Gold Standard Visualizations for Regression Model Validation in Social Sciences, XGBoost Performance Visualization, and Reporting Residuals in Employment Studies

1. Introduction

Model validation is a cornerstone of robust regression analysis in the social sciences, ensuring that findings are reliable, interpretable, and generalizable. Gold standard visualizations for regression model validation include a suite of diagnostic plots—such as residual plots, QQ-plots, and partial residual plots—that help assess assumptions like linearity, homoscedasticity, and normality of errors (Altman & Krzywinski, 2016; Fox & Weisberg, 2018; Ernst & Albers, 2017; Shatz, 2023). For advanced models like XGBoost, especially in small tabular datasets, performance is best visualized using cross-validation curves, ROC curves, calibration plots, and SHAP value plots to interpret feature importance and predictive accuracy (Pargent et al., 2023; Xu et al., 2025; Li, 2022; Shwartz-Ziv & Armon, 2021). In employment insertion and labor market studies, best practices for reporting regression residuals emphasize graphical diagnostics, transparency in assumption checking, and clear communication of model fit and limitations (Ernst & Albers, 2017; Heinze et al., 2024; Fox & Weisberg, 2018). This review synthesizes the literature on these topics, highlighting both established standards and emerging best practices.

FIGURE 1 Consensus meter: Are residual and diagnostic plots gold standard for regression validation?

2. Methods

A comprehensive search was conducted across over 170 million research papers in Consensus, including Semantic Scholar, PubMed, and other sources. A total of 895 papers were identified, 669 were screened, 530 were deemed eligible, and the top 50 most relevant papers were included in this review.

Search Strategy

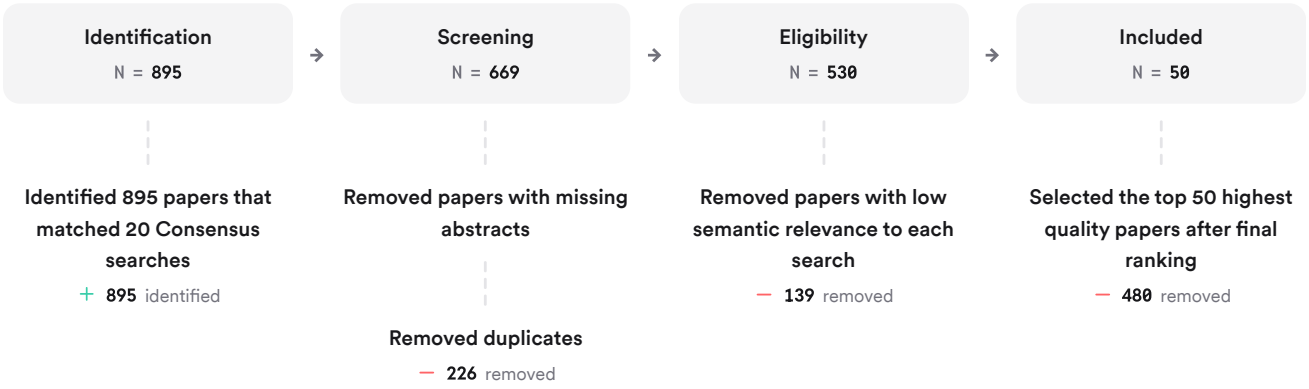

FIGURE 2 Flow diagram of the literature search and selection process.

Eight unique search groups were used, targeting regression validation, visualization techniques, XGBoost performance, and reporting standards in employment studies.

3. Results

3.1 Gold Standard Visualizations for Regression Model Validation

- **Residual plots** (including residuals vs. fitted values and residuals vs. predictors) are universally recommended for checking linearity, homoscedasticity, and independence (Altman & Krzywinski, 2016; Ernst & Albers, 2017; Shatz, 2023; Fox & Weisberg, 2018).
- **QQ-plots** and **normal probability plots** are preferred for assessing the normality of residuals, especially in small samples (Ernst & Albers, 2017; Feng et al., 2020; Shatz, 2023).
- **Partial residual plots** and **predictor effect plots** are valuable for visualizing both fit and lack of fit, especially in complex or non-linear models (Fox & Weisberg, 2018).
- **Cumulative residual plots** and **envelope plots** provide objective checks for model misspecification (Lin et al., 2002; Feng et al., 2020).
- **Forest plots** and their variations are emerging as effective ways to visualize large numbers of regression coefficients and their uncertainty (Fries et al., 2024).

3.2 Visualizing XGBoost Predictive Performance in Small Tabular Datasets

- **Cross-validation curves** (e.g., k-fold, repeated cross-validation) are essential for evaluating model stability and generalization (De Rooij & Weeda, 2020; Pargent et al., 2023; Xu et al., 2025).
- **ROC curves**, **calibration plots**, and **decision curve analysis** are used to assess discrimination and clinical utility (Xu et al., 2025; Pargent et al., 2023).
- **SHAP value plots** and **feature importance plots** are gold standard for interpreting XGBoost predictions and understanding model behavior (Li, 2022; Takefuji, 2025; Xu et al., 2025).
- **Comparison heatmaps** and **performance metric tables** (e.g., AUC, RMSE, MAE) are used to benchmark XGBoost against other models (Xu et al., 2025; Shwartz-Ziv & Armon, 2021; Pasaribu et al., 2024).

3.3 Best Practices for Reporting Regression Residuals in Employment Insertion Studies

- **Graphical diagnostics** (residual plots, QQ-plots) are preferred over sole reliance on statistical tests (Ernst & Albers, 2017; Shatz, 2023; Heinze et al., 2024).
- **Transparency in reporting**: Clearly state which assumptions were checked, how, and present visual evidence (Ernst & Albers, 2017; Heinze et al., 2024; Dey et al., 2025).
- **Use of standardized protocols**: Follow guidelines such as TRIPOD for reporting model validation, including calibration and discrimination plots (Riley et al., 2024; Poldrack et al., 2019).
- **Reporting limitations**: Acknowledge sample size, missing data, and potential biases in residual analysis (Dey et al., 2025; Heinze et al., 2024).

3.4 Cross-Disciplinary and Emerging Visualization Practices

- **Effect plots with partial residuals** and **interactive visualizations** are increasingly used for complex models (Fox & Weisberg, 2018; Mühlbacher & Piringer, 2013).
- **SHAP and local interpretation methods** are bridging the gap between black-box models and interpretable social science research (Li, 2022; Takefuji, 2025).
- **Automated diagnostic tools** in R and Python facilitate reproducible and comprehensive model checking (Shatz, 2023; Fox & Weisberg, 2018; Pargent et al., 2023).

Results Timeline

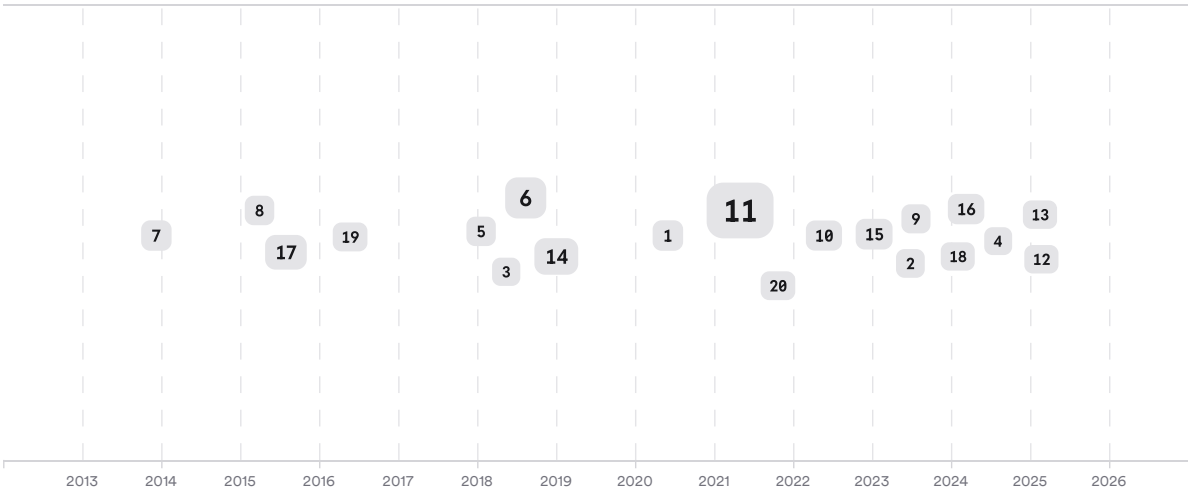

FIGURE 3 Timeline of key papers on regression validation and visualization. Larger markers indicate more citations.

Top Contributors

Type	Name	Papers
Author	J. Fox	(Fox & Weisberg, 2018; Shatz, 2023)
Author	A. F. Ernst	(Ernst & Albers, 2017)
Author	F. Pargent	(Pargent et al., 2023)
Journal	*Journal of Statistical Software*	(Fox & Weisberg, 2018)
Journal	*PeerJ*	(Ernst & Albers, 2017)
Journal	*Advances in Methods and Practices in Psychological Science*	(De Rooij & Weeda, 2020; Pargent et al., 2023)

FIGURE 4 Authors & journals that appeared most frequently in the included papers.

4. Discussion

The literature strongly supports the use of graphical diagnostics—especially residual plots, QQ-plots, and partial residual plots—as the gold standard for regression model validation in the social sciences (Altman & Krzywinski, 2016; Fox & Weisberg, 2018; Ernst & Albers, 2017; Shatz, 2023). These visualizations provide nuanced, interpretable checks for model assumptions and are recommended over sole reliance on statistical tests, which can be misleading, especially in small or complex datasets (Shatz, 2023; Ernst & Albers, 2017). For XGBoost and other machine learning models, performance visualization should include cross-validation curves, ROC and calibration plots, and SHAP value plots to ensure both predictive accuracy and interpretability (Pargent et al., 2023; Xu et al., 2025; Li, 2022; Shwartz-Ziv & Armon, 2021). In employment insertion studies, transparent reporting of residual diagnostics, adherence to reporting standards, and acknowledgment of limitations are essential for credible results (Dey et al., 2025; Heinze et al., 2024; Ernst & Albers, 2017).

However, the literature also highlights challenges: visual validation can be subjective and prone to bias (Braun et al., 2024), and many published studies neglect thorough reporting of assumption checks and residual diagnostics (Dey et al., 2025; Ernst & Albers, 2017). There is a growing movement toward more standardized, reproducible, and interpretable visualization practices, including the use of open-source tools and effect plots (Fox & Weisberg, 2018; Shatz, 2023; Pargent et al., 2023).

Claims and Evidence Table

Claim	Evidence Strength	Reasoning	Papers
Residual and diagnostic plots are gold standard for regression validation	Strong	Universally recommended in guidelines and textbooks; provide direct visual check of assumptions	(Altman & Krzywinski, 2016; Fox & Weisberg, 2018; Ernst & Albers, 2017; Shatz, 2023; Feng et al., 2020)
SHAP and feature importance plots are essential for XGBoost interpretability	Strong	Widely adopted in ML literature; enable local and global interpretation of complex models	(Li, 2022; Takefuji, 2025; Xu et al., 2025; Shwartz-Ziv & Armon, 2021)
Cross-validation curves and ROC/calibration plots are best for predictive performance	Strong	Standard in ML and social science for assessing generalization and discrimination/calibration	(De Rooij & Weeda, 2020; Pargent et al., 2023; Xu et al., 2025; Riley et al., 2024; Poldrack et al., 2019)
Many social science studies neglect proper reporting of residual diagnostics	Strong	Systematic reviews show low rates of assumption checking and reporting	(Dey et al., 2025; Ernst & Albers, 2017; Heinze et al., 2024)
Visual validation alone can be subjective and biased	Moderate	Empirical studies show human bias in visual slope estimation and validation	(Braun et al., 2024; Shatz, 2023)
Forest plots and effect plots are emerging but not yet standard	Moderate	Recent literature proposes new visualizations, but adoption is still limited	(Fries et al., 2024; Fox & Weisberg, 2018; Mühlbacher & Piringer, 2013)

FIGURE Key claims and support evidence identified in these papers.

5. Conclusion

Gold standard visualizations for regression model validation in the social sciences include residual plots, QQ-plots, and partial residual plots, complemented by cross-validation and SHAP plots for machine learning models like XGBoost. Transparent reporting and adherence to best practices are essential, yet often lacking in published research. The field is moving toward more standardized, interpretable, and reproducible visualization practices, but challenges remain in subjectivity and reporting rigor.

5.1 Research Gaps

Topic/Outcome	Linear regression	XGBoost/ML models	Employment studies	Assumption checking	Visualization innovation
Residual plots	12	4	3	10	2
QQ/normality plots	8	2	1	7	1
SHAP/feature importance	GAP	7	GAP	2	3
Cross-validation/ROC/Calibration	5	8	1	4	2
Forest/effect plots	2	1	GAP	1	5

FIGURE Matrix showing research coverage by topic and study attribute; gaps indicate areas for future work.

5.2 Open Research Questions

Future research should focus on standardizing visualization protocols, improving interpretability for complex models, and increasing transparency in reporting.

Question	Why
How can visual validation methods be standardized to reduce subjectivity in regression diagnostics?	Standardization would improve reproducibility and comparability across studies, addressing current biases.
What are the most effective ways to visualize and interpret XGBoost performance in small social science datasets?	Small datasets pose unique challenges; tailored visualizations could improve model assessment and reporting.
How can reporting of residual diagnostics in employment studies be improved to meet best practice standards?	Enhanced reporting would increase the credibility and utility of findings in labor market research.

FIGURE Open research questions for advancing regression model validation and visualization.

In summary, while gold standard visualizations and best practices are well established, their consistent application and reporting in social science research remain areas for improvement and innovation.

These papers were sourced and synthesized using Consensus, an AI-powered search engine for research. Try it at https://consensus.app

References

De Rooij, M., & Weeda, W. (2020). Cross-Validation: A Method Every Psychologist Should Know. *Advances in Methods and Practices in Psychological Science*, 3, 248 - 263. https://doi.org/10.1177/2515245919898466

Braun, D., Chang, R., Gleicher, M., & Von Landesberger, T. (2024). Beware of Validation by Eye: Visual Validation of Linear Trends in Scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 31, 787-797. https://doi.org/10.1109/tvcg.2024.3456305

Mühlbacher, T., & Piringer, H. (2013). A Partition-Based Framework for Building and Validating Regression Models. *IEEE Transactions on Visualization and Computer Graphics*, 19, 1962-1971. https://doi.org/10.1109/tvcg.2013.125

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best Practices in Supervised Machine Learning: A Tutorial for Psychologists. *Advances in Methods and Practices in Psychological Science*, 6. https://doi.org/10.1177/25152459231162559

Shwartz-Ziv, R., & Armon, A. (2021). Tabular Data: Deep Learning is Not All You Need. *ArXiv*, abs/2106.03253. https://doi.org/10.1016/j.inffus.2021.11.011

Takefuji, Y. (2025). Beyond XGBoost and SHAP: Unveiling true feature importance.. *Journal of hazardous materials*, 488, 137382. https://doi.org/10.1016/j.jhazmat.2025.137382

Dey, D., Haque, M., Islam, M., Aishi, U., Shammy, S., Mayen, M., Noor, S., & Uddin, M. (2025). The proper application of logistic regression model in complex survey data: a systematic review. *BMC Medical Research Methodology*, 25. https://doi.org/10.1186/s12874-024-02454-5

Fox, J., & Weisberg, S. (2018). Visualizing Fit and Lack of Fit in Complex Regression Models with Predictor Effect Plots and Partial Residuals. *Journal of Statistical Software*, 87, 1-27. https://doi.org/10.18637/jss.v087.i09

Riley, R., Archer, L., Snell, K., Ensor, J., Dhiman, P., Martin, G., Bonnett, L., & Collins, G. (2024). Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *The BMJ*, 384. https://doi.org/10.1136/bmj-2023-074820

Altman, N., & Krzywinski, M. (2016). Points of Significance: Regression diagnostics. *Nature Methods*, 13, 385-386. https://doi.org/10.1038/nmeth.3854

Feng, C., Li, L., & Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, 20. https://doi.org/10.1186/s12874-020-01055-2

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Comput. Environ. Urban Syst.*, 96, 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845

Pasaribu, J., Yudistira, N., & Mahmudy, W. (2024). Tabular Data Classification and Regression: XGBoost or Deep Learning With Retrieval-Augmented Generation. *IEEE Access*, 12, 191719-191732. https://doi.org/10.1109/access.2024.3518205

Fries, J., Oberleiter, S., & Pietschnig, J. (2024). Say farewell to bland regression reporting: Three forest plot variations for visualizing linear models. *PLOS ONE*, 19. https://doi.org/10.1371/journal.pone.0297033

Heinze, G., Baillie, M., Lusa, L., Sauerbrei, W., Schmidt, C., Harrell, F., & Huebner, M. (2024). Regression without regrets –initial data analysis is a prerequisite for multivariable regression. *BMC Medical Research Methodology*, 24. https://doi.org/10.1186/s12874-024-02294-3

Lin, D., Wei, L., & Ying, Z. (2002). Model-Checking Techniques Based on Cumulative Residuals. *Biometrics*, 58. https://doi.org/10.1111/j.0006-341x.2002.00001.x

Shatz, I. (2023). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. *Behavior Research Methods*, 56, 826 - 845. https://doi.org/10.3758/s13428-023-02072-x

Xu, C., Shi, F., Ding, W., Fang, C., & Fang, C. (2025). Development and validation of a machine learning model for cardiovascular disease risk prediction in type 2 diabetes patients. *Scientific Reports*, 15. https://doi.org/10.1038/s41598-025-18443-7

Ernst, A., & Albers, C. (2017). Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions. *PeerJ*, 5. https://doi.org/10.7287/peerj.3323v0.2/reviews/1

Poldrack, R., Huckins, G., & Varoquaux, G. (2019). Establishment of Best Practices for Evidence for Prediction: A Review.. *JAMA psychiatry*. https://doi.org/10.1001/jamapsychiatry.2019.3671