Hospital_j) + \beta_j x_{ij}(Marker_{ij}) + \epsilon_{ij}$$

μ_j : is the random intercept for each hospital

$$\mu_j = \mu + \alpha_j$$

$$\mu_j \sim N(\mu, \tau^2)$$

β_j : is the random slope for each hospital

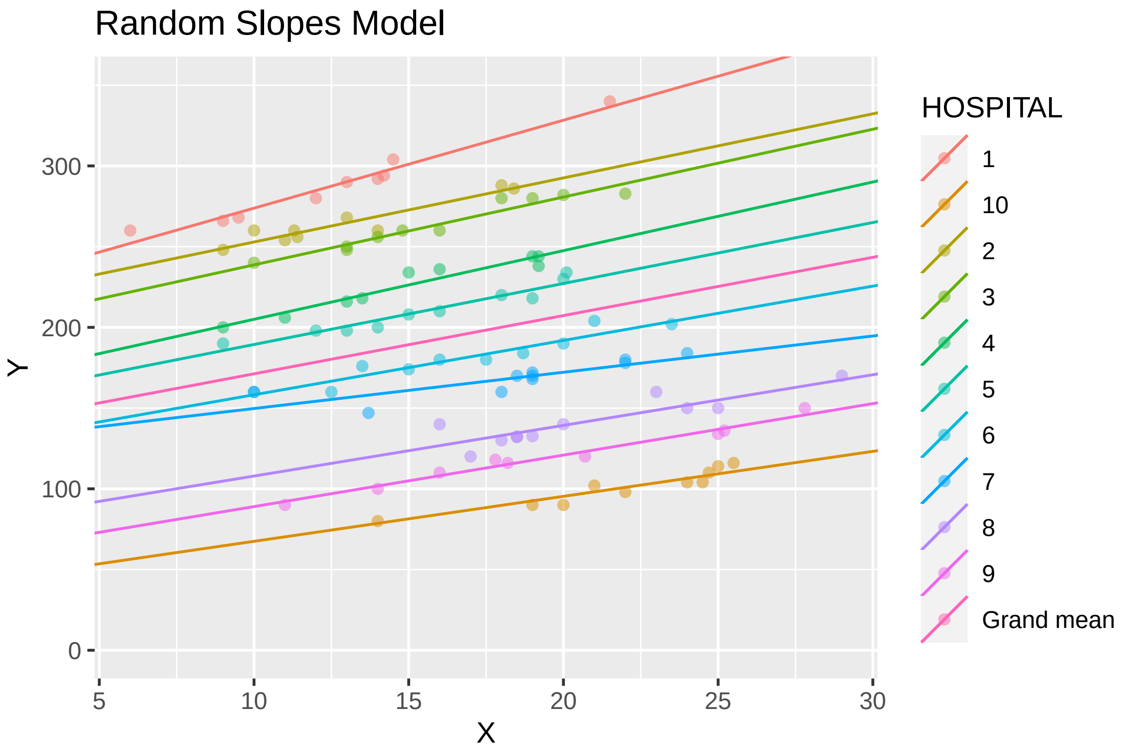
$$\beta_j \sim N(0, \tau_\beta^2)$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

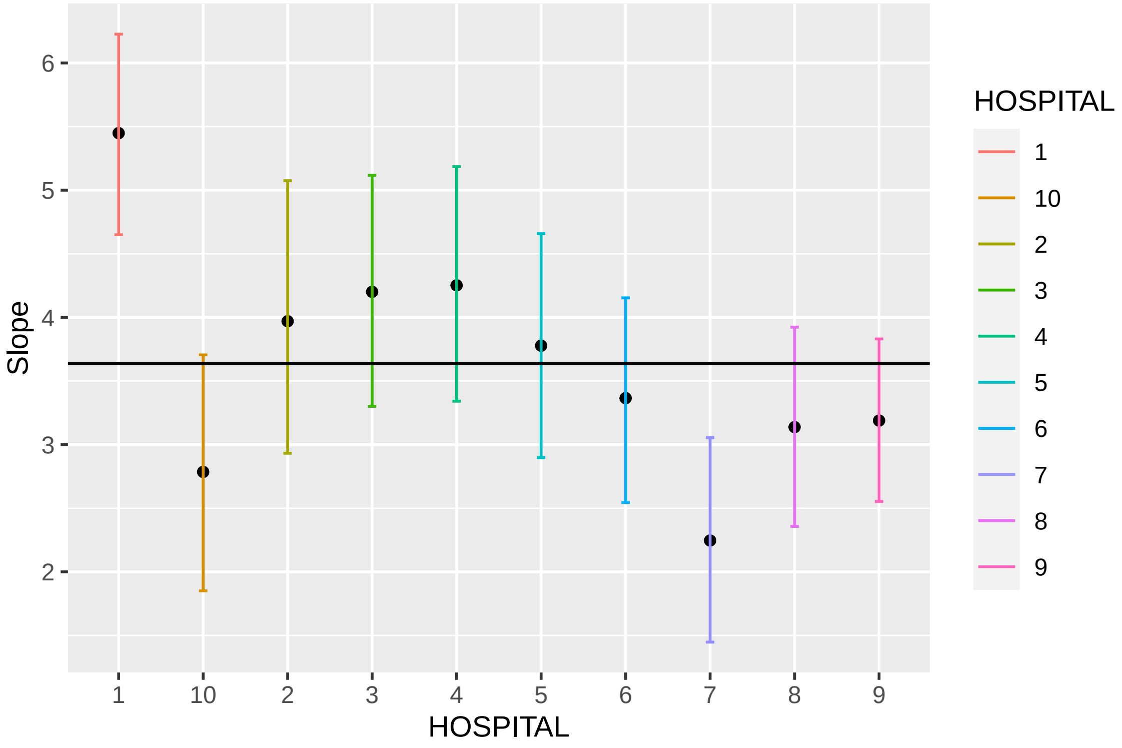
b) The estimates below are calculated by extracting coefficients from the posterior and calculating the respective difference values for each hospital.

Hospital	Point_Est	Lower_2.5	Upper_97.5
1	49.54319	42.28079	56.61610
2	36.09878	26.66439	46.14676
3	38.20010	30.01607	46.52415
4	38.66349	30.38283	47.14966
5	34.35223	26.34891	42.35730
6	30.60150	23.13750	37.77126
7	20.42666	13.16358	27.77571
8	28.53174	21.43251	35.67528
9	28.99829	23.21046	34.83096
10	25.32778	16.82874	33.69380

c)



d) The plot below suggests that there are likely significant differences between the association of DNA marker and cardiometabolic risk across hospitals. While most of the hospitals have overlapping 95% intervals for slope, hospital 1 and 7 stray from the pack, with hospital 1 exhibiting higher association and hospital 7 exhibiting lower association. We can see this significant difference by observing that the intervals for hospital 1 and 7 do not overlap with the intervals of many of the other hospitals. Additionally, the intervals do not contain the population mean slope value.



Question 5

To account for this pollutant, a natural and simple extension to the model in question 4 would be to add a fixed effect for the PM2.5 pollutant. This can be seen in statistical notation below:

$$y_{ij} = \mu_j(Hospital_j) + \beta_j x_{ij}(Marker_{ij}) + \gamma Z_j(PM2.5\ Pollutant_j) + \epsilon_{ij}$$

$$\mu_j \text{ : is the random intercept for each hospital}$$

$$\mu_j = \mu + \alpha_j$$

$$\mu_j \sim N(\mu, \tau^2)$$

$$\beta_j \text{ : is the random slope for each hospital}$$

$$\beta_j \sim N(0, \tau_\beta^2)$$

$$\gamma \text{ : is the change expected in in cardiametabolic risk per one unit of the polluntant}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

Question 6

In question 5 the model was modified to include a fixed effect for PM2.5, which makes sense in the case that PM2.5 behaves the same at each hospital. To test the hypothesis that this is not the case and that the chemical composition of PM2.5 varies from hospital to hospital, the model only needs one small modification. Now instead of a fixed effect, there would be one for each hospital. In other words, the model would now include a random slope for PM2.5 for each hospital. This can be seen below:

$$y_{ij} = \mu_j(Hospital_j) + \beta_j x_{ij}(Marker_{ij}) + \gamma_j Z_j(PM2.5\ Pollutant_j) + \epsilon_{ij}$$

As it currently stands only the average PM2.5 measurement for each hospital exists in the data. In order to adapt the model above the study would need to be modified in such a way that offers multiple measurements of the pollutant at each hospital. Then, much like our earlier exploration, we can use this model to explore the effect of the PM2.5 pollutant between and within each hospital.