Ukrainian Catholic University

Faculty of Applied Sciences

Business Analytics & Computer Science Programmes

Cancer incidence analysis in Ukraine

Econometrics final project report

Authors:
Maksym Zhuk
Anton Valihurskyi

May 18, 2025

1 Topic and motivation

Today cancer is one of the hardest diseases to cure. As the World Health Organization estimated [2] states, only 10 types of cancer comprised two-thirds of all death worldwide in 2022, furthermore, incidence rates only accelerate and it is estimated that by 2050 around 35 million of new cases will be diagnosed. Causes of such epidemiological situation can be divided into several categories.

- **Standards of living:** Studies in Japan [3] and USA [6] shows that standard of living measured in **income**, **education** highly affect cancer incidence and mortality.
- **Environment:** Several studies [7], [5] found correlation between water and air quality and cancer incidence and mortality.
- **Medical care:** Direct access to medical facilities is also crucial for cancer prevention and cure. This observation is concluded by study in the USA [1]

Considering how these factors affect incidence rates are crucial for **government** to understand how to allocate resources properly and which policies to make to deccelerate illnes spread.

2 The aim and the tasks

The main **aim** of this project is to find out which and how these factors stated above affect diagnosing cancer on late stages, looking at **objective** to find illness at the start of development

- Clean and merge datasets.
- Fit linear regression model based on air and environment quality, income levels and attributes of medical system (number of laboratories, number of doctors, etc.) to estimate the effect of those parameters on cancer incidence.
- Analyze the results of the modeling. Analyze and interpret the results of the modeling, give explanation to them

The goal of this research is to answer the following questions:

- Does and how medical development of region affect diagnosing on late stages?
- Can GDP explain diagnosing cancer at late stages?
- Do air quality affect this metric?

The cases of diagnosis of multiple tumors are of particular interest in this research, authors believe that this can be better explained by the economic and medical development of the regions, rather than lots of other unobserved factors that may affect baseline cancer incidence.

3 Literature review

A large number of comprehensive works have already been written on this topic. We found that is essential to estimate influence of easiness of access to diagnosing tools[1] and socio-economic [1][6][3]. It is also important to divide population into age groups and race/ethnicity [6]. However, as used data does not include ethnicity as variable we will not test it. Different air pollution elements may vary in how sever they are, it is important to identify those that have the greatest impact and where we can make decisions to address the problem. [7]. We can also consider number of production capacities (e.g., thermal power plants, factories).

4 Data collection

In this analysis we will combine data from several sources: State Statistics Service of Ukraine, Medical Statistics Service of Ukraine

From **Medical Statistics** the following data was obtained.

- Yearly cancer incidence grouped by regions (Example from 2023)
- Yearly visitings to the doctors divided regions (Example from 2023, page **F470100**)
- Number of hospitals where certain types of equipment (MRI, X-Ray) are available in respective regions (Example from 2023, page **F471700**)
- Number of doctors grouped by region (Example from 2023, page **F473210**)

On the other hand from **State Statistics Service of Ukraine** the following data was gathered:

- Yearly data on air pollution grouped by regions and different types of pollutants (Example from 2022: pages **27-31** for air pollution data by regions,
- Yearly data on water pollution grouped by regions and type of pollution (pages **49-50** for water pollution data by regions)
- Region's GRP per capita to capture standards of living (data from 2004-2021)
- Population size and its composition (data from 2004-2025)

After collection and merging the following data was selected: total incidence, multiple tumor cases (cancer at late stage), number of doctor and nursing personnel, numbers of medical equipment and labolatories, air pollution, mean age at region and gpd per capita

In this analysis periods of COVID19 and full-scale invasion are not considered, because those events bring a lot more unobserved effects (such as fear to go to hospital, because of possibility of being infected by COVID or highly increased migration after full-scale invasion). As well Donetsk, Luhansk and Crimea region weren't included into analysis, because of russian invasion into these territories.

5 Data analysis

5.1 Feature description

- nx_ray_pht - number of of X-ray cabinets.
- nflurography_pht - number of fluorography departments
- nct_pht - computer tomography rooms
- nendoscop_pht - endoscopic departments
- ndiaglab_pht - clinical and diagnostic laboratories
- nultrasound_pht - ultrasound diagnostics rooms
- ndoctors_pht - number of doctors
- nnursing_pht - number of nurses
- nill_pht - number of ill patients
- dvisits_pht - number of doctor visits
- air_pollution - emissions of pollutants into the atmosphere air in thousands of tons in the region
- gdp - gross regional product per individual
- mean_age - mean age of people in the region
- mtumors_pti - number of diagnosed multiple tumor cancer cases per thousand of total cancer incidence

Where *pht* means "per hundred thousand people"

5.1.1 Feature Selection

Variables can be divided into 5 following groups

1. Equipment
2. Medical personnel
3. Environmental
4. Socioeconomic

As visible in Fig. 1 equipment, personnel, and illness variables posses multicollinearity among themselves. This may be resolved in the following way. First, the index for availability of equipment - **mdevindex** is created, which is equal to the average of ndiaglab_pht, nx_ray_pht, nultrasound_pht, nendoscop_pht, nultrasound_pht. nct_pht is left aside, firstly, because it is barely correlated with other columns, secondly, because we are particularly interested in the effect of number of CT on mtumors_pti

Also it is visible that **ndoctors_pht** and **nnursing_pht** are highly correlated with **dvisits_pht** and all equipment variables (thus will be highly correlated with *mdevindex*), thus it was decided to drop both personnel variables, but divide **dvisits_pht** and **mdevindex** by **ndoctors_pht**

The resulting correlation table's heat map looks much better

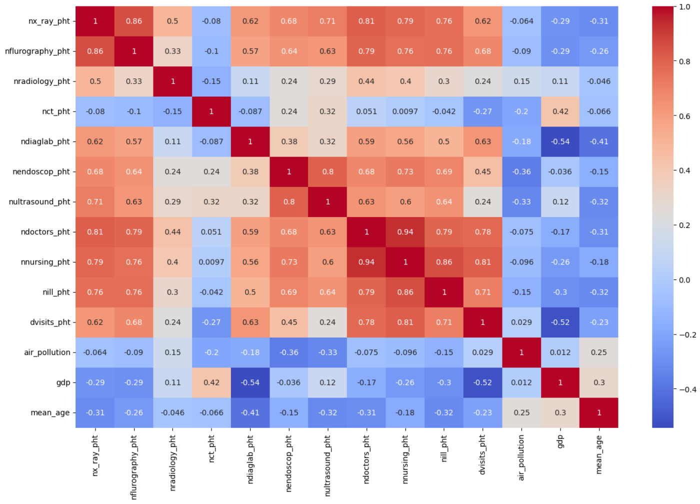


Figure 1: Initial heat map.

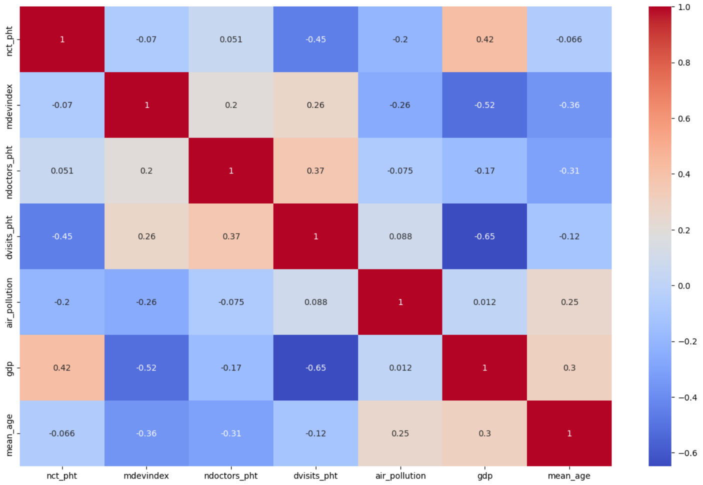


Figure 2: Heat map after EDA.

5.1.2 Demeaning

In the years of observations medical system had undergone lots of reforms, which are really hard to model. To get rid of these effects, we focus analysis only on the differences between regions. Thus for each year of observation mean of respective feature across all regions was subtracted. Plots of some of explanatory variables before and after detrending are shown on Fig. 4.

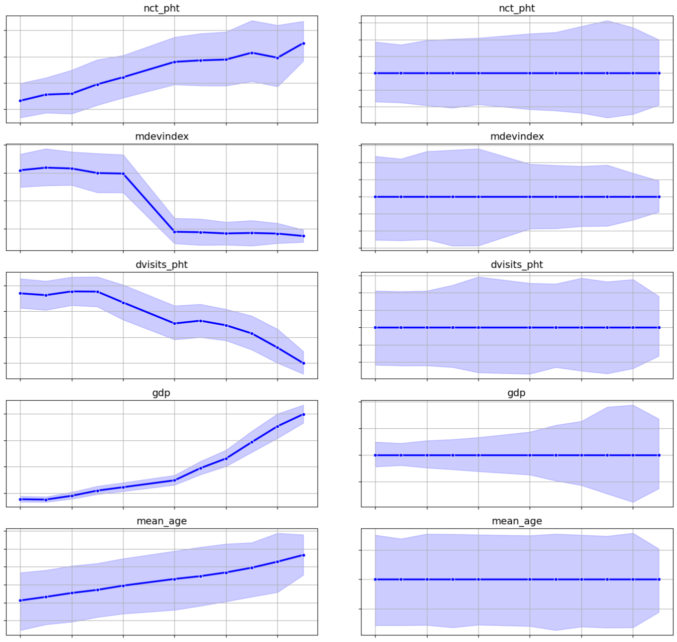
5.2 Regression Analysis

As result the following model (5.2) was used

$$log(mtumors\_pti) = \beta_0 + \beta_1 \times nct\_pht +$$

$$\beta_1 \times mdevindex + \beta_2 \times dvisits\_pht + \beta_3 \times mean\_age$$

The log-level model was used was to increase normality of residuals and decrease effects of heteroskedasticity. Several features such as *gdp* and *air_pollution* turned out to be insignificant given other explanatory variables. This choice not only made model more simple, but made Durbin-Watson



(a) Before demeaning (b) After demeaning

Figure 3

statistics more close to 2 and increased p-value of Jarque-Bera test.

From the model, we note, that on average, holding all the other factors fixed, each year there is 0.2 percent more incidences with multiple cancerous tumors. Also, we see, that with one additional computer tomography (the most precise way of diagnosing oncology) room, we get 0.2 percent less incidences. Coefficient near mdevindex_pht variable, shows us, that in general having 1 additional clinical and diagnostic laboratory, X-ray cabinet, ultrasound room, endoscopic deparment and fluorography room decreases by astonishing 7 percent.

We also note somewhat obvious fact: older people are more susceptible to late stage cancer. We conclude it by looking on coefficient near mean_age variable.

6 Methods

6.1 Regression modeling

The analysis commenced with descriptive statistics to summarize regional variations in predicting variables. Multivariable regression model was employed to quantify associations between environmental, socioeconomic, and healthcare accessibility factors and cancer outcomes. Log-linear regression was employed to estimate annual rates of change and examine associations between environmental, socioeconomic, and healthcare accessibility factors and cancer outcomes. Statistical significance will be evaluated at the 0.05 level, with effect sizes presented as rate ratios and absolute differences.

6.2 Multicollinearity

To avoid multicollinearity, we were mainly guided by correllation table. If the variable was highly correlated with more than one variable it was dropped.

6.3 No serial correlation assumption

We ran Durbin-Watson statistic to make final conclusion concerning autocorrelation.

$$DB = \frac{\sum_{t=2}^{T} e_t - e_{t-1}}{\sum_{t=1}^{T} e_t^2}$$

where e_t is residual at time point t. It can be shown that for large T, the statistic converges to $2(1-\hat{p})$, where $\hat{p}$ is autocorrelation coefficient. Thus, if Durbin-Watson statistic is close to 2, we cannot reject hypothesis that autocorrelation is absent.

6.4 Homoskedasticity assumption

Breusch-Pagan test was used to test heteroskedasticity[4]. The null hypothesis is that variance of error term is independent of exogenous variables.

$$H_0 : Var(u := error\_term|x_1, x_2, \dots, x_k) = \sigma^2$$

Under zero conditional mean this hypothesis translates to

$$H_0 : E(u^2|x_1, x_2, \dots, x_k) = E(u^2) = \sigma^2$$

We want u^2 to be independent of x_i. To test this we fit the model

$$u^2 = \delta_0 + \delta_1 x_1 + \dots + \delta_k x_k + \mu$$

Then F-statistic [4] statistic is applied to test overall significance of the model.

We use heteroskedasticity robust standard errors to be able run t-test, to test significance of the variable (see Wooldridge p. 245). In fact, we need only to estimate variance of the coefficient, so it is asymptotically independent of the observation:

$$\hat{\text{Var}}(\hat{\beta}_j) = \frac{n-k}{n} \frac{\sum_{i=1}^{n} r_{ij}^2 u_i^2}{SSR_j^2}$$

where n - number of observations, k - number of independent variables, r_{ij} - i'th residual from regressing x_j on all the other variables, u_i is i'th residual and SSR_j sum of squared residuals from regressing x_j on all the other variables.

6.5 Normality assumption

Normality assumption also holds by Jarque-Bera normality test. It compares skewness and kurtosis of the residuals, to that of standard normal one(i.e. skewness should be 0, kurtosis should be 3). The null hypothesis is that data follow normal distribution. Jarque-Bera statistic is Chi-squared distributed with 2 degress of freedom.

$$JB = \frac{n}{6}\left(Skewness^2 + \frac{1}{4}(Kurtosis^2 - 3)\right)$$

Dep. Variable:	mtumors_pti	**R-squared:**	0.632
Model:	OLS	**Adj. R-squared:**	0.625
Method:	Least Squares	**F-statistic:**	100.2
Date:	Thu, 24 Apr 2025	**Prob (F-statistic):**	2.15e-58
Time:	19:45:01	**Log-Likelihood:**	171.25
No. Observations:	264	**AIC:**	-330.5
Df Residuals:	258	**BIC:**	-309.0
Df Model:	5		
Covariance Type:	HC1		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	4.0488	0.013	302.942	0.000	4.023	4.075
t	0.0263	0.002	13.902	0.000	0.023	0.030
nct_pht	-0.2014	0.069	-2.904	0.004	-0.337	-0.065
mdevindex	-7.6133	3.201	-2.378	0.017	-13.887	-1.340
dvisits_pht	0.0002	2.98e-05	6.339	0.000	0.000	0.000
mean_age	0.0525	0.004	13.240	0.000	0.045	0.060

Omnibus:	0.120	**Durbin-Watson:**	2.002
Prob(Omnibus):	0.942	**Jarque-Bera (JB):**	0.251
Skew:	-0.003	**Prob(JB):**	0.882
Kurtosis:	2.849	**Cond. No.**	1.28e+05

Notes:

[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 1.28e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Log-level model was used, because, taking log of variable mtumors_pht - number of incidences of diagnosing multiple tumors, makes it more normally distributed

7 Results

Unfortunately, we could not conclude that air quality or gdp of region affects diagnosing cancer on late stages. However, we have successfully concluded, that if we increase the number of medical equipment we can significantly decrease rate of late stage cancer.

Here github repository with all used data and code can be found, main pipeline are present in file *modeling.ipynb*

8 Conclusions

After analysis a strong evidence is present that by increasing medical development of region one can decrease diagnosing cancer at late stages, thus it is strong evidence for the **government** to invest more at medical equipment that can substantially prolong someone's life.

References

[1] . Decreased cancer mortality-to-incidence ratios with increased accessibility of federally qualified health centers. *J Community Health*, .

[2] Global cancer burden growing, amidst mounting need for services. `https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounti`

[3] Chaojing Duan, Siheng Chen, and Jelena Kovačević. Socioeconomic factors and cancer incidence, mortality, and survival in a metropolitan area of japan: A cross-sectional ecological study. ., .

[4] Sarah Lee. A comprehensive guide to interpreting the breusch-pagan test `https://www.numberanalytics.com/blog/a-comprehensive-guide-to-interpreting-breusch-pagan-te` 2025.

[5] R D Morris. Drinking water and cancer. *Environmental Health Perspectives*, .

[6] Gopal K. Singh. Socioeconomic and racial/ethnic disparities in cancer mortality, incidence, and survival in the united states, 1950–2014: Over six decades of changing patterns and widening inequalities. *Environmental Public Health*, 2017.

[7] Anita Nath Thilagavathi Ramamoorthy. Assessing the global impact of ambient air pollution on cancer incidence and mortality: A comprehensive meta-analysis. *JCO Global Oncology*, 2024.