Lecture 31 Worksheet

Chrysafis Vogiatzis

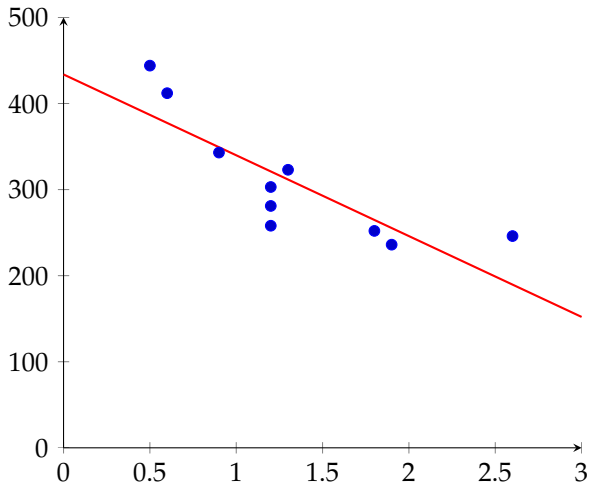
Every worksheet will work as follows.

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 This is no longer valid.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: Fluoride levels and cavities

In the previous worksheet, you were asked to find the regression line for fluoride levels (x) and the resulting cavities (y). After doing all necessary calculations, we came up with the following line:

$$\hat{y} = 433.75 - 93.9 \cdot x.$$



Using the line, we must have also calculated all residuals ($\hat{y}_i - y_i$) as in the following table:

i	x_i	y_i	$\hat{y}_i$	$\hat{y}_i - y_i$
1	1.9	236	255.34	-19.34
2	2.6	246	189.61	56.39
3	1.8	252	264.73	-12.73
4	1.2	258	321.07	-63.07
5	1.2	281	321.07	-40.07
6	1.2	303	321.07	-18.07
7	1.3	323	311.68	11.32
8	0.9	343	349.24	-6.24
9	0.6	412	377.41	34.59
10	0.5	444	386.8	57.2

Problem 1: Calculating the SS_E

As we noted in this lecture, the sum of squares of error SS_E is an immensely useful quantity. Use the values in the last column of the table to calculate the sum of squares of error for the fluoride level regression.

Answer to Problem 1.

Problem 2: Calculating the noise variance

As has become increasingly clear, the noise plays a fundamental role on how well our regression will behave. The variance of the noise can be estimated as the mean square error, which in turn is based on the sum of squares of the error. What is the noise variance in this case?

Answer to Problem 2.

Problem 3: Significant regression?

Combine your answer in Problem 2 (where you got the estimator for the noise variance) with your calculation of $S_{xx} = \sum (x_i - \overline{x})^2$ [1] to decide whether the regression is significant or not using $\alpha = 5\%$.

[1] Remember that x is the fluoride level.

Answer to Problem 3.

How about for $\alpha = 0.2\%$? We do not need to recalculate everything, right? By the way, this can prove interesting for identifying P-values, even using the T distribution critical values!

Worksheet 2: Grades and effort

A professor is interested in whether their exams are unfair: they have come up with a plan to check that. They will ask students to note on their exams (for extra credit perhaps?) how many hours the students spent preparing for the exam. Here are the student answers and the grade they received in the exam.

Student	Hours of study	Grade
1	12	80
2	13	86
3	10	94
4	8	77
5	2	45
6	5	87
7	5	60
8	6	98
9	7	95
10	5	50

In any question that looks like this, our answer is always the same: hypothesis testing. In regression specifically, we will always follow the next steps.

1. Calculate the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
2. Use the regression line to estimate the residuals $y_i - \hat{y}_i$.
3. Calculate the sum of squares of error, SS_E, and the mean square error, MS_E.
4. Calculate S_{xx}.
5. Combine all steps to do a t-test with $n - 2$ degrees of freedom.
   - If we are able to reject, the regression is significant.
   - If we fail to reject, then we lack evidence to claim that the regression is significant.

Let's put this to the test in the data that have been provided to us.

Problem 4: Significant regression?

Using $\alpha = 5\%$, do you have enough evidence that your regression is significant?

Answer to Problem 4.

Problem 5: P-values

When studying the significance of a regression, P-values can help. The smaller the P-value, the more confident we become that the regression is significant. Like in hypothesis testing, if P-value $< \alpha$, then we may reject and claim that the regression is significant. What is the P-value here? How would you go about calculating it? [2]

Answer to Problem 5.

[2] Maybe.. go through the critical values for the T distribution with $n-2$ degrees of freedom and find a value that is *close*?

Worksheet 3: A "qualitative" regression

Continuing (again) from the last worksheet, we saw two ways to quantify the educational level for the purposes of a regression. First, we present the level with integer numbers $1, 2, 3, \ldots, 8$.

Level	Number	Income
Less than high school	1	\$520
High school	2	\$712
Some college	3	\$774
Associate degree	4	\$836
Bachelor's degree	5	\$1173
Master's degree	6	\$1401
Professional degree	7	\$1836
Doctoral degree	8	\$1743

The regression line can be found as

$$\hat{y} = 195.23 + 245.84 \cdot x.$$

Then, we present the level as the "number of years of education" as in $9, 12, 14, \ldots, 22$.

Level	Number	Income
Less than high school	9	\$520
High school	12	\$712
Some college	14	\$774
Associate degree	15	\$836
Bachelor's degree	16	\$1173
Master's degree	18	\$1401
Professional degree	20	\$1836
Doctoral degree	22	\$1743

The line is now becoming:

$$\hat{y} = -630.18 + 111.4 \cdot x.$$

Problem 6: Education level as a numeric level

Is the first regression significant or not?

Answer to Problem 6.

Problem 7: Education level as number of years in education

Is the second regression significant or not?

Answer to Problem 7.

Problem 8: Compare and contrast

Be very careful with this next one. Which one of the two lines appears to provide us with a **more significant** regression? Why? [3]

[3] Would you prefer a bigger or a smaller T statistic?

Answer to Problem 8.

Interesting! We may be able to use this comparison approach soon.