Determinants of economic growth

FEM11149 - Introduction to Data Science

Maya Archer 623099 (xx%), Nicolas Gonzalez Gort 780037 (xx%)
Rishi Ashok Kumar 560527 (xx%), Saumya Bothra 595402 (xx%)

Introduction

Economic growth, the sustained rise in real Gross Domestic Product (GDP), reflects a country's living standards, tax capacity, and debt sustainability. Governments, lenders, and firms base budgets, borrowing, and trade plans on credible growth expectations. Theory points to several drivers: Solow's growth theory points to technology and capital as important drivers. The Demographic Transition Model (DTM) explains the role declining mortality and fertility rates affecting population structures, which in turn impact economic growth. Lastly, the open-economy theory highlights the importance of openness to trade as a contributing factor. Yet countries differ widely, and policymakers need evidence on which factors actually track growth across economies. This leads to the central research question guiding this analysis: **To what extent can economic and demographic variables, such as trade balance, unemployment, and population characteristics, be used to predict GDP growth?** The results of such an analysis will help clarify the structural factors that shape economic performance and improve forecasting models. It will also help enable policymakers anticipate risks, develop sustainable fiscal strategies, and allocate resources more efficiently, especially those that may be allocated towards ineffective or low impact policies.

Data

The dataset for this analysis is obtained from the *World Bank's Global Jobs Indicators Database* and *Balance of Payments statistics*. The dataset combines 73 macroeconomic, demographic, and labor-market indicators (independent variables) across 150 countries, allowing for the observation of structural patterns that inform policy and strategy. The dependent variable is GDP growth. Most variables are expressed as percentages or ratios to ensure consistency and reduce the need for additional data transformations to bring them to a similar scale. The initial descriptive statistics find that on average, the annual GDP growth rate is 2.6%, but country-level outcomes vary widely, ranging from severe contractions of over –34.3% to strong expansions of more than 13%. The average unemployment rate is 8.1%, with some countries at a low 0.2% while others face levels above 27%. Population growth is generally positive, averaging at 1.4%, but ranges from slight declines (–1.9%) to rapid expansions at 5% per year.

Methods

Regression analysis examines the relationship between a dependent variable and explanatory variables. In economics, it often helps quantify the impact of structural and demographic factors on performance. Ordinary Least Squares (OLS) estimates coefficients by minimizing the squared differences between observed and predicted values, and is expressed as $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$, where y_i is the dependent variable, x_{ij} are the predictors, β_j are coefficients, and ε_i is the error term. OLS assumes linearity of relationships, independence of errors, homoskedasticity (constant variance of errors), and limited multicollinearity. In

practice, economic data often strain OLS assumptions. Variables like unemployment, labor participation, and population growth are highly correlated, inflating standard errors and weakening inference.

Penalized regression methods address multicollinearity by shrinking the coefficents. LASSO (Least Absolute Shrinkage and Selection Operator) adds an L_1 penalty, which stabilizes estimation by setting some coefficients exactly to zero, performing variable selection. Mathematically, LASSO minimizes the sum of squared errors plus a penalty term, expressed as $\min_\beta \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$, where λ controls the strength of the penalty. Larger values of λ produce simpler models with fewer predictors, while smaller values allow more variables to remain in the model. Ridge regression applies an L_2 penalty, minimizing the sum of squared residuals plus $\lambda \sum_j \beta_j^2$, which also shrinks the coefficients towards zero, distributing the effect across many correlated predictors instead of removing them. Elastic Net blends the L_1 and L_2 penalties into one objective, $\min_\beta \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2\right]$, where $\alpha \in [0, 1]$ controls the balance between LASSO and Ridge. This flexibility makes Elastic Net particularly effective when predictors are highly correlated, as it can select groups of variables together while preserving stability.

Model diagnostics help determine predictive power. While the coefficient of determination (R^2) shows how much variation in the dependent variable is explained, a high R^2 does not indicate strong predictive accuracy. Therefore, penalized regressions rely on cross-validation to evaluate generalization. The performance of each model is assessed using a loss function, the mean squared error (MSE) or the mean absolute error (MAE), two values of the penalty parameter λ are often reported, λ_{min}, minimizes the chosen prediction error, and λ_{1se} selects a simpler model where its error is within one standard error of the minimum.

While λ_{min} achieves the lowest estimated prediction error, it can lead to overfitting by tailoring the model too closely to the specific folds of the cross-validation procedure. The "1-SE rule" mitigates this risk by selecting the largest λ within one standard error of the minimum error. This results in a simpler model, trading a small loss in accuracy for improved robustness and generalization. In practice, the stability of cross-validation also depends on how the folds are generated. Setting a random seed (e.g., using `set.seed()` in R) ensures that results are reproducible, since different random splits of the data can otherwise lead to slightly different values of λ_{min} and λ_{1se}. Different seeds can yield different models, especially with correlated predictors, so results should be interpreted carefully and seen as indicators of broader economic patterns, relying on methods like Elastic Net improving stability.

LASSO and Elastic Net are complementary. LASSO selects the most relevant predictors, while Elastic Net adds stability under collinearity. Used together, they balance interpretability and predictive accuracy, yielding models that are both theoretically informative and practically useful.

Results

Table 5 (Appendix A) presents two OLS regression models estimated with a theory-guided predictor set, with and without Net Trade. With an insignificant coefficent of 0.000, Net Trade did not improve the model. The model's fit and explanatory power also worsened with an adjusted R^2 declining from by 0.006, and an unchanged residual standard error (RSE). Significant predictors across the models were life expectancy, number of employers, and GDP per capita.

Table 7 (Appendix A) shows the results of the same model with nonlinear and interaction terms added, changing only marginally. Adjusted R^2 increased by 0.006, improving the explanatory power of the model. The coefficient for GDP per capita^2 is 0.003 significant at the 10 percent level. A 1 unit increase in GDP per capita increases GDP growth by 0.006 percentage points. Population growth^2 and the interaction term between unemployment and the labor force participation were not significant. Since the RSE and overall fit in both models are similar, while nonlinear terms enrich the interpretation of certain variables (especially GDP per capita), they did not drastically improve model performance.

Diagnostic plots show most residuals centered around zero with a few outliers, and heteroskedasticity in the scale-location plot (Appendix B Figure 3). The outliers suggest that some countries may influence the results more than others. The plot highlights the variance of errors changing with fitted values affecting the model's

reliability as it explains some parts of the data more consistently than others. Robust standard errors were examined to ensure no single country dominated results. Findings should therefore be interpreted as general patterns rather than precise point estimates.

As a robustness check, the model was re-estimated with LASSO. At the CV minimum (λ_{min}) it attached a moderate penalty to employment, life expectancy, unemployment, export value, and GDP per capita, shrinking the rest to zero due to their limited predictive power. At the stronger penalty (λ_{1se}), it shrunk toward the intercept (average growth rate) implying that the data does not support many predictors. Compared with OLS, LASSO isolates true drivers and removes noise, yielding a simpler, more stable model with similar accuracy and lower overfitting risk.

Standardizing predictors is essential for LASSO because the penalty acts on coefficient size. Putting variables on a common scale makes the penalty fair, so only informative predictors are retained. Without standardization, large-unit variables would be over penalized and coefficients can shrink to zero. Table 1 shows with standardization only meaningful predictors remain.

LASSO was estimated on a reduced predictor set to illustrate variable selection and the role of standardization. To balance shrinkage and selection, the full feature set and other regularization methods like Ridge and Elastic Net were used, also aiming for greater stability and equal or better predictive accuracy. CV selected $\lambda_{min} = 0.1827$ with a minimized prediction error and $\lambda_{1se} = 2$ for a simpler model. At λ_{min}, LASSO keeps a broad set of predictors ranging from demographic (adolescent fertility, population structure, life expectancy), labor (industry employment, LFP, unemployment), and institutional (export value, tax, credit) variables, suggesting several socioeconomic factors may influence GDP growth when the model is flexible. At λ_{1se}, all coefficients shrink to zero. This contrast shows why relying only on λ_{min} can overfit, while the 1-SE rule trades a small loss in accuracy for a far more robust model. Many human capital and trade factors look relevant under weak penalization, but are not strong enough to survive stricter regularization in this dataset.

Cross-validation for Elastic Net selected a mixing parameter of $\alpha = 0.4$, blending LASSO (selection) and Ridge (shrinkage) with a slight tilt toward the Ridge model. The penalties were $\lambda_{min} = 0.3163$ with a minimized mean squared error and $\lambda_{1se} = 2.2320$ for a more parsimonious model. Compared to LASSO, Elastic Net is more stable under collinearity, shrinking groups of related predictors rather than arbitrarily picking one. In the given data set this is helpful given the correlated labor and demographic variables. In practice, it offers a robust compromise with a similar accuracy to LASSO, better group retention, and improved generalization.

Table 1: Optimal tuning parameters for LASSO and Elastic Net

Method	Metric	Value
LASSO	λ.min	0.1827
LASSO	λ.1se	2.0000
Elastic Net	α	0.4000
Elastic Net	λ.min	0.3163
Elastic Net	λ.1se	2.2320

Table 2: Test Performance (RMSE, MAE, R^2, SD_y)

Model	RMSE	MAE	R2	SD_y
Baseline (mean of train)	6.570	2.925	-0.053	6.486
Lasso (lambda.min)	6.584	3.002	-0.057	6.486
Lasso (lambda.1se)	6.570	2.925	-0.053	6.486
Elastic Net =0.4 (lambda.min)	**6.552**	**3.057**	**-0.047**	**6.486**
Elastic Net =0.4 (lambda.1se)	6.570	2.925	-0.053	6.486

A regular regression removing insignificant variables can be unreliable as it struggles when variables are highly related to each other, or when there are more variables compared to the number of observations and results in estimates jumping around with small changes in data. Penalized models add a complexity penalty, producing stable, parsimonious models that reduce overfitting and focus on signals that generalize. Overfitting can show high accuracy on the training data, but is limited in predictive ability in new data. In these cases, LASSO and Elastic Net deliver interpretable models with similar or better out of sample accuracy, making them a safer choice for policy and business decisions.

To check whether some countries grow faster than others, a binary target, Growing More (1 if GDP growth $> 2.7\%$) was defined, and fit to a Ridge logistic model. Table 3 shows that between the CV choices, the λ_{1se} model outperformed λ_{min} with higher accuracy 0.525 vs 0.475, and lower LogLoss 0.6871 vs 0.8054, and improving Brier 0.2471 vs 0.2774, and AUC 0.563 vs 0.464.

Table 3: Ridge (logistic) — Test Metrics

Model	Accuracy	LogLoss	Brier	AUC
Ridge (lambda.min)	0.475	0.8054	0.2774	0.464
Ridge (lambda.1se)	0.525	0.6871	0.2471	0.563

Table 4 shows a repeated analysis with an alternative 110/40 split. The results find that Run 1 favors λ_{min} on LogLoss/Brier/AUC (0.6553, 0.2321, 0.619), while λ_{1se} is slightly higher on accuracy (0.575 vs 0.525). Run 2 mirrors Table 3, with λ_{1se} best across metrics. The results are consistent with expectations, given the modest sample and a threshold near many cases. On average across the two runs, λ_{1se} delivers better accuracy and calibration (lower LogLoss/Brier), whereas λ_{min} retains a small AUC advantage. Hence, λ_{1se} is preferred for stability and well calibrated probabilities.

If data collection were expanded, a priority would be to collect more countries to reduce variance, followed by more predictors (institutions, investment, trade exposure etc.). A continuous GDP growth variable is preferred when estimating magnitudes for budgeting analyses, whereas the Growing More prediction is preferred when a threshold probability ("will this country exceed 2.7%?") is required.

Table 4: Ridge (logistic) — Test Metrics (different seeds)

	Run 1				Run 2			
Model	Accuracy	LogLoss	Brier	AUC	Accuracy	LogLoss	Brier	AUC
Ridge (lambda.min)	0.525	0.6553	0.2321	0.619	0.475	0.8054	0.2774	0.464
Ridge (lambda.1se)	0.575	0.6821	0.2445	0.500	0.525	0.6871	0.2471	0.563

Conclusion and Discussion

The results show that a small, consistent set of economic and demographic variables explain a meaningful but limited share of cross-country GDP growth. Across OLS and penalized models, higher life expectancy and export value are linked to stronger growth, while unemployment and larger share of employers reduce it. Net trade adds little explanatory power, and penalized regression reaches accuracy comparable to OLS while shrinking most variables to zero. Using the 1-SE rule, the model often picks a very simple set of variables, which means there isn't much consistent signal beyond the core predictors. The ridge model performs best at the 1-SE penalty in answering whether a country exceeds 2.7% growth, yielding better calibration and similar or higher accuracy across data splits. In conclusion, the evidence supports the conclusion that a handful of human-capital, labor-market, openness, and income-level indicators provide the most reliable basis for forecasting, but with large residual variation.

The main limitations are the modest sample size, cross-sectional design, and measurement noise, which make estimates sensitive to outliers and heteroskedasticity. Future research should expand the data to a country-year panel with fixed effects, add better proxies for sectors, investments, R&D, and shocks, and standardize nonlinear terms for comparability. Endogeneity in labor or income variables should also be addressed to improve inference. These steps would strengthen generalization and yield more robust answers to the research question.

Appendix A

Table 5: Comparison of OLS Models: With vs. Without Net Trade

	Dependent variable:	
	GDP growth (annual %)	
	With Net Trade	Without Net Trade
	(1)	(2)
Fertility rate	−0.140	−0.136
	(0.643)	(0.639)
Employers (%)	−0.477***	−0.477***
	(0.123)	(0.122)
Labor force part. (%)	0.010	0.010
	(0.040)	(0.040)
Life expectancy (male)	0.261***	0.262***
	(0.093)	(0.093)
Unemployment (%)	−0.066	−0.066
	(0.064)	(0.064)
Export value index	0.002*	0.002*
	(0.001)	(0.001)
Population growth (%)	0.202	0.196
	(0.479)	(0.474)
GDP per capita (PPP)	−0.0001**	−0.0001**
	(0.00003)	(0.00003)
Net trade (BoP, US$)	0.000	
	(0.000)	
Wage & salaried workers (%)	−0.010	−0.010
	(0.026)	(0.026)
Intercept	−12.514	−12.574
	(8.524)	(8.475)
Observations	150	150
R^2	0.189	0.189
Adjusted R^2	0.131	0.137
Residual Std. Error	4.250 (df = 139)	4.235 (df = 140)
F Statistic	3.242*** (df = 10; 139)	3.627*** (df = 9; 140)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 7: Comparison of LASSO Coefficients With and Without Standardization

	No.Standardization	Standardization
(Intercept)	2.609816	-6.626631

	No.Standardization	Standardization
Fertility rate, total (births per woman)	0.000000	0.000000
Employers, total (% of total employment) (modeled ILO estimate)	0.000000	-0.398982
Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)	0.000000	0.002144
Life expectancy at birth, male (years)	0.000000	0.164192
Unemployment, total (% of total labor force) (modeled ILO estimate)	0.000000	-0.061668
Export value index (2000 = 100)	0.000000	0.001245
Population growth (annual %)	0.000000	0.000000
Population growth (annual %)^2	0.000000	0.000000
GDP per capita, PPP (constant 2011 international $)	0.000000	-0.000045
GDP per capita, PPP (constant 2011 international $)^2	0.000000	0.000000
Net trade in goods and services (BoP, current US$)	0.000000	0.000000
Wage and salaried workers, total (% of total employment) (modeled ILO estimate)	0.000000	-0.000340
Unemp x LFP	0.000000	0.000000

Table 8: Variable selection across LASSO and Elastic Net

Variable	L_min	L_1se	ENet_min	ENet_1se
(Intercept)	T	T	T	T
Adolescent fertility rate (births per 1,000 women ages 15-19)	T	F	T	F
Contributing family workers, female (% of female employment) (modeled ILO estimate)	T	F	T	F
Depth of credit information index (0=low to 8=high)	T	F	T	F
Employers, male (% of male employment) (modeled ILO estimate)	T	F	T	F
Employment in industry, male (% of male employment) (modeled ILO estimate)	T	F	T	F
Export value index (2000 = 100)	T	F	T	F
GDP per capita, PPP (constant 2011 international $)	T	F	T	F
Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate)	T	F	T	F
Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate)	T	F	T	F
Life expectancy at birth, male (years)	T	F	T	F
Own-account workers, total (% of male employment) (modeled ILO estimate)	T	F	T	F
Population ages 15-64, total	T	F	T	F
Population ages 65 and above (% of total)	T	F	T	F
Rural population	T	F	T	F
Tax payments (number)	T	F	T	F
Time required to start a business (days)	T	F	T	F
Time to prepare and pay taxes (hours)	T	F	T	F
Unemployment, male (% of male labor force) (modeled ILO estimate)	T	F	T	F
Unemployment, youth male (% of male labor force ages 15-24) (modeled ILO estimate)	T	F	T	F
Access to electricity (% of population)	F	F	T	F
Contributing family workers, total (% of total employment) (modeled ILO estimate)	F	F	T	F

Variable	L_min	L_1se	ENet_min	ENet_1se
Employers, female (% of female employment) (modeled ILO estimate)	F	F	T	F
Employment in industry (% of total employment) (modeled ILO estimate)	F	F	T	F
Export volume index (2000 = 100)	F	F	T	F
Labor force, total	F	F	T	F
Own-account workers, male (% of male employment) (modeled ILO estimate)	F	F	T	F
Time required to enforce a contract (days)	F	F	T	F

Table 9: Variables Selected under .min vs .1se

Variable	Coef_min	Coef_1se
(Intercept)	-1.7794971	3.001157
Adolescent fertility rate (births per 1,000 women ages 15-19)	-0.0039180	NA
Contributing family workers, female (% of female employment) (modeled ILO estimate)	-0.0067494	NA
Depth of credit information index (0=low to 8=high)	0.0269231	NA
Employers, male (% of male employment) (modeled ILO estimate)	-0.1109801	NA
Employment in industry, male (% of male employment) (modeled ILO estimate)	0.0328390	NA
Export value index (2000 = 100)	0.0013346	NA
GDP per capita, PPP (constant 2011 international $)	-0.0000529	NA
Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate)	-0.0413561	NA
Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate)	0.0076160	NA
Life expectancy at birth, male (years)	0.1067299	NA
Own-account workers, total (% of male employment) (modeled ILO estimate)	0.0079536	NA
Population ages 15-64, total	0.0000000	NA
Population ages 65 and above (% of total)	-0.0107295	NA
Rural population	0.0000000	NA
Tax payments (number)	0.0261170	NA
Time required to start a business (days)	-0.0061819	NA
Time to prepare and pay taxes (hours)	-0.0031672	NA
Unemployment, male (% of male labor force) (modeled ILO estimate)	-0.0353272	NA
Unemployment, youth male (% of male labor force ages 15-24) (modeled ILO estimate)	-0.0287594	NA

Table 6: OLS Regression with Nonlinear and Interaction Effects

	Dependent variable:
	GDP growth (annual %)
Fertility rate	−0.127
	(0.651)
Employers (%)	−0.500***
	(0.123)
Labor force participation (%)	−0.009
	(0.058)
Life expectancy (male)	0.303***
	(0.096)
Unemployment (%)	−0.262
	(0.342)
Export value index	0.002**
	(0.001)
Population growth (%)	0.011
	(0.658)
Population growth $(\%)^2$	0.049
	(0.186)
GDP per capita (PPP)	−0.0002**
	(0.0001)
X GDP per capita $(PPP)^2$	0.003*
	(0.001)
Wage & salaried workers (%)	0.007
	(0.028)
Unemp × LFP	0.003
	(0.006)
Constant	−13.929
	(8.753)
Observations	150
R^2	0.206
Adjusted R^2	0.137
F Statistic	2.967*** (df = 12; 137)
Note:	*p<0.1; **p<0.05; ***p<0.01

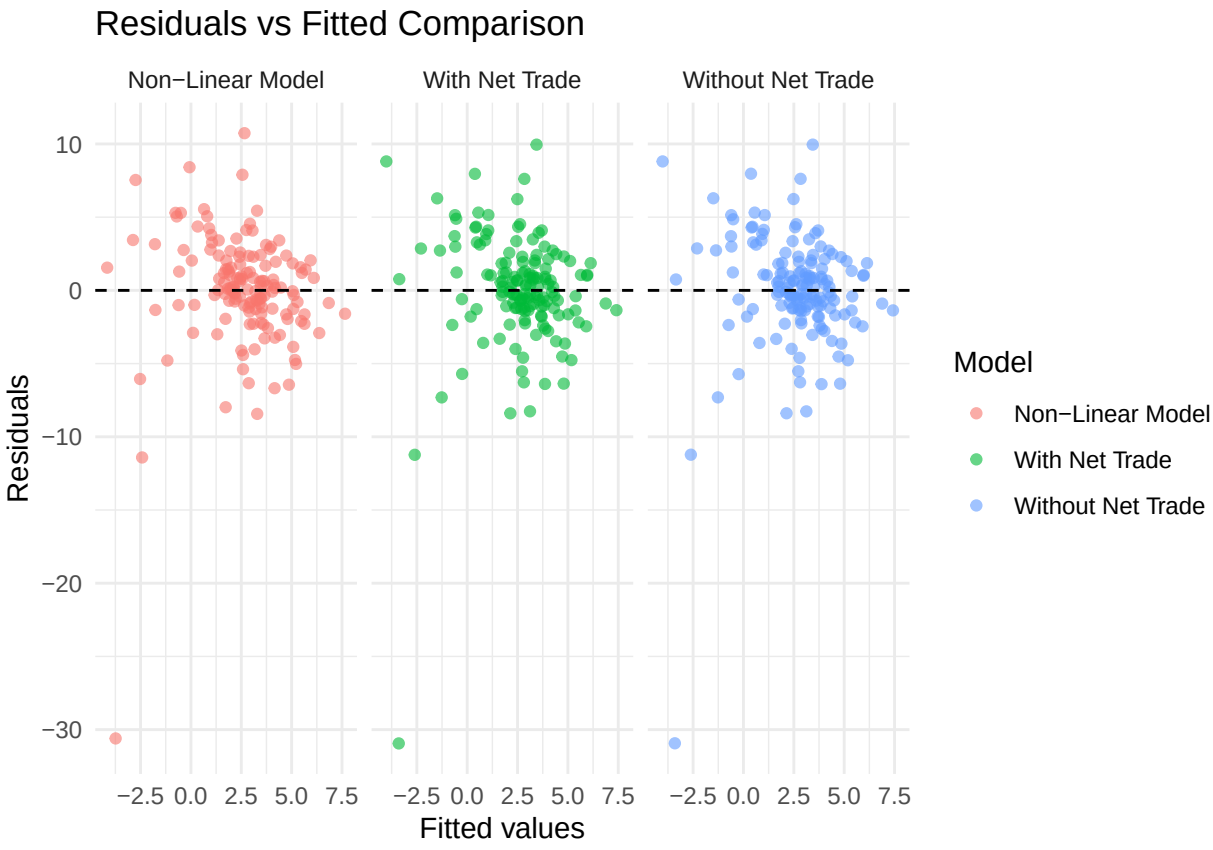

Figure 1: Residuals vs Fitted

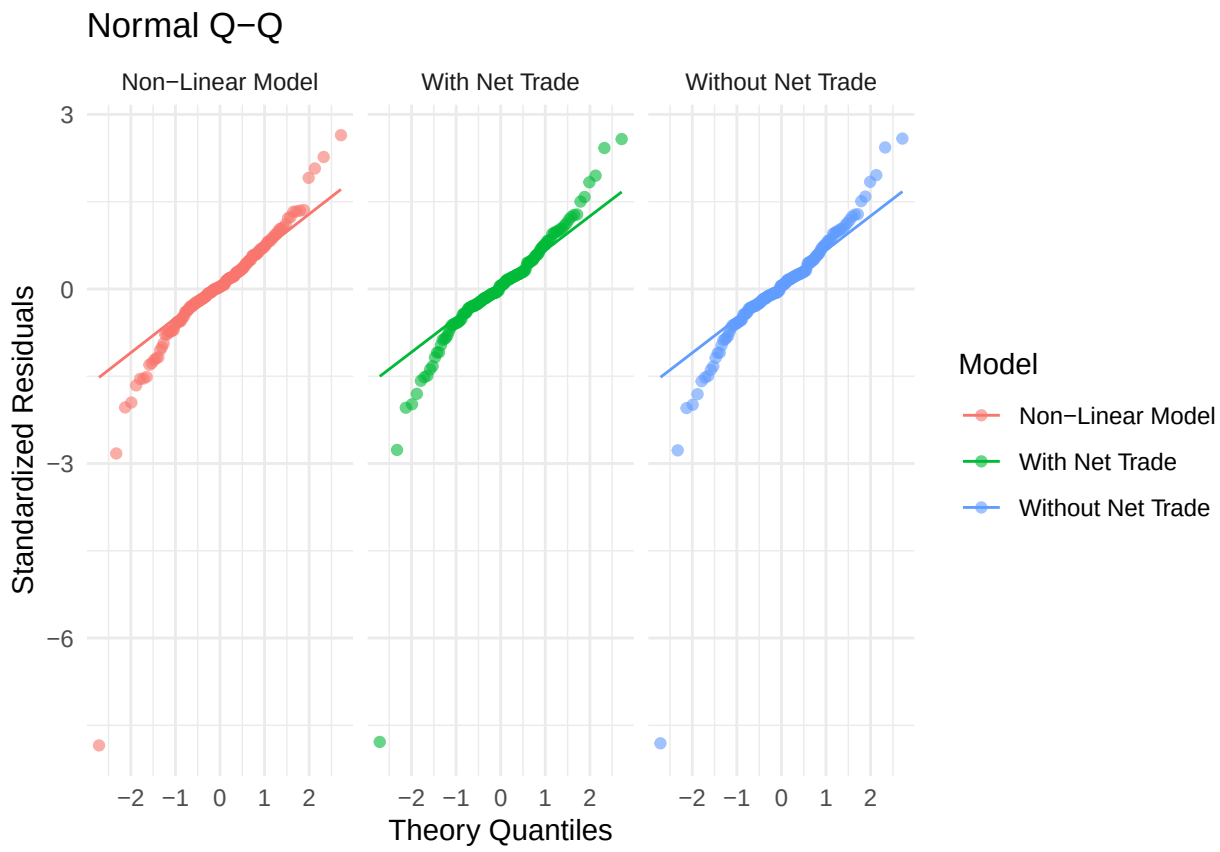

Figure 2: Normal QQ

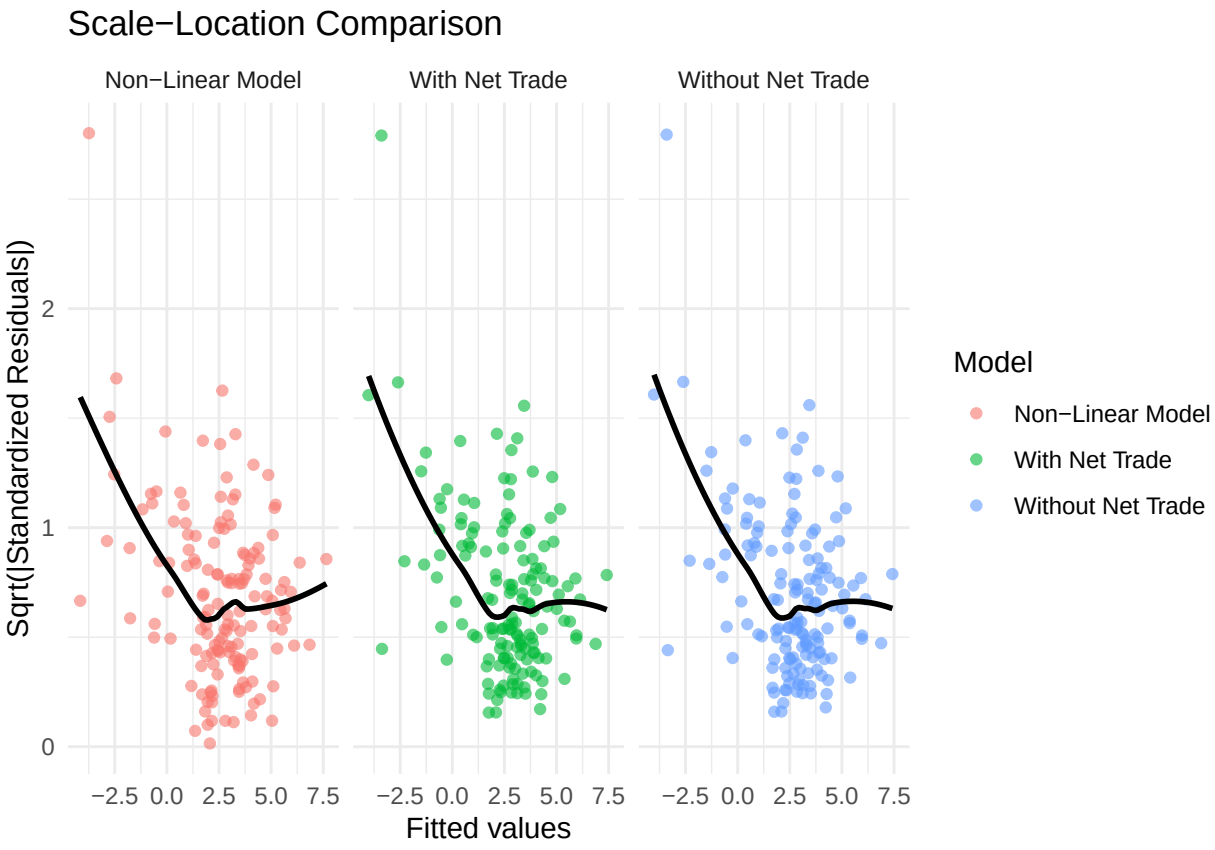

Figure 3: Scale Location Comparison

Appendix B

Appendix C

```
# Data Splitting
set.seed(254646); n1<-110; idx<-sample(nrow(df), n1)  # draw n1 rows
data_clean<-df; dfm_train<-data_clean[idx,]; dfm_test<-data_clean[-idx,]  # train/test
Xvars_mat_train<-model.matrix(`GDP growth (annual %)`~.-Country, data=dfm_train)[,-1]  # drop intercept
Yvar_train<-dfm_train$`GDP growth (annual %)`  # target train
Xvars_mat_test<-model.matrix(`GDP growth (annual %)`~.-Country, data=dfm_test)[,-1]   # drop intercept
Yvar_test<-dfm_test$`GDP growth (annual %)`    # target test

# OLS Regressions without trade and with trade
model <- lm(`GDP growth (annual %)`~`Fertility rate, total (births per woman)`+`Employers, total (% of total employment) (modele
+`Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)` +`Life expectancy at birth, mal
+`Unemployment, total (% of total labor force) (modeled ILO estimate)`+`Export value index (2000 = 100)`+`Population growth (ann
+`GDP per capita, PPP (constant 2011 international $)`+`Net trade in goods and services (BoP, current US$)`
+`Wage and salaried workers, total (% of total employment) (modeled ILO estimate)`,data = df,singular.ok = FALSE)
model_without <- update(model, . ~ . - `Net trade in goods and services (BoP, current US$)`, singular.ok = FALSE)

# OLS Regression with non linear and interaction effects
df2<-within(df,{`GDP per capita, PPP (constant 2011 international $)^2`<-(`GDP per capita, PPP (constant 2011 international $)`)
`Population growth (annual %)^2`<-(`Population growth (annual %)`^2)
`Unemp x LFP`<-`Unemployment, total (% of total labor force) (modeled ILO estimate)`*`Labor force participation rate, total (% o
model_ext <- lm(`GDP growth (annual %)`~`Fertility rate, total (births per woman)`+`Employers, total (% of total employment) (mo
+`Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)`+`Life expectancy at birth, male
+`Unemployment, total (% of total labor force) (modeled ILO estimate)`+`Export value index (2000 = 100)`
+`Population growth (annual %)`+`Population growth (annual %)^2` # nonlinear
+`GDP per capita, PPP (constant 2011 international $)`+`GDP per capita, PPP (constant 2011 international $)^2` # convergence
+`Wage and salaried workers, total (% of total employment) (modeled ILO estimate)`+`Unemp x LFP`,data = df2,singular.ok = FALSE)

models <-list("With Net Trade"=model,"Without Net Trade"=model_without,"Non-Linear Model"=model_ext) #putting in list
df_models <-lapply(names(models),function(name){augment(models[[name]]) %>% mutate(Model=name)}) %>% bind_rows() #tidy dataframe
gg_resfit <-ggplot(df_models, aes(.fitted, .resid, color = Model)) + geom_point(alpha=0.6)+
geom_hline(yintercept=0,linetype="dashed") + facet_wrap(~Model)+labs(title="Residuals vs Fitted Comparison",
x="Fitted values",y="Residuals")+theme_minimal() #Residuals vs Fitted plot
gg_qq <- ggplot(df_models,aes(sample=.std.resid,color=Model))+stat_qq(alpha=0.6)+stat_qq_line()+facet_wrap(~Model)+
labs(title="Normal Q-Q",x ="Theory Quantiles",y="Standardized Residuals")+theme_minimal() # Normal Q-Q plot
gg_scale <- ggplot(df_models, aes(.fitted, sqrt(abs(.std.resid)),color=Model))+geom_point(alpha = 0.6)+
geom_smooth(se = FALSE, color = "black")+facet_wrap(~Model)+labs(title = "Scale-Location Comparison",
x="Fitted values",y="Sqrt(|Standardized Residuals|)")+theme_minimal() #Scale-Location plot

# LASSO Regression
X_lasso <- model.matrix(`GDP growth (annual %)`~`Fertility rate, total (births per woman)`+`Employers, total (% of total employm
`Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)`+`Life expectancy at birth, male
`Unemployment, total (% of total labor force) (modeled ILO estimate)`+`Export value index (2000 = 100)`+
`Population growth (annual %)`+`Population growth (annual %)^2`+`GDP per capita, PPP (constant 2011 international $)`+
`GDP per capita, PPP (constant 2011 international $)^2`+`Net trade in goods and services (BoP, current US$)`+
`Wage and salaried workers, total (% of total employment) (modeled ILO estimate)`+`Unemp x LFP`, # Interaction
data = df2)[, -1] # drop intercept
y_lasso <- df2$`GDP growth (annual %)`
#Cross-validated LASSO
set.seed(2354245);cv_lasso_nl_int<-cv.glmnet(X_lasso,y_lasso,alpha = 1,nfolds = 10,standardize = TRUE)
cv_lasso_nl_int$lambda.min;cv_lasso_nl_int$lambda.1se
coef(cv_lasso_nl_int, s = "lambda.min");coef(cv_lasso_nl_int, s = "lambda.1se")

set.seed(2134);cv_no_std <- cv.glmnet(X_lasso, y_lasso, alpha = 1, nfolds = 10, standardize = FALSE)
coef(cv_no_std, s = "lambda.min") # LASSO without standardization
set.seed(2134);cv_std <- cv.glmnet(X_lasso, y_lasso, alpha = 1, nfolds = 10, standardize = TRUE)
coef(cv_std, s = "lambda.min") # LASSO with standardization (default in glmnet)

coef_no_std<-as.matrix(coef(cv_no_std,s="lambda.min"));coef_std<-as.matrix(coef(cv_std,s="lambda.min"))
comparison <- data.frame(Variable = rownames(coef_no_std),`No Standardization` = round(coef_no_std[,1], 6),
Standardization = round(coef_std[,1], 6)) # Combining into a comparison table

# LASSO Grid search for lambda
```

```r
set.seed(254646);lambda_grid<-seq(0.001,2,length.out=100)
cv_lasso<-cv.glmnet(Xvars_mat_train,Yvar_train,family="gaussian",alpha=1,lambda=lambda_grid,
nfolds = 10,type.measure="mse",standardize=TRUE)
best_lambda_min<-cv_lasso$lambda.min; best_lambda_1se<-cv_lasso$lambda.1se

coef_min<-as.matrix(coef(cv_lasso,s="lambda.min"));coef_1se<-as.matrix(coef(cv_lasso,s="lambda.1se"))
sel_min<-coef_min[coef_min[,1] !=0, ,drop=FALSE] #Get selected variables from extracted coefficients
sel_1se<-coef_1se[coef_1se[,1] != 0, ,drop=FALSE] #non-zero coefficients
df_min<-data.frame(Variable=rownames(sel_min),Coef_min=sel_min[,1]) #Create data frames
df_1se<-data.frame(Variable=rownames(sel_1se),Coef_1se=sel_1se[,1])
comparative_tbl<-merge(df_min,df_1se,by="Variable",all=TRUE)

rmse<-function(y,p) sqrt(mean((y-p)^2)); mae<-function(y,p) mean(abs(y-p))
r2<-function(y,p) 1-sum((y-p)^2)/sum((y-mean(y))^2) # metrics functions
row_metrics<-function(name,y,p){data.frame(Model=name,RMSE=rmse(y,p),
MAE=mae(y,p),R2=r2(y,p),SD_y=sd(y))}
nz_coefs<-function(cv_obj,s){cf<-as.matrix(coef(cv_obj,s=s))
out<-data.frame(Variable=rownames(cf),Coef=cf[,1],row.names=NULL);subset(out,Coef!=0)}
# Lasso Predictions
pred_lasso_min <- as.numeric(predict(cv_lasso, newx = Xvars_mat_test, s = "lambda.min"))
pred_lasso_1se <- as.numeric(predict(cv_lasso, newx = Xvars_mat_test, s = "lambda.1se"))
met_lasso_min  <- row_metrics("Lasso (lambda.min)",  Yvar_test, pred_lasso_min)
met_lasso_1se  <- row_metrics("Lasso (lambda.1se)",  Yvar_test, pred_lasso_1se)
sel_lasso_min  <- nz_coefs(cv_lasso, "lambda.min"); sel_lasso_1se  <- nz_coefs(cv_lasso, "lambda.1se")
pred_base <- rep(mean(Yvar_train), length(Yvar_test)) # Baseline (mean)
met_base  <- row_metrics("Baseline (mean of train)", Yvar_test, pred_base)

# ELASTIC NET, grid search for alpha and training the model
set.seed(2546); alpha_grid<-seq(0.1,0.9,by=0.1);
cv_list<-lapply(alpha_grid,function(a) cv.glmnet(Xvars_mat_train,Yvar_train,family="gaussian",
alpha=a,nfolds=10,type.measure="mse",standardize=TRUE)); cv_mins<-sapply(cv_list, function(cv) min(cv$cvm)); best_i<-which.min(c
best_cv <- cv_list[[best_i]];cat("\nElastic Net - best alpha:",best_a,"| lambda.min:",signif(best_cv$lambda.min,4), "| lambda.1s

lam_tbl<-data.frame(Method=c("LASSO","LASSO"),Metric=c("$ $.min","$ $.1se"),
Value=c(signif(best_lambda_min,4),signif(best_lambda_1se,4))) # existing lasso table
elastic_tbl<-data.frame(Method=c("Elastic Net","Elastic Net","Elastic Net"),Metric=c("$ $","$ $.min","$ $.1se"),
Value=c(best_a,signif(best_cv$lambda.min,4),signif(best_cv$lambda.1se,4))) #adding elastic net results
lam_tbl <- rbind(lam_tbl, elastic_tbl) #combine
knitr::kable(lam_tbl, caption = "Optimal tuning parameters for LASSO and Elastic Net") #display

# Elastic Net Predictions.
pred_en_min<-as.numeric(predict(best_cv, newx = Xvars_mat_test, s = "lambda.min"))
pred_en_1se<-as.numeric(predict(best_cv, newx = Xvars_mat_test, s = "lambda.1se"))
met_en_min<-row_metrics(paste0("Elastic Net  =", best_a, " (lambda.min)"), Yvar_test, pred_en_min)
met_en_1se<-row_metrics(paste0("Elastic Net  =", best_a, " (lambda.1se)"), Yvar_test, pred_en_1se)
sel_en_min<-nz_coefs(best_cv, "lambda.min"); sel_en_1se<-nz_coefs(best_cv, "lambda.1se")

metrics_all<-rbind(met_base, met_lasso_min, met_lasso_1se, met_en_min, met_en_1se) #metric comparison
metrics_tbl<-metrics_all%>%dplyr::select(Model, RMSE, MAE, R2, SD_y) %>%
mutate(across(where(is.numeric),~round(.,3))); best_rmse_idx<-which.min(metrics_all$RMSE)
kable(metrics_tbl,caption="Test Performance (RMSE, MAE, $R^2$, $SD_y$)")%>%
kable_styling(full_width=FALSE,position="center",bootstrap_options=c("striped","hover","condensed"))%>%
row_spec(best_rmse_idx, bold = TRUE)

top_k<-15 #Best features used
top_lasso_min <- head(sel_lasso_min[order(abs(sel_lasso_min$Coef), decreasing = TRUE), ], top_k)
top_lasso_1se <- head(sel_lasso_1se[order(abs(sel_lasso_1se$Coef), decreasing = TRUE), ], top_k)
top_en_min    <- head(sel_en_min[order(abs(sel_en_min$Coef),   decreasing = TRUE), ], top_k)
top_en_1se    <- head(sel_en_1se[order(abs(sel_en_1se$Coef),   decreasing = TRUE), ], top_k)
count_nz <- function(df) sum(df$Variable != "(Intercept)") # amount of variables selected in each model
cat("\n Amount of Variables -> Lasso min:", count_nz(sel_lasso_min),"| Lasso 1se:", count_nz(sel_lasso_1se),
   "| ENet min:", count_nz(sel_en_min), "| ENet 1se:", count_nz(sel_en_1se), "\n")
vars_lasso_min <-sel_lasso_min$Variable; vars_lasso_1se<-sel_lasso_1se$Variable
vars_en_min<-sel_en_min$Variable; vars_en_1se<-sel_en_1se$Variable # extract variable names
all_vars<-unique(c(vars_lasso_min,vars_lasso_1se,vars_en_min,vars_en_1se)) # combined into 1 master vector
comparative_df<-data.frame(Variable=all_vars,L_min=ifelse(all_vars%in%vars_lasso_min,"T","F"),
L_1se=ifelse(all_vars %in% vars_lasso_1se,"T","F"),ENet_min=ifelse(all_vars %in% vars_en_min,"T","F"),
```

```r
ENet_1se = ifelse(all_vars %in% vars_en_1se,"T","F"))

# Ridge and Logistic
df_logistic<-data_clean[ , !(names(data_clean) %in% "Country")]
df_logistic$GrowingMore <- as.integer(df_logistic$`GDP growth (annual %)` > 2.7)
set.seed(1111); train_index<-sample.int(nrow(df_logistic),110) # Data Splitting (Seed One)
train_set<-df_logistic[train_index, ]; test_set  <- df_logistic[-train_index, ]
y_train<-train_set$GrowingMore; y_test<-test_set$GrowingMore
X_train <- model.matrix(GrowingMore ~ . -`GDP growth (annual %)`, data=train_set)[, -1]
X_test  <- model.matrix(GrowingMore ~ . - `GDP growth (annual %)`, data=test_set )[,-1]

cv_ridge<-cv.glmnet(X_train,y_train,family="binomial",alpha=0,nfolds=10,type.measure="deviance",standardize=TRUE)
lam_min<-cv_ridge$lambda.min; lam_1se<-cv_ridge$lambda.1se
cat("lambda.min =",signif(lam_min,4)," | lambda.1se =",signif(lam_1se,4), "\n")
# Prediction on Test Using both variables.
p_min <- as.numeric(predict(cv_ridge, newx = X_test, s = "lambda.min", type = "response"))
p_1se <- as.numeric(predict(cv_ridge, newx = X_test, s = "lambda.1se", type = "response"))
pred_min <- ifelse(p_min >= 0.5, 1, 0); pred_1se <- ifelse(p_1se >= 0.5, 1, 0) # Classify at 0.5

logloss<-function(y,p){p<-pmin(pmax(p,1e-15),1-1e-15);-mean(y*log(p)+(1-y)*log(1-p))}
brier<-function(y,p) mean((p-y)^2) # Metrics for performance
acc_min<-mean(pred_min==y_test); acc_1se<-mean(pred_1se==y_test); ll_min<-logloss(y_test,p_min)
ll_1se<-logloss(y_test,p_1se); br_min<-brier(y_test,p_min); br_1se<-brier(y_test,p_1se)
get_auc<-function(y,p){if(requireNamespace("pROC",quietly=TRUE))as.numeric(pROC::auc(y,p))else NA_real_}
auc_min<-get_auc(y_test,p_min); auc_1se<-get_auc(y_test,p_1se) # optional AUC
results<-data.frame(Model=c("Ridge (lambda.min)","Ridge (lambda.1se)"),Accuracy=c(acc_min,acc_1se),
LogLoss=c(ll_min,ll_1se),Brier=c(br_min,br_1se),AUC=c(auc_min,auc_1se)); print(results,row.names=FALSE)

results_tbl<-results%>%transmute(Model,Accuracy=round(Accuracy,4),LogLoss=round(LogLoss,4),
Brier=round(Brier,4),AUC=ifelse(is.na(AUC),NA,round(AUC,3)))
kable(results_tbl,caption="Ridge (logistic) - Test Metrics")%>%
kable_styling(full_width=FALSE,position="center",bootstrap_options = c("striped","hover","condensed"))

# Final Model Coefficients, and example showing 10 largest (by |coef|) at lambda.1se
coef_min<-coef(cv_ridge,s="lambda.min"); coef_1se<-coef(cv_ridge,s="lambda.1se")
ix<-order(abs(as.numeric(coef_1se)),decreasing = TRUE); print(coef_1se[ix[1:10], ,drop=FALSE])

set.seed(9999); train_index<-sample.int(nrow(df_logistic),110) # Data Splitting (Seed Two)
train_set<-df_logistic[train_index, ]; test_set<-df_logistic[-train_index, ]
y_train<-train_set$GrowingMore; y_test<-test_set$GrowingMore
X_train<-model.matrix(GrowingMore ~ . -`GDP growth (annual %)`, data=train_set)[, -1]
X_test<-model.matrix(GrowingMore ~ . - `GDP growth (annual %)`, data=test_set )[,-1]

# Ridge and Elastic Net.
cv_ridge<-cv.glmnet(X_train,y_train,family="binomial",alpha=0,nfolds = 10,type.measure="deviance",standardize=TRUE)
lam_min<-cv_ridge$lambda.min; lam_1se<-cv_ridge$lambda.1se
cat("lambda.min =", signif(lam_min,4), " | lambda.1se =", signif(lam_1se,4), "\n")
# Prediction on Test Using both variables.
p_min <- as.numeric(predict(cv_ridge, newx = X_test, s = "lambda.min", type = "response"))
p_1se <- as.numeric(predict(cv_ridge, newx = X_test, s = "lambda.1se", type = "response"))
pred_min<-ifelse(p_min>=0.5,1,0); pred_1se<-ifelse(p_1se>=0.5,1,0) # Classify at 0.5

logloss<-function(y,p){p<-pmin(pmax(p,1e-15),1-1e-15); -mean(y*log(p)+(1-y)*log(1-p))} # Metrics for performance
brier<-function(y,p) mean((p-y)^2); acc_min<-mean(pred_min==y_test); acc_1se<-mean(pred_1se==y_test)
ll_min<-logloss(y_test,p_min); ll_1se<-logloss(y_test,p_1se); br_min<-brier(y_test,p_min); br_1se<-brier(y_test,p_1se)
get_auc<-function(y,p){if(requireNamespace("pROC",quietly=TRUE))as.numeric(pROC::auc(y, p))else NA_real_}
auc_min<-get_auc(y_test,p_min); auc_1se<-get_auc(y_test,p_1se) # optional AUC
results1<-data.frame(Model=c("Ridge (lambda.min)","Ridge (lambda.1se)"),Accuracy=c(acc_min,acc_1se),LogLoss=c(ll_min,ll_1se),
Brier=c(br_min,br_1se),AUC=c(auc_min,auc_1se)); print(results, row.names = FALSE)
results_tbl_1<-results1%>%transmute(Model,Accuracy=round(Accuracy,4),LogLoss=round(LogLoss,4),Brier=round(Brier,4),
AUC=ifelse(is.na(AUC),NA,round(AUC,3))); kable(results_tbl_1,caption="Ridge (logistic) - Test Metrics") %>%
kable_styling(full_width = FALSE, position = "center", bootstrap_options = c("striped","hover","condensed"))

coef_min<-coef(cv_ridge,s="lambda.min"); coef_1se<-coef(cv_ridge,s="lambda.1se")
ix<-order(abs(as.numeric(coef_1se)),decreasing=TRUE); print(coef_1se[ix[1:10], ,drop=FALSE])

tblA<-results_tbl_1%>% select(Model,Accuracy,LogLoss,Brier,AUC)%>% mutate(across(-Model,~round(.,4)))
```

```
tblB<-results_tbl%>% select(Model,Accuracy,LogLoss,Brier,AUC)%>% mutate(across(-Model,~round(.,4)))
wide<-inner_join(tblA,tblB,by="Model",suffix=c("_run1","_run2"))
kable(wide,col.names=c("Model","Accuracy","LogLoss","Brier","AUC","Accuracy","LogLoss","Brier","AUC"),
caption="Ridge (logistic) - Test Metrics (different seeds)") %>% kable_styling(full_width=FALSE,position="center",
bootstrap_options=c("striped","hover","condensed")) %>% add_header_above(c(" " =1, "Run 1" =4, "Run 2" =4))

kable(comparison[ , -1],caption="Comparison of LASSO Coefficients With and Without Standardization") # Pretty table
knitr::kable(comparative_df,caption="Variable selection across LASSO and Elastic Net") # Display nicely
knitr::kable(comparative_tbl, caption = "Variables Selected under  .min vs  .1se") # Show the table


gg_resfit


gg_qq


gg_scale
```