ECON6190 – Project

Predicting Annual Medical Costs

Eren Darici

Fall 2025

Contents

Dataset and Code

The dataset for Part B and all accompanying code files for both parts are publicly available at the following GitHub repository: `https://github.com/eren-darici/ECON6190`

Part A

Question 1

1.i.

Question: Make sure R can read it correctly and report descriptive statistics (including variance) of the two variables in the dataset using R. Provide a scatter plot of the variables using R.

Answer: Table 1. shows the descriptive statistics of the `"data1.csv"`. Table is created using the R Library `stargazer` [1].

Statistic	y	x
Min.	-2.346	-1.261
1st Qu.	-0.833	-0.362
Median	-0.547	-0.067
Mean	-0.337	0.116
3rd Qu.	0.353	0.780
Max.	1.084	1.609
Variance	0.817	0.753

Table 1: Question 1.1 – Descriptive Statistics for data1.csv

Figure 1. shows the scatterplot of x versus y.

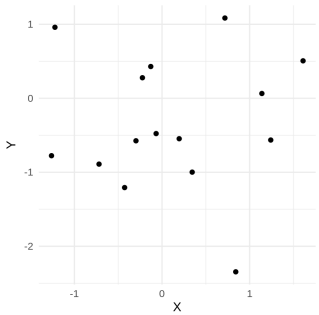


Figure 1: Scatterplot of X vs Y

1.ii.

Question: Run a simple two-sided t-test to check for the equality of the means of the y and the x variables in the data. Use 5% and 1% levels of significance. Comment on the results - whether you reject the null of equal mean or not at 5% level, report test statistic, p-value for the test and 95% confidence interval. How does it differ from the 99% CI? Use R and report the output.

Table 2: Comparison of 95% and 99% Confidence Intervals for Two-Sample t-Test

Statistic	95% CI	99% CI
Lower Bound	-0.2092	-0.4406
Upper Bound	1.1164	1.3477
t-statistic	1.4019	
Degrees of Freedom	27.954	
p-value	0.1719	
Mean of x	0.1163	
Mean of y	-0.3373	

Answer: t-distribution allocates less area in tails ($\alpha/2$) when α is smaller ($\alpha = 0.01$ in 99%), therefore the critical value becomes larger ($t_{0.005,df} > t_{0.025,df}$, which expands the margin of error. Therefore the CI gets wider in 99% when compared to 95%. Moreover, test fails to reject in both significance levels as both CI's contains 0, indicating no evidence in the difference of means.

Question 2

Use seed as the last 4 digits of your student ID number for both (i) and (ii) below. Use R and report the output only as asked for.

Seed used: 8900

2.i.

Question: Generate 10 random numbers from continuous Uniform distribution with range [0,1] and plot the histogram. Compare this to the histogram based on 1000 random numbers generated from the same uniform distribution. Report the sample mean and plot of histogram in each case.

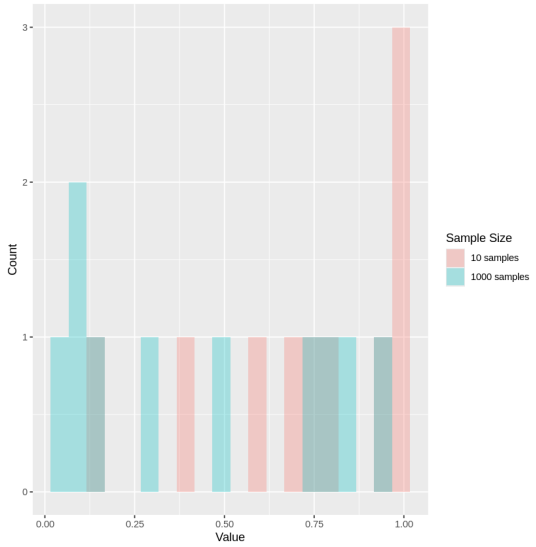


Figure 2: Histogram of Two Generated Samples

Answer: Figure 2. shows the histogram of two generated random samples.

Table 3: Sample Means for Different Sample Sizes

Sample Size	Mean
10	0.7230633
1000	0.4479066

Table 3. shows the means of the generated samples.

2.ii.

Question: Generate 25 random numbers from a standard normal distribution and plot a histogram. Now repeat the process and generate 250 random numbers from the same distribution and plot the histogram. Plot kernel density for these 250 random numbers with (approx.) optimal bandwidth. Increase and decrease the bandwidth sufficiently to illustrate (with a short comment) how the plots changes with bandwidth (make sure to present all plots from R in your output.

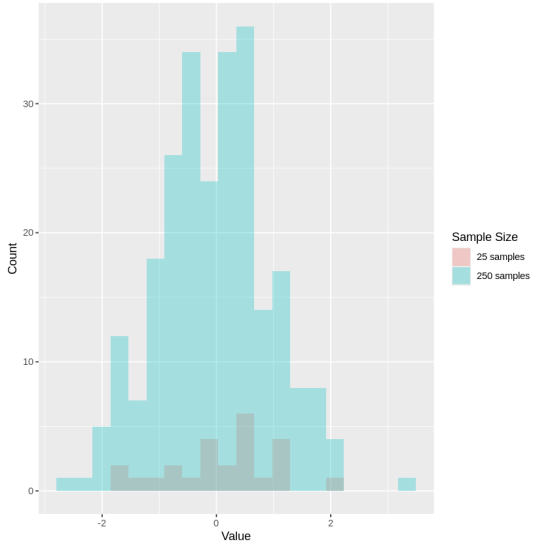


Figure 3: Histogram of Two Generated Samples

Answer: Figure 3. shows the histogram of two generated normal random samples.

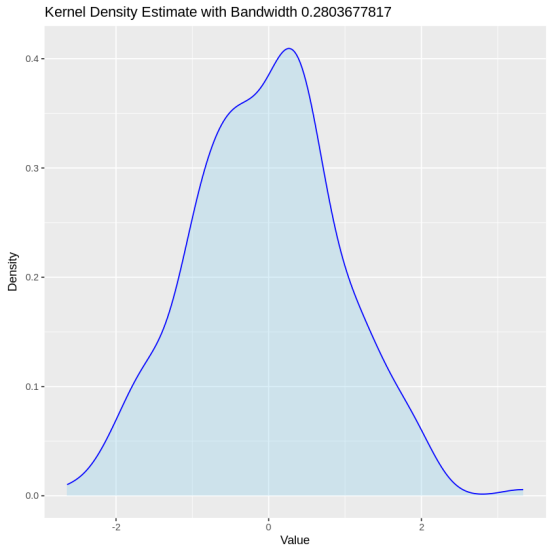
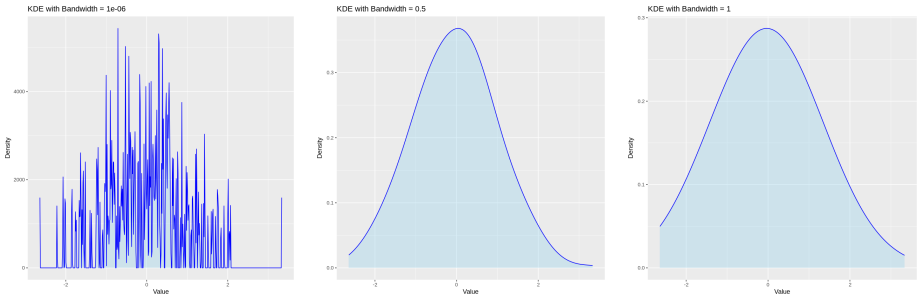


Figure 4: KDE for 250 Samples with Optimal Bandwidth of 0.2803678



Small Bandwidth Optimal Bandwidth Large Bandwidth

Figure 5: KDE Comparisons Across Bandwidths

Figure 4. shows the KDE with optimal bandwidth. Figure 5 shows the KDE's with different bandwidths. As the KDE bandwidth increases above the optimal level, the estimate becomes smoother; as the bandwidth decreases, the estimate becomes more wiggly and captures more noise.

Optimal KDE is calculated using "Silverman's Rule of Thumb" [3].

Question 3

Use the last 4 digits of your student ID as seed value.

Seed used: 8900

3.i.

Question: Generate any 5 numbers. Now resample with replacements from these 3 times. Make each resample size the same as the original sample size. Use R and report the original and the resampled values.

Answer:

Original Generated Numbers = [14, 31, 11, 21, 73]

First Resample = [14, 31, 14, 11, 73]

Second Resample = [21, 21, 73, 21, 21]

Third Resample = [21, 21, 11, 11, 11]

Part B

For this question, the selected dataset is the "Annual Medical Cost" dataset [4]. In this analysis, the **annual medical cost** of each individual is used as the dependent variable (Y). This represents total yearly healthcare expenditure. Six predictors are chosen as independent variables (X), each representing meaningful factors that may influence medical spending:

- **age** – Older individuals may incur higher medical costs due to increased health risks.
- **bmi** – Body Mass Index, an indicator of weight-related health risk.
- **chronic_count** – Number of chronic medical conditions.
- **visits_last_year** – Medical visits in the past year.
- **medication_count** – Number of medications currently taken.
- **hospitalizations_last_3yrs** – Hospitalizations in the past three years.

Question 4.i

Question: Provide descriptive statistics for all variables including their variances. Run OLS regression of Y on all X variables and report the results in a Table. Comment on statistical significance of regression coefficients at 10%, 5% and 1% levels? How is the overall in-sample fit of the model? Justify.

Table 4: Descriptive Statistics of Model Variables

	age	bmi	chronic_count	visits_last_year	medication_count	hospitalizations_last_3yrs	y
Mean	47.522	26.991	0.725	1.928	1.236	0.094	3009.452
SD	15.989	4.995	0.806	1.738	1.209	0.305	3127.463
Variance	255.640	24.949	0.649	3.020	1.463	0.093	9781023.705
Min	0.000	12.000	0.000	0.000	0.000	0.000	55.550
25%	37.000	23.600	0.000	1.000	0.000	0.000	1175.118
Median	48.000	27.000	1.000	2.000	1.000	0.000	2082.575
75%	58.000	30.400	1.000	3.000	2.000	0.000	3707.957
Max	100.000	50.400	6.000	25.000	11.000	3.000	65724.900

Answer: Table 4 presents the descriptive statistics for the dataset. The variable y corresponds to `annual_medical_cost`.

Table 5: OLS Regression Results

	Dependent variable:
	y
age	(0.579298)
	p = 0.000000***
bmi	(1.837097)
	p = 0.000000***
chronic_count	(13.906250)
	p = 0.000000***
visits_last_year	(5.979674)
	p = 0.000000***
medication_count	(8.261830)
	p = 0.141535
hospitalizations_last_3yrs	(30.150560)
	p = 0.000000***
Constant	(58.338020)
	p = 0.00000002***
Observations	100,000
R^2	0.139229
Adjusted R^2	0.139177
Residual Std. Error	2,901.677000 (df = 99993)
F Statistic	2,695.623000*** (df = 6; 99993)
Note:	*p<0.1; **p<0.05; ***p<0.01

Based on the OLS results (Table 5.), five of the six predictors (`age, bmi, chronic_count, visits_last_year, hospitalizations_last_3yrs`) are statistically significant at all (1%, 5%, 10%) significance levels. `medication_counts` is not significant at any of the significance levels.
The overall model is statistically significant, as evidenced by a very large F-statistic (F = 2695.623, $p < 0.01$), indicating that the regressors collectively contribute to explaining variation in the dependent variable. However, the in-sample goodness of fit is relatively weak. The model explains only about 14% of the variation in `annual_medical_costs` ($R^2 \approx 0.139$), which suggests underfitting and implies that important determinants of medical costs may be missing from the model.

Question 4.ii

Question: Generate in-sample predictions and compare it to observed Y. [You may use kernel density plots for a quick comparison. Provide a brief comment].

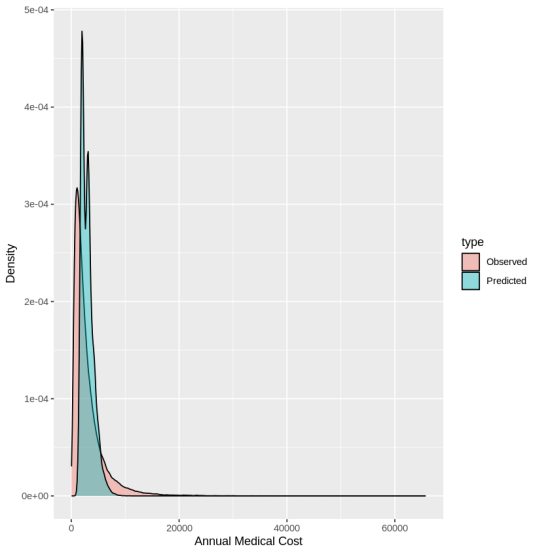


Figure 6: Overlaid KDEs for Observed and Predicted Values

Answer: Figure 6. compares the KDE of the observed and in-sample predicted `annual_medical_cost` values. Predicted values are noticeably concentrated more (peak) on the lower cost values with less variability than the observed distribution. These results are consistent with the low in-sample fit ($R^2 \approx 0.139$). Overall, the OLS model captures broad tendencies but fails to account for extreme values and skewness inherent in the dataset.

Question 4.iii

Question: Run a heteroscedasticity test and report your result with conclusion.

Answer: To assess the heteroscedasticity of the regression errors in the OLS model, a studentized Breusch–Pagan test was conducted. The test produced a statistic of $BP = 1997$ with a corresponding p-value of $p = 2.2 \times 10^{-16}$. Since the p-value is extremely small ($p < 0.01, 0.05, 0.1$), there is sufficient evidence to reject the null hypothesis of homoscedasticity. **This indicates that the error terms exhibit heteroscedasticity.**

Question 4.iv

Question: Run a multicollinearity test and report your result with conclusion.

Table 6: Variance Inflation Factors and Diagnostics

Term	VIF	VIF CI Low	VIF CI High	SE Factor	Tolerance	Tol CI Low	Tol CI High
age	1.0189	1.0134	1.0266	1.0094	0.9815	0.9741	0.9867
bmi	1.0000	1.0000	Inf	1.0000	0.99997	0.0000	1.0000
chronic_count	1.4903	1.4782	1.5027	1.2208	0.6710	0.6655	0.6765
visits_last_year	1.2824	1.2728	1.2922	1.1324	0.7798	0.7738	0.7856
medication_count	1.1857	1.1774	1.1944	1.0889	0.8434	0.8373	0.8494
hospitalizations_last_3yrs	1.0034	1.0005	1.0215	1.0017	0.9967	0.9789	0.9995

Answer: To evaluate multicollinearity among the predictors, Variance Inflation Factor (VIFs) and their diagnostics were computed (Table 6.). All VIF values are below the commonly used thresholds ($VIF > 5$ and$VIF > 10$) [2]. **There is no evidence of problematic multicollinearity among the six predictors.**

Question 4.v

Question: Drop any 2 independent variables of your choice and run an F-test to explain whether you should prefer the restricted (0-type restriction as explained in class) or the unrestricted model. Make sure to report your F-test results.

Answer: After removing **medication_count** (non-significant) and **visits_last_year**, a new "restricted" OLS model was fitted, as presented in Table 7.

Table 7: OLS Regression Results

	Dependent variable:
	y
age	(0.580665)
	p = 0.000000***
bmi	(1.841425)
	p = 0.000000***
chronic_count	(11.544700)
	p = 0.000000***
hospitalizations_last_3yrs	(30.221730)
	p = 0.000000***
Constant	(57.645250)
	p = 0.000000***
Observations	100,000
R^2	0.135142
Adjusted R^2	0.135108
Residual Std. Error	2,908.528000 (df = 99995)
F Statistic	3,906.289000*** (df = 4; 99995)
Note:	*p<0.1; **p<0.05; ***p<0.01

Nested F-test results are presented in Table 8.

Table 8: OLS Regression Results

Statistic	N	Mean	St. Dev.	Min	Max
Res.Df	2	99,994.000000	1.414214	99,993	99,995
RSS	2	843,912,568,573.000000	2,826,300,097.000000	841,914,072,609.000000	845,911,064,537.000000
Df	1	2.000000		2	2
Sum of Sq	1	3,996,991,928.000000		3,996,991,928.000000	3,996,991,928.000000
F	1	237.358700		237.358700	237.358700
Pr(>F)	1	0.000000		0	0

Based on individual R^2 results ($R^2_{full} \approx 0.139$, $R^2_{restricted} \approx 0.135$), and the p-value of the nested F-test (Table 8.), it can be seen that the restricted model performs significantly worse than the full model.

Question 4.vi

To further improve the regression model, the following modifications may be incorporated:

- **Model specification.** Model specification can be refined by considering nonlinear models to capture potential nonlinear relationships between the predictors and the response variable.
- **Variable selection.** Instead of arbitrarily selecting predictors, systematic variable selection methods such as stepwise procedures, LASSO, or information-criterion–based approaches may help identify a more parsimonious and better-performing model.
- **Additional variables.** Including additional predictors may improve accuracy, as health conditions and insurance costs are influenced by a broad range of socioeconomic factors and their interactions.
- **Alternative modeling approaches.** Based on the no-free-lunch theorem [5], different modeling alternatives such as random forests, gradient boosting, or support vector machines can be evaluated to determine whether they offer better predictive performance.
- **Training methodology.** Training methodology can be enhanced through techniques such as cross-validation and hyperparameter tuning (e.g. grid search or Bayesian optimization) to maximize predictive accuracy and model robustness.

Bibliography

[1] Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. R package version 5.2.3. Social Policy Institute. Bratislava, Slovakia, 2022. URL: https://CRAN.R-project.org/package=stargazer.

[2] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R and Python*. 2nd ed. Includes Python edition. New York, NY: Springer, 2023.

[3] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall/CRC, 1986, p. 45. ISBN: 978-0-412-24620-3.

[4] Mohan Krishna Thalla. *Medical Insurance Cost Prediction [Dataset]*. https://www.kaggle.com/datasets/mohankrishnathalla/medical-insurance-cost-prediction. Accessed: 2025-11-26. 2025.

[5] D.H. Wolpert and W.G. Macready. "No free lunch theorems for optimization". In: *IEEE Transactions on Evolutionary Computation* 1.1 (1997), pp. 67–82. DOI: 10.1109/4235.585893.