Lecture 30 Worksheet

Chrysafis Vogiatzis

Every worksheet will work as follows.

1. You will be entered into a Zoom breakout session with other students in the class.
2. Read through the worksheet, discussing any questions with the other participants in your breakout session.
   - You can call me using the "Ask for help" button.
   - Keep in mind that I will be going through all rooms during the session so it might take me a while to get to you.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students in a breakout session is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work. This is no longer valid. All worksheets are now flexibly due 2 days before the corresponding exam.

Worksheet 1: Regression fundamentals

Problem 1: Dependent vs. independent variables

What do we mean by dependent (or response) and independent (or predictor) variables? Provide one example of such a pair.

Answer to Problem 1.

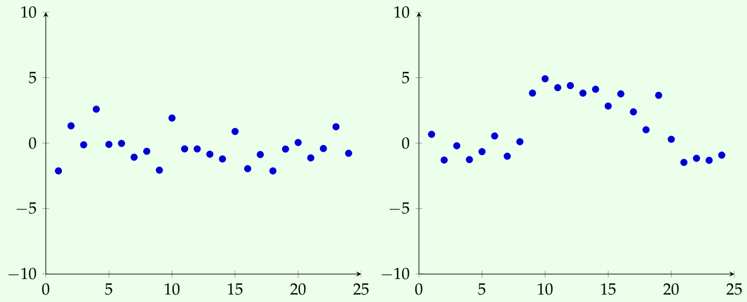
When performing a linear regression, there are some underlying assumptions that we make. They are:

1. **Linearity**: i.e., that the relationship between the independent variable x and the dependent variable y is indeed linear.
2. **Homoscedasticity**: i.e., that the variance of the residuals is the same for any value of x.
3. **Independence**: it implies that the observations (data points) collected are independent from one another.
4. **Normality**: i.e., that the residuals are normally distributed. Equivalently, for any fixed value of the independent variable x, the independent variable y is normally distributed.

Problem 2: Linearity

Recommend a way to check for the linearity assumption. If you were to plot the residuals $y_i - \hat{y}_i$, what should the plot look like in order for the linearity assumption to hold? To help, consider the following two figures plotting the residuals. Which one of the two appears to portray a linear relationship between x and y and which one does not?

Answer to Problem 2.

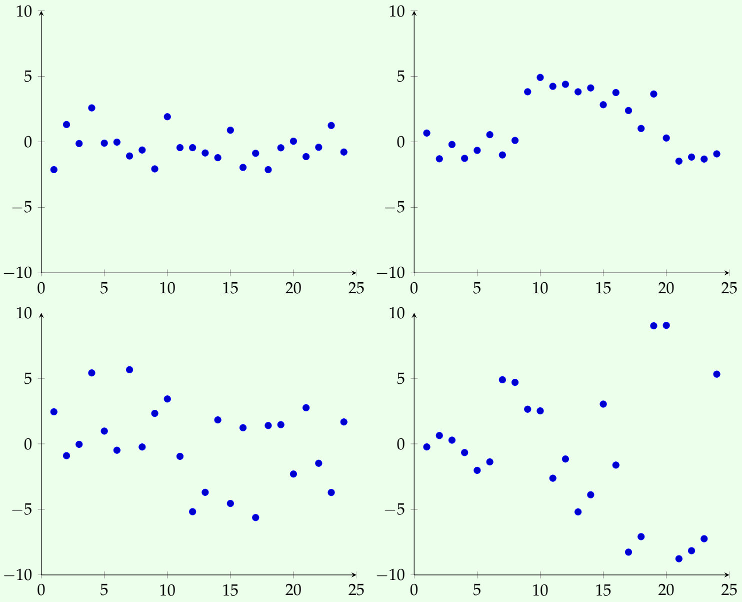


To check the linearity assumption, plotting the residuals should show them be distributed around a horizontal line. A "bow"-like plot signals nonlinearity!

Problem 3: Homoscedasticity

Recommend a way to check for the homoscedasticity assumption. If you were to plot the residuals $y_i - \hat{y}_i$ again, what should the plot look like in order for the homoscedasticity assumption to hold? To help, consider the following two figures plotting the residuals. Which one of the two appears to portray that the variance of the residuals is the same throughout?

Answer to Problem 3.



Problem 4: Regression model vs. regression line?

If our data satisfies the linearity assumption, then we assume there exists a slope β_1 and an intercept β_0 such that: $y = \beta_0 + \beta_1 x$. However, when doing our regression, we write $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Explain in a sentence why we use *observed* or *estimated* values in the regression line, and we do not do that in the original model. What do we have to add to the regression model to be realistic?

Answer to Problem 4.

The model is missing the **noise**! The “true” values would follow:

$$\boxed{y = \beta_0 + \beta_1 x + \epsilon}$$

Worksheet 2: Fluoride levels

Water fluoridation has been a great way to prevent tooth decay (especially in children). The World Health Organization recently published **a report** where the recommended level of fluoride in a community water supply is ranging between 0.5 to 1.5 mg/L. The following data comes from studying children in 10 cities across the United States. Researchers studied the number of cavities per 100 children versus the level of fluoride in the water supply.

City	Fluoride	Cavities
1	1.9	236
2	2.6	246
3	1.8	252
4	1.2	258
5	1.2	281
6	1.2	303
7	1.3	323
8	0.9	343
9	0.6	412
10	0.5	444

Problem 5: Finding the line

Find the least squares regression line using the data of the 10 cities given in the table.

Answer to Problem 5.

Problem 6: Using the line

Use the line you calculated to find the expected number of cavities per 100 children for a city with a fluoride level of 1.5.

Answer to Problem 6.

Problem 7: Calculating residuals

The *residuals* or *errors* are a very important measure in regression. They are typically calculated as

$$y_i - \hat{y}_i$$

with y_i being the observed value for city i and $\hat{y}_i$ being the fitted value had we used the regression line for city i.

Calculate the (10) residuals, one for each city. It is useful to do this calculation in table format: so a table is given to you for convenience!

Answer to Problem 7.

City (i)	Fluoride (x_i)	Cavities (y_i)	Fitted value ($\hat{y}_i$)	Residual ($y_i - \hat{y}_i$)
1	1.9	236		
2	2.6	246		
3	1.8	252		
4	1.2	258		
5	1.2	281		
6	1.2	303		
7	1.3	323		
8	0.9	343		
9	0.6	412		
10	0.5	444		

Worksheet 3: A "qualitative" regression

The Bureau of Labor Statistics has offered the following data on median weekly income per educational level attained.

Level	Income	Level	Income
Less than high school	\$520	Bachelor's degree	\$1173
High school	\$712	Master's degree	\$1401
Some college	\$774	Professional degree	\$1836
Associate degree	\$836	Doctoral degree	\$1743

Problem 8: Translation

Observe how x (our independent variable) is *qualitative*: we need this to become numeric. Think of a way to "translate" the educational level into numbers. Write your idea down.

Answer to Problem 8.

Problem 9: Educational level as level

Let us treat the educational level as $x =$ level. We would then have 8 levels as in: 1 (less than high school), 2 (high school), 3 (some college), 4 (associate degree), 5 (bachelor's degree), 6 (master's degree), 7 (professional degree), 8 (doctoral degree). What is the regression line? What would be the expected salary of a person who quits a year after starting their Master's program (that would be, "some master's education")? [1]

Answer to Problem 9.

[1] You may assume that a Master's degree is typically two years long and takes place after obtaining a Bachelor's degree (so a year of Master's education would lie "between" a Bachelor's and a Master's degree).

Problem 10: Educational level as years of education

Let us treat the educational level as $x =$ years of education. We would then have 9 (less than high school), 12 (high school), 14 (some college), 15 (associate degree), 16 (bachelor's degree), 18 (master's degree), 20 (professional degree), 22 (doctoral degree). What is the regression line? What would be the expected salary of a person who quits a year after starting their Master's program (that would be, "some master's education")?

Answer to Problem 10.