“Modeling pIC50 in Carbonic Anhydrase II: an exploratory, reproducible study in R, Python, and Excel”

M. Osvaldo Hernández Montoya 13/05/2025

Abstract

This project explores the use of simple linear regression to model the inhibitory activity of Carbonic Anhydrase II (pIC50) from the molecular descriptor AlogP, using curated bioactivity data from ChEMBL. The pipeline included filtering, cleaning, and duplicate resolution, followed by a reproducible implementation of the model in R, Python, and Excel. Results showed consistent parameter estimates across platforms, but an R^2 close to zero and a negative adjusted R^2, indicating that Alog does not explain the variability in pIC50. Statistical diagnostics supported model assuptioms (residuals without systematic patterns, reasonable normality, and approximate homoscedasticity), yet predictive utility is essentially null due to the lack of information in the explanatory variable. Overall, This análisis is a transparent an didactic modeling exercise highlighting the importance of the data curation and critical interpretation of metrics.

Introduction

Carbonic Anhydrase II (CA II) plays a key phisiological role in acid-base balance and carbon dioxide regulation. From a pharmacological perspective, CA II inhibition is relevant for developing compounds with therapeutic potential. This project investigates–at an exploratory level–the relashionship between molecular descriptors and inhibitory activity, measured as pIC50. A simple linear regression was used with AlogP (lipophilicity) as the single explanatory variable to evaluate whether it can account for variability in CA II bioactivity. Beyond numerical results, the primary goal is to demonstrate and end-to-end workflow for data curation, model fitting, diagnostics, and critical interpretation as part of data science portfolio focused on computational biophysicochemistry.

Data Source

Data were obtained form CheMBL (https://www.ebi.ac.uk/chembl/), a widely used public repository in medicinal chemistry and bioinformatics. The target enzyme Carbonic Anhydrase II corresponds to CheBML target ID CHEML205. Experimental

bioactivity records associated with this target were downloaded and curated according to the criterio described below.

Curation and processing

The initial dataset contained 1,215 observations. To improve consistency and statistical validity, records were filtered using the following criteria:

- Standard type: IC50
- Standard units: nM
- Standard relation: "=" (excluding relations such as <, >, ~ to avoid censoring and non-comparable measurements).
- Target organism: *Homo sapiens* (some records may still appear as 'none' despite filtering).

Duplicates, inconsistent values, and extreme records were removed. After curation, the final dataset comprised 543 observations, a reduction of ~55% from the original extract. Such reductions are common for experimental bioactivity data and help ensure a more reliable input for modeling. Notably, acetazolamide–aclassical CA II inhibitor–appears frequently, suporting the biological plausability of the selected records.

Model

To evaluate the relashionship between AlogP and CA II inhibitory activity (pIC50), we fit an Ordinary Least Squares (OLS) simple linear regression model: $pIC50 = \beta_0 + \beta_1 * AlogP + \varepsilon$. The model was implemented reproducibly in three enviroments: R(lm), Python (statsmodels OLS), and Excel (Data Analysis Toolpak). Agreement of parameter estimates cross platforms provides an additional sanity check for correct implementation.

Numerical results

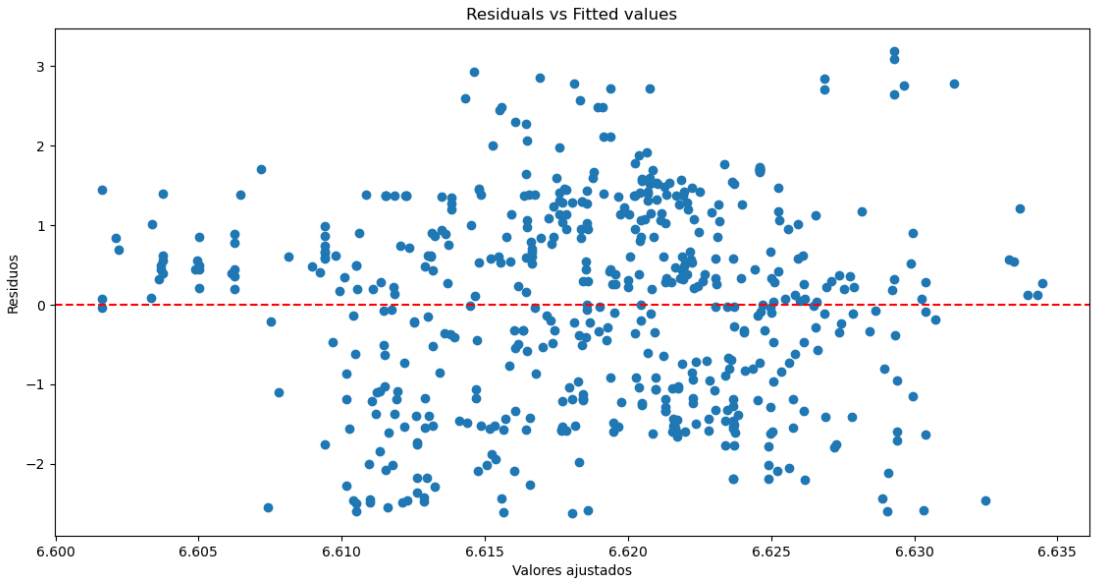
The simple linear regression yielded the following parameter estimates and summary metrics (consistent across R, Python, and Excel):

Parameter / Metric	Estimated value
Intercept	6.61
Slope (AlogP)	0.0032
Residual variance (σ^2, OLS)	1.65
Residual standard error	1.286
R^2	≈ 0.00
Adjusted R^2	< 0

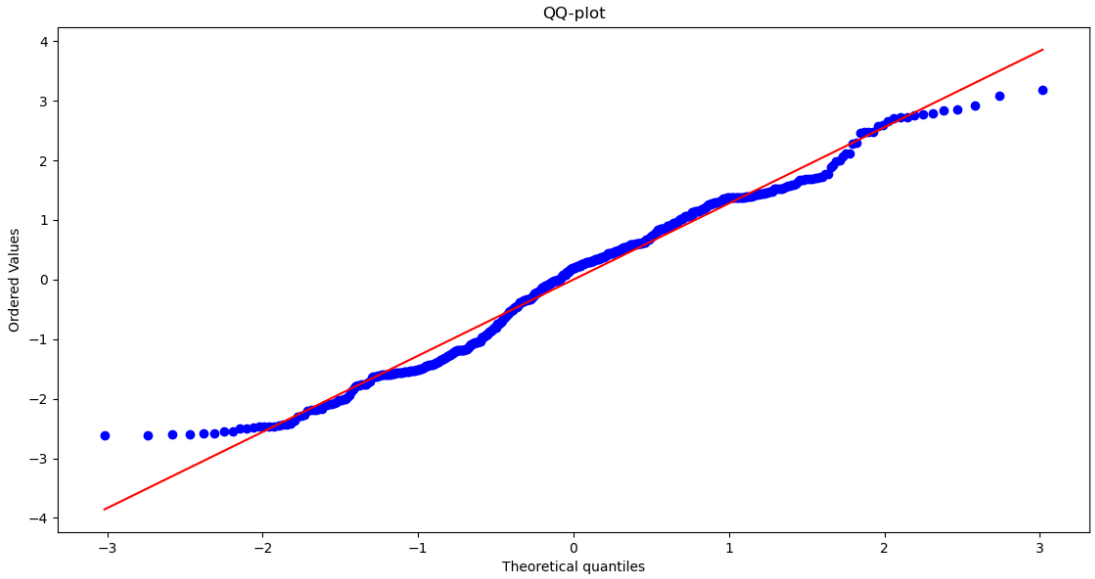
Although the pipeline was correctly implemented and parameters match across platforms, the explanatory power of Alog Pis essentially null. The intercept domanates the fit and the slope is near zero, indicating that the best predictor is almost a constant.

Model diagnostics

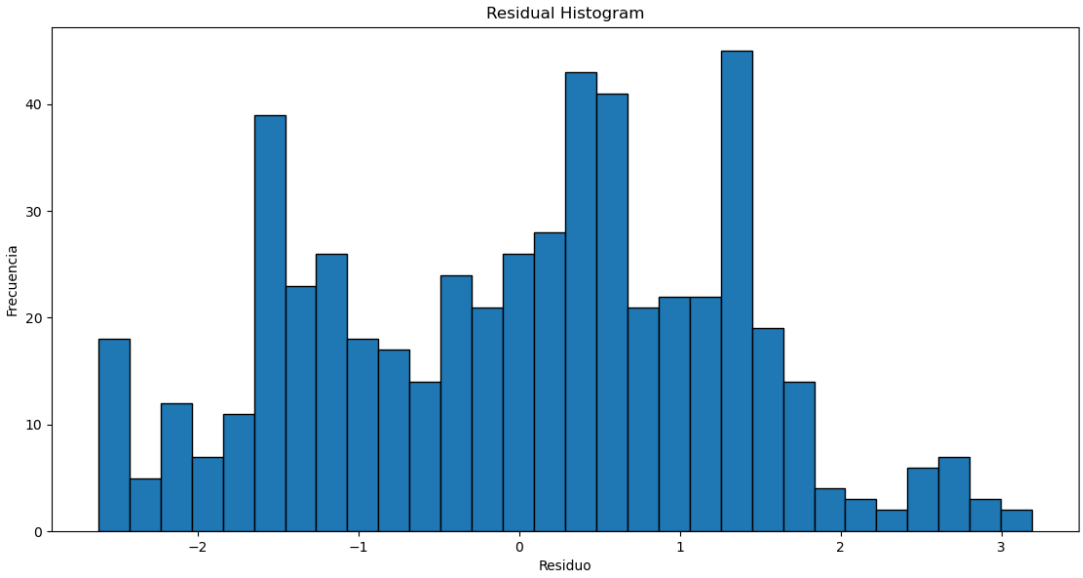
To assess regression assumptions, residual-based diagnostics were reviewed. Overall, diagnostic evidence supported the basic OLS assumptions:



- Residuals scatter randomly around zero without a clear systematic pattern.
- No strong visual indication of heteroscedasticity was observed at this baseline level.



- A Q–Q plot showed approximate aligment with the reference line, with mild deviations at the tails.
- A residual histogram suggested a roughly symmetric distribution around zero.



Limitations

Even though the fitted model is statistically valid in the sense of meeting core assumptions, it has important limitations:

- Low explanatory power: $R^2 \approx 0$ indicates Alog Pdoes not explain pIC50 does not explain pIC50 variability for CA II in this dataset.
- Unreliable predictions: prediction intervals would be very wide, making individual pIC50 estimates based on AlogP alone unhelpful.
- Insufficient descriptors: a single descriptor cannot capture the complexity of pretin-ligand interactions (e.g., polarity, H-bonding, size, shape).
- Experimental noise: bioactivity measurements vary across assay conditions even after curation.

Conclusion

This Project evaluated wheter AlogP alone can explain CA II inhibitory activity (pIC50) using a simple OLS regression model. After extracting and curating ChEBML bioactivity records and validating the workflow across R, Python, and Excel, results showed that AlogP provides virtually no explanatory power ($R^2 \approx 0$; adjusted $R^2 < 0$). In practice, the mean pIC50 is better predictor than the fitted linear model. Beyond predictive performancem the Project demonstrates reproducibility, careful data curation, and honest interpretaion−skills that are essential in applied data science.

Future work

This baseline motivates several natural extesions:

- Multivariate models: include additional descriptors (e.g., H-bond donors/acceptors, polarity, TPSA, rotable bonds).
- Machine learning: explore nonlinear models (Random Forest, SVM, neural networks) and compare against linear baselines.
- External validation: test models on independent dataset sor related Carbonic Anhydrase isoenzymes.

- Integration with docking/dybamics: incorpórate structure-based descriptor sor binding energy surrogates.