

## Exercise, Thursday 8, 11-12 Linear regression

Ram/Ayushi

The dataset provided to you is from an imaginary prospective pregnancy cohort in India. The pregnant women in this cohort were enrolled in the first trimester of their pregnancy and followed up till delivery. At delivery, the birth weight of the baby and gestational age at delivery were documented.

"**Demo\_dataset.dta**" dataset contains the following variables:

1. **id**: This variable represents the unique identifier for each participant in the dataset.
2. **gestational\_age**: The gestational age of the pregnancy, measured from conception to delivery.
3. **birth\_weight**: The weight of the baby at delivery.
4. **low\_birth\_weight**: A binary variable indicating whether the baby's weight was less than 2500 grams (low birth weight).
5. **bmi**: The pre-pregnancy body mass index (BMI) of the mother at enrollment.
6. **bmi\_group**: Categorical variable representing the pre-pregnancy BMI categories at enrollment. It includes the following categories:
  - "normal" (corresponding to a BMI group of 1)
  - "underweight" (corresponding to a BMI group of 2)
  - "overweight" (corresponding to a BMI group of 3)
  - "obese" (corresponding to a BMI group of 4)
7. **preterm\_birth**: A binary variable indicating whether the baby was born before 37 weeks of gestational age (preterm birth).
8. **passive\_smoking**: A binary variable indicating whether the mother was exposed to passive smoking during pregnancy.
9. **biomass\_fuel**: A binary variable indicating whether the household used non-LPG fuel sources for cooking (biomass fuel use).
10. **parity**: A categorical variable indicating the parity of the mother (number of previous pregnancies). It includes the following categories:
  - 0 (nulliparous, meaning no previous pregnancies)
  - 1 (parous, meaning at least one previous pregnancy)
11. **part\_education**: The number of years of education completed by the mother.
12. **family\_inc**: The total family income in Indian Rupees (INR).

### Exercise:

1. Load and describe the "**Demo\_dataset.dta**" dataset into Stata.

```
use "Demo_dataset.dta",clear
// Describe dataset
sum gestational_age ,d
sum birth_weight ,d
tab low_birth_weight
sum bmi ,d
```

```

tab bmi_group
tab preterm_birth
tab passive_smoking
tab biomass_fuel
tab parity
sum part_education ,d
sum family_inc ,d

```

2. Visualize the relationship between gestational age and birth weight using a scatter plot.  
**scatter birth\_weight gestational\_age**

3. Perform a linear regression analysis to investigate the association between gestational age and birth weight.

- Q1. Are these two characteristics associated with each other?
- Q2. What seems to be the influence of gestational age at delivery on birth weight?
- Q3. How do you interpret  $R^2$  in the test report?
- Q4. What does regression analysis inform beyond a correlation analysis for the relationship between two continuous variables?
- Q5. What is residual and how is it different from random error?

```

//correlation
cor birth_weight gestational_age
pwwcorr birth_weight gestational_age, sig

```

```

//linear regression model
glm birth_weight gestational_age, family(gaussian)
*or
regress birth_weight gestational_age

```

```

//Plot residuals
rvpplot gestational_age
*or histogram of residuals
predict resid_ga, r
histogram resid_ga

```

4. Conduct another linear regression analysis to examine the relationship between birth weight and preterm birth

- Q1. How do you interpret the results?
- Q2. Can you relate this with t test?

```

glm birth_weight i.preterm_birth, family(gaussian)
*or
regress birth_weight i.preterm_birth
ttest birth_weight, by( preterm_birth)

```

5. Visualize the relationship between gestational age and pre-pregnancy BMI using a scatter plot. What are your thoughts on the scatterplot that you see? How are these two measures related to each other?

**scatter gestational\_age bmi**

6. Perform a linear regression analysis to assess the association between gestational age and pre-pregnancy BMI.

Q1. Interpret the results.

Q2. Can you interpret the association between these two variables as you did for the previous example (gestational age and birthweight)? If not, why?

Q3. What are the solutions to overcome this problem?

**glm gestational\_age bmi, family(gaussian)**

**\*or**

**regress gestational\_age bmi**

**//Plot residuals**

**rvpplot bmi**

**\*or histogram of residuals**

**predict resid\_bmi, r**

**histogram resid\_bmi**

7. Treat pre-pregnancy BMI as a categorical variable and conduct a linear regression analysis.

Q1. How do you interpret the results?

**glm gestational\_age i.bmi\_group, family(gaussian)**

**\*or**

**regress gestational\_age i.bmi\_group**

8. Additional: Conduct non-linear analysis between gestational age and pre-pregnancy BMI using GAM in R.

**R code**

```
library(ggplot2)
library(mgcv)
model<-gam(gestational_age ~ s(bmi),family = gaussian(), data = dat_for_model)
plot( model)
#or
dat%>%ggplot( aes( x=bmi,y=gestational_age))+
geom_point()+
geom_smooth(method = "gam")+
scale_x_continuous(limits = c(5,40),n.breaks = 10)+
ylab( "Gestational Age")+xlab("BMI")
```



