Lecture 32 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 This is no longer valid.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: The ANOVA identity and R^2

In this first set of activities, we will use the ANOVA identity and calculate R^2. As a reminder, the **analysis of variance** identity states that the total variance can be attributed to either the regression or the error (noise):

$$SS_T = SS_R + SS_E.$$

This is used in the calculation of R^2, which can be computer as

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Now, let us return to the fluoride level dataset that we first saw in Worksheet 30.

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_T

Recall that the total sum of squares is needed for calculating the variance. It is defined as $SS_T = \sum(y_i - \overline{y})^2$. What is the total sum of squares in our dataset?

Answer to Problem 1.

Problem 2: Calculating the SS_R

During Worksheet 31, you must have calculated the sum of squares of the error as

$$SS_E = \sum(y_i - \hat{y}_i)^2 = 14261.26.$$

Use this fact as well as the ANOVA identity to calculate the sum of squares of the regression.

Answer to Problem 2.

Problem 3: R^2

What is R^2?

Answer to Problem 3.

Problem 4: An F test for significance

During Lecture 32, we saw that an F test is a valid test when checking for regression significance. That said, when we only have one predictor variable (x), then we saw that a T test can also do the same thing. Are those two equivalent? Do an F test between the mean square of the regression and the mean square of the error and report whether you'd reject (that is, find the regression to be significant) or not.

Answer to Problem 4.

Interesting! We–again–rejected, like we did in Worksheet 31. This is not unexpected. But what is truly fascinating is that we get *exactly* the same P-values! In Worksheet 31, you must have gotten that $T_0 = -4.23$, which translates to a P-value of 0.003. The F test statistic you got here (equal to 17.89) leads to the same P-value of 0.003 also!

Overall: doing a proper F test or a proper T test on a **simple linear regression model** will lead to exactly the same conclusion. This is not true in the case of multiple linear regression, where the F test provides us information about the regression as a whole ("are all predictors insignificant or is there even one significant among them") and the T test provides us information on each individual predictor ("is this specific predictor insignificant or not").

Worksheet 2: Predicting the yield

We have collected the following data on $n = 16$ observations for two factors (x_1, x_2) and their effects on the yield (y) of a crop.

Factor 1 (x_1)	Factor 2 (x_2)	Yield (y)
42	29	251
43	29	251
44	30	248
45	3	268
47	30	273
48	30	277
50	31	270
53	31	285
53	31	290
57	32	297
57	32	303
64	32	305
65	32	309
71	32	322
77	33	331
78	33	349

The line obtained from multiple linear regression is

$$\hat{y} = 157.09 + 2.46 \cdot x_1 - 0.18 \cdot x_2.$$

On your way to find the least squares line, you would have built

$$X = \begin{bmatrix} 1 & 42 & 29 \\ 1 & 43 & 29 \\ 1 & 44 & 30 \\ \vdots & \vdots & \vdots \\ 1 & 78 & 33 \end{bmatrix}.$$

You are also given that

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 1.9298 & -0.0203 & -0.0249 \\ -0.0203 & 0.0006 & -0.0004 \\ -0.0249 & -0.0004 & 0.0016 \end{bmatrix}.$$

Problem 5: Multiple linear regression significance

After some calculations, we found that $\sum(y_i - \hat{y}_i)^2 = 708$ and $\sum(\hat{y}_i - \overline{y})^2 = 12447$. Is the regression significant or not? Use $\alpha = 5\%$.

Answer to Problem 5.

So, we got a significant regression in our hands. But are both factors significant?

Problem 6: More significant

Which of the two factors (x_1 or x_2) is more significant? Why?

Answer to Problem 6.

Worksheet 3: A full case study

Problem 7: Wildfire prediction

Forest fires and wildfires have been a major problem in many parts around the world. Unfortunately, this is an issue that is exacerbated over the last few years. With everything that has happened in 2020, it is very difficult to remember that in the beginning of the year, Australia experienced some of the worst wildfires (see a BBC article here).

In this worksheet, we will try to use **multiple linear regression** to predict the level of severity of a fire depending on the underlying conditions. Among all conditions, we picked the following 4:

- The Duff Moisture Code (DMC).
- The Drought Code (DC).
- The Relative Humidity (RH).
- The wind speed (W).

As our dependent variable, we consider y to be the area burned. Out of a huge dataset, we isolated 10 fires to study.

	DMC	DC	RH	W	Area (in hectares)
1	95	670	26	3.1	64.1
2	83	530	43	4	71.3
3	130	720	21	4.5	88.49
4	150	730	27	3.1	95.18
5	130	700	43	2.7	103.39
6	70	670	36	3.1	105.66
7	120	670	25	3.1	154.88
8	140	600	41	5.8	196.48
9	120	650	46	4.5	200.94
10	150	730	40	4.6	212.88

A person you are working with has let you know that DMC and DC are supposed to give similar indications, so not both of them are necessary. **Between the two multiple linear regression models (one with DMC, RH, W and one with DC, RH, W), which one of the two behaves best and why? Compare them based on their F tests, as well as their R^2 and R^2_{adj} coefficients.**

Before you start solving, here is a small roadmap of what you should do. This will hopefully guide you in all multiple linear regression you try to build.

1. Build matrix X.
2. Calculate matrix $\left(X^T X\right)^{-1}$ and vector $X^T y$.
3. Calculate the coefficients as $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$.
4. Use the line to calculate SS_E.
5. Use the data to calculate SS_T.
6. Use ANOVA to calculate SS_R.
7. Estimate the F_0 test statistic as MS_R / MS_E and reject/fail to reject.
8. Estimate the individual T_0 test statistics for each coefficient and reject/fail to reject.
9. Calculate R^2 and R^2_{adj}.

This may take a while. That said, you have probably noticed by now how useful regression can be, so hopefully you will enjoy comparing these two models and helping us solve a real-life, large-scale societal issue.

Answer to Problem 7.

Answer to Problem 7 (cont'd).