Problem Set 5

Cem Kozanoglu

November 11, 2025

Collaborators: Wooyong Park, Roberto Gonzalez-Tellez, Hanniel Ho and Aileen Wu.

1 Solutions

First we load the data and generate summary statistics for the dataset. The summary statistics can be found in the appendix in the end.

One interesting observation we can glean off this table is that the maximum value for the earnings columns seem to be arbitrary, suggesting that the data was winsorised. As we will see later on, this ends up meaning that many datapoints have very similar leverage values, even though they probably wouldn't have without the winsorisation.

1.1 Problem 1

The regression model for earnings in 1978 on earnings in 1975 is given by:

$$= \text{earnings78} = \beta_0 + \beta_1 \cdot \text{earnings75} + u \tag{1}$$

The regression table for this model is as follows:

	Coefficient	Std. Error
Intercept	5352.7055	101.0181
re75	0.6955	0.0061

And the corresponding formula for leverage will thus be:

$$h_i = \frac{1}{n} + \frac{(\text{earnings75}_i - \bar{\text{earnings75}})^2}{\sum_{j=1}^{n}(\text{earnings75}_j - \bar{\text{earnings75}})^2} \tag{2}$$

Therefore, we would expect that the leverage of an observation will be higher if the earnings in 1975 are further away from the mean earnings in 1975. This will either be values that are exceptionally high, or exceptionally low at 0.

Empirically we observe that the column earnings75 has minimum value of 0, maximum value of 25243 and mean 13650, which means that the observation with earnings75 equal to 0 will have the highest leverage, followed by similarly small values.

The top 5 observations with the highest leverage are:

Table 1: Top 5 observations with highest leverage

treat	age	education	black	hispanic	married	nodegree	re74	re75	re78	leverage
0	45	12	0	0	1	0	25862.32	0.0	3924.842	0.000198
0	49	10	0	0	1	1	391.8534	0.0	0.0	0.000198
0	17	10	0	0	0	1	0.0	0.0	4755.324	0.000198
0	28	16	0	0	1	0	16500.95	0.0	10267.24	0.000198
0	26	17	1	0	1	0	0.0	0.0	4753.846	0.000198

And as predicted, the re75 vs leverage curve has a U-shape:

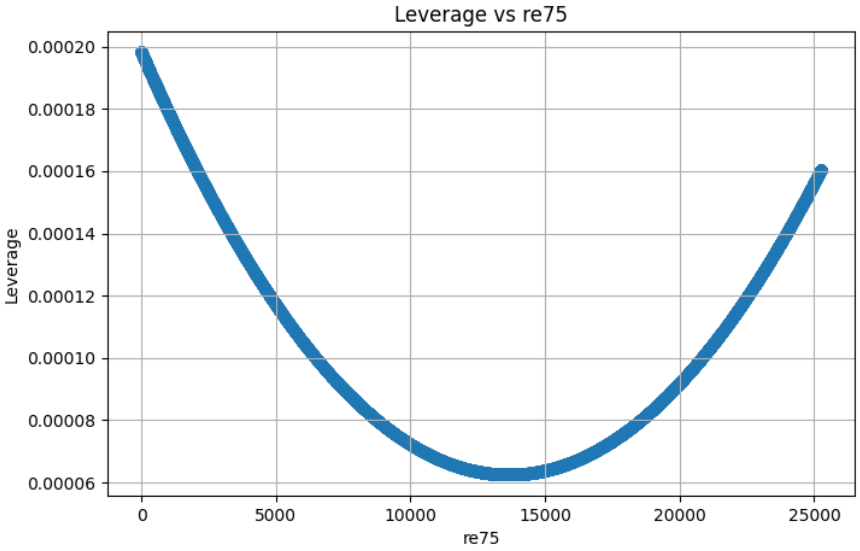

Figure 1: Leverage vs re75

1.2 Problem 2

Our next model also includes covariates earnings in 1974, education, indicators for Black and Hispanic, and age. The regression table for this model is as follows:

	Coefficient	Std. Error
const	6461.5815	321.4952
age	-102.6443	5.5756
education	113.9149	20.2147
black	-826.2451	214.5650
hispanic	-209.0108	219.0637
re74	0.2913	0.0120
re75	0.4702	0.0121

The formula for leverage in this case will be:

$$h_i = \frac{1}{n} + \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \tag{3}$$

Where $\mathbf{x}_i$ is the vector of covariates for observation i, and $\mathbf{X}$ is the matrix of covariates for all observations.

The top 5 observations with the highest leverage in this model are:

Table 2: Top 5 observations with highest leverage (full model)

treat	age	education	black	hispanic	married	nodegree	re74	re75	re78	leverage
0	51	18	0	0	1	0	0.0	25241.76	23052.53	0.002982
0	52	18	0	0	1	0	0.0	25243.55	25564.67	0.002717
0	28	12	0	1	1	0	25862.32	1090.306	21777.26	0.002641
0	38	12	0	1	1	0	0.0	25243.55	25564.67	0.002589
0	29	18	1	0	0	0	0.0	25243.55	25564.67	0.002552

Since we didn't normalise our variables in this case, it makes sense that the observations with the highest leverage are those furthest off the mean when it comes to re74 and re75, which is consistent with our previous model. However peculiarly in this case, the values furthest off the mean are those who had the highest earnings in 1975 and the lowest earnings in 1974, with one exception being the row with high earnings in 1974 and low earnings in 1975. In either case, these people have high leverage because their earnings in both years were far off the mean.

The economic interpretation of this is that earnings between years are likely to be correlated, such that earnings in 1974 are likely to be similar to those in 1975 and those in 1978. Therefore, the observations with the highest leverage are those that had the most extreme earnings in 1974 and 1975, which are likely to be more influential in determining the regression line.

The threshold for 'high' leverage where we should begin to be concerned is given by $3p/N$, where N is the number of observations. In this case, with 15,992 observations and 6 covariates, the threshold for high leverage is approximately 0.00113.

We can see that the top 5 observations with the highest leverage all have leverage values above this threshold, which indicates that they are influential observations that could potentially have a significant impact on the regression results. A total of 968 rows

(6.05% of all rows) have leverage values above this threshold, which is a significant portion of the dataset. This suggests that there are many influential observations in the dataset that could potentially affect the regression results.

1.3 Problem 3

The following are the 5 observations with the highest normalised residuals:

Table 3: Top 5 observations with highest normalized residuals

treat	age	education	black	hispanic	married	nodegree	re74	re75	re78	leverage	std_residuals
0	29	18	0	0	1	0	25862.32	25243.55	0.00	0.00043	3.568122
0	24	12	0	0	1	0	25862.32	25243.55	0.00	0.00032	3.543759
0	25	13	0	0	0	0	25713.42	25243.55	19.21045	0.000285	3.536417
0	29	16	0	0	1	0	25862.32	25243.55	0.00	0.000299	3.535523
0	26	13	0	0	1	0	25862.32	25243.55	0.00	0.000271	3.530685

We can see that the top 5 observations with the highest normalised residuals all have leverage values below the threshold of 0.00113, likely because they aren't actually that far from the mean. However, they have very high normalised residuals because their actual earnings in 1978 are much lower than what the model would predict based on their earnings in 1974 and 1975. This suggests that these observations are outliers in terms of their earnings in 1978 (likely because of an employment shock), and they could potentially have a significant impact on the regression results if they were included in the analysis.

While these observations have high normalised residuals, this doesn't necessarily mean we should remove them from our analysis. In this case, these datapoints correspond to people who were unemployed in 1978, so removing them from our dataset would be discounting a substantial demographic of people who are relevant to our analysis of the effects of the job program.

1.4 Problem 4 & 5

The formula for the Root Mean Squared Error (RMSE) is given by:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4}$$

Where y_i is the actual value of the dependent variable for observation i, and $\hat{y}_i$ is the predicted value from the regression model for observation i.

To calculate the out-of-sample RMSE I use leave-one-out cross-validation (LOOCV), which involves fitting the regression model on all observations except one, and then using that model to predict the value for the left-out observation. This process is repeated for each observation in the dataset, and the RMSE is calculated based on the predictions and actual values. The overall formula for LOOCV RMSE is given by:

$$\text{LOOCV RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_{-i})^2} \tag{5}$$

Where $\hat{y}_{-i}$ is the predicted value for observation i from the model fitted on all observations except i.

In our case we observed the following RMSE values for the simple and full models:

Model	Out of Sample RMSE
Simple	7177.1950
Full	6991.9675

The RMSE for the full model is lower than that of the simple model, which indicates that the full model has a better fit to the data. This is because the full model includes additional covariates that help explain more of the variation in the dependent variable, leading to smaller residuals and thus a lower RMSE. But we can also see that the improvement is fairly marginal, likely because earnings in 1975 already contains most of the important data for predicting earnings in 1978, and the additional covariates do not add much explanatory power.

	statistic	treat	age	education	black	hispanic	married	nodegree	re74	re75	re78
0	count	15992.0000	15992.0000	15992.0000	15992.0000	15992.0000	15992.0000	15992.0000	15992.0000	15992.0000	15992.0000
1	null$_count$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	mean	0.0000	33.2252	12.0275	0.0735	0.0720	0.7117	0.2958	14016.8003	13650.8034	14846.6597
3	std	0.0000	11.0452	2.8708	0.2610	0.2586	0.4530	0.4564	9569.7959	9270.4032	9647.3915
4	min	0.0000	16.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	256	507	758	max	0.0000	55.0000	18.0000	1.0000	1.0000	1.0000	1.0000
25862.3200	25243.5500	25564.6700									

Table 4: Summary Statistics