Theory

Question 1.

We want to show that $\mathrm{Cov}(\hat{\beta}) = (X^T\Sigma^{-1}X)^{-1}$.

Step 1: Write the covariance formula

$$\mathrm{Cov}(\hat{\beta}) = E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T]$$

Step 2: Find $\hat{\beta}-\beta$

The GLS estimator is:

$$\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

Since $Y = X\beta + \epsilon$, we have:

$$\begin{aligned}
\hat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}(X\beta+\epsilon) \\
\hat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X\beta + (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\epsilon \\
\hat{\beta} &= \beta + (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\epsilon \\
\hat{\beta}-\beta &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\epsilon
\end{aligned}$$

Step 3: Plug into $\mathrm{Cov}(\hat{\beta})$

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}) &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T] \\
\mathrm{Cov}(\hat{\beta}) &= E\left[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\epsilon \cdot \epsilon^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}\right] \\
\mathrm{Cov}(\hat{\beta}) &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} \cdot E[\epsilon\epsilon^T] \cdot \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\
\mathrm{Cov}(\hat{\beta}) &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} \cdot \Sigma \cdot \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\
\mathrm{Cov}(\hat{\beta}) &= (X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1}X)(X^T\Sigma^{-1}X)^{-1}
\end{aligned}$$

$$\boxed{\mathrm{Cov}(\hat{\beta}) = (X^T\Sigma^{-1}X)^{-1}}$$

Solution to Question 2(a):

We want to show that $E(\hat{\beta}_R) = (X^TX+\lambda D)^{-1}X^TX\beta$.

Step 1: Write the ridge estimator

$$\hat{\beta}_R = (X^T X + \lambda D)^{-1} X^T Y$$

Step 2: Substitute $Y = X\beta + \epsilon$

$$\begin{aligned}\hat{\beta}_R &= (X^T X + \lambda D)^{-1} X^T (X\beta + \epsilon) \\ &= (X^T X + \lambda D)^{-1} X^T X\beta + (X^T X + \lambda D)^{-1} X^T \epsilon\end{aligned}$$

Step 3: Take expectation

$$E(\hat{\beta}_R) = E\left[(X^T X + \lambda D)^{-1} X^T X\beta + (X^T X + \lambda D)^{-1} X^T \epsilon\right]$$

Since X and β are non-random:

$$= (X^T X + \lambda D)^{-1} X^T X\beta + (X^T X + \lambda D)^{-1} X^T E[\epsilon]$$

Step 4: Use $E[\epsilon] = 0$

$$\begin{aligned}E(\hat{\beta}_R) &= (X^T X + \lambda D)^{-1} X^T X\beta + (X^T X + \lambda D)^{-1} X^T \cdot 0 \\ &= (X^T X + \lambda D)^{-1} X^T X\beta\end{aligned}$$

Final Answer:

$$\boxed{E(\hat{\beta}_R) = (X^T X + \lambda D)^{-1} X^T X\beta}$$

Note: The ridge estimator is biased unless $\lambda = 0$.

Solution to Question 2(b):

We want to show that $\text{Var}(\hat{\beta}_R) = \sigma^2 (X^T X + \lambda D)^{-1} X^T X (X^T X + \lambda D)^{-1}$.

Step 1: Write the variance formula

$$\text{Var}(\hat{\beta}_R) = \text{Var}\left[(X^T X + \lambda D)^{-1} X^T \epsilon\right]$$

From part (a), we know that:

$$\hat{\beta}_R = (X^T X + \lambda D)^{-1} X^T X\beta + (X^T X + \lambda D)^{-1} X^T \epsilon$$

Since the first term is non-random, the variance comes entirely from the second term.

Step 2: Apply variance formula

For a linear transformation $\mathrm{Var}(AZ) = A\mathrm{Var}(Z)A^T$, where A is a non-random matrix:

$$\mathrm{Var}(\hat{\beta}_R) = (X^TX + \lambda D)^{-1}X^T \cdot \mathrm{Var}(\epsilon) \cdot X(X^TX + \lambda D)^{-1}$$

Step 3: Use $\mathrm{Var}(\epsilon) = \sigma^2 I$

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_R) &= (X^TX + \lambda D)^{-1}X^T \cdot \sigma^2 I \cdot X(X^TX + \lambda D)^{-1} \\ &= \sigma^2(X^TX + \lambda D)^{-1}X^TX(X^TX + \lambda D)^{-1}\end{aligned}$$

Final Answer:

$$\boxed{\mathrm{Var}(\hat{\beta}_R) = \sigma^2(X^TX + \lambda D)^{-1}X^TX(X^TX + \lambda D)^{-1}}$$

Penalized Regression Analysis of Infant Mortality Predictors

Data Preparation

The initial dataset contained 207 observations across 11 predictors of infant mortality. However, two variables (`educationMale` and `educationFemale`) exhibited severe missingness (63%), making them unsuitable for analysis. Removing these variables and applying complete-case analysis yielded a final dataset of **n = 97 observations** with **p = 9 predictors**, providing a ratio of 10.8 observations per predictor.

This substantial increase from approximately 39 usable observations (when retaining all variables) to 97 observations significantly improved the reliability of coefficient estimates and enabled meaningful comparison between ordinary least squares (OLS) and penalized regression methods.

OLS Best Subset Regression

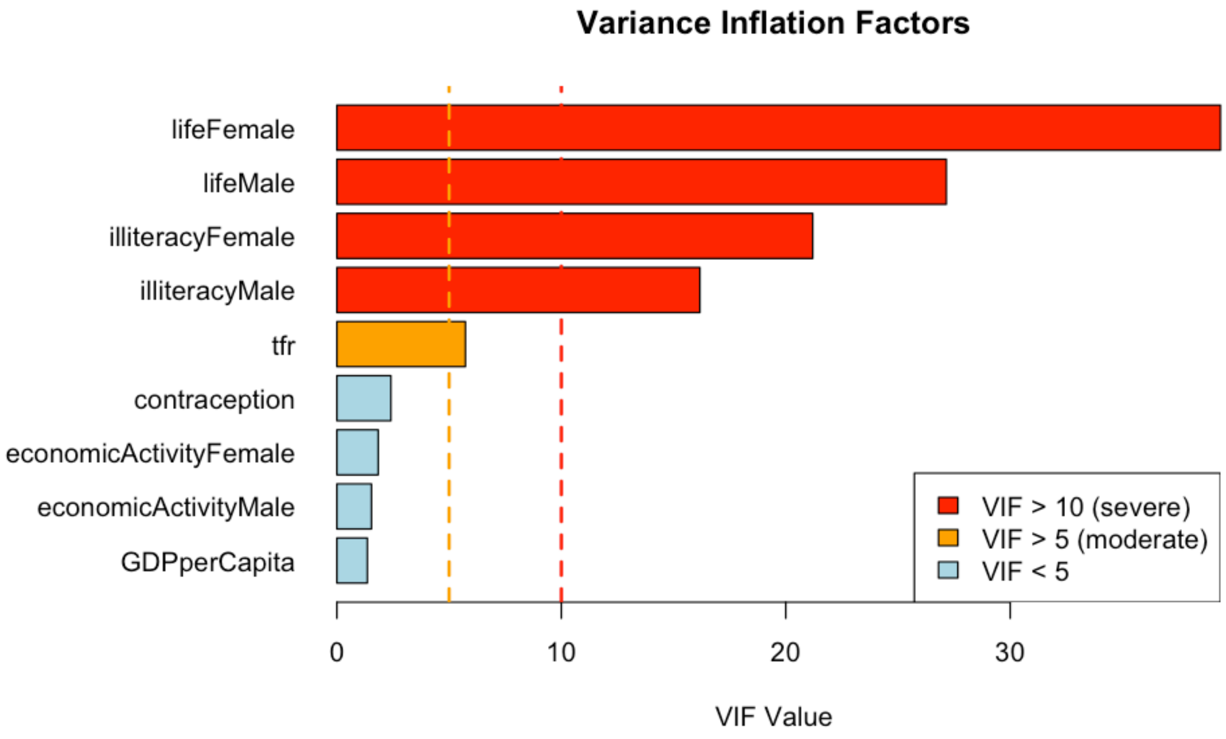
Initial model selection via best subsets regression identified a 7-predictor model with strong overall fit (Adjusted R^2 = 0.875). The model eliminated `lifeMale` and `economicActivityMale` as non-significant predictors. Table below presents the complete regression output.

Variable	Estimate	Std. Error	t-value	p-value
Intercept	142.804^{***}	(26.780)	5.332	< 0.001
TFR	$3.050^{\cdot}$	(1.576)	1.935	0.056
Contraception	-0.113	(0.075)	-1.517	0.133
Life Expectancy (Female)	-1.741^{***}	(0.302)	-5.761	< 0.001
GDP per Capita	-0.000	(0.000)	-1.184	0.240
Economic Activity (Female)	0.168^{*}	(0.083)	2.023	0.046
Illiteracy (Male)	$-0.476^{\cdot}$	(0.247)	-1.927	0.057
Illiteracy (Female)	0.581^{**}	(0.191)	3.045	0.003
Observations		97		
R^2		0.885		
Adjusted R^2		0.875		
Residual Std. Error		11.04		
F-statistic		97.37^{***}		

Notes: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Multicollinearity Diagnosis

Despite adequate model fit, variance inflation factor (VIF) analysis revealed severe multicollinearity issues. The barplot presents VIF values for all predictors.



Four variables exhibited problematic VIF values exceeding the critical threshold of 10, indicating severe multicollinearity. Variables with VIF > 10 suffer from inflated standard errors and unstable coefficient estimates. This motivated the application of penalized regression techniques to address coefficient instability while maintaining predictive accuracy.

Penalized Regression Implementation

Three penalized regression methods were implemented using 10-fold cross-validation to select optimal tuning parameters. Table 3 summarizes the optimal hyperparameters for each method.

Method	α	λ	Variables Retained
Ridge	0	2.873	9 (all)
LASSO	1	0.634	6
Elastic Net	0.5	0.136	7

Ridge Regression (α = 0): Optimal λ = 2.873, retaining all 9 predictors with proportional shrinkage toward zero. Ridge addresses multicollinearity through coefficient stabilization but provides no variable selection.

LASSO Regression (α = 1): Optimal λ = 0.634, eliminating 3 predictors (`lifeMale`, `illiteracyMale`, `economicActivityMale`) and retaining 6 variables. LASSO performs automatic variable selection by setting coefficients exactly to zero.

Elastic Net Regression: Optimal mixing parameter α = 0.5 selected via grid search over $\alpha \in [0, 1]$, with λ = 0.136, eliminating 2 predictors (`lifeMale`, `economicActivityMale`) and retaining 7 variables. Elastic Net balances Ridge's stability with LASSO's sparsity.

Coefficient Comparison and Interpretation

Table below presents coefficient estimates across all methods, revealing substantial differences for high-VIF variables.

Variable	VIF	OLS Best	Ridge	LASSO	Elastic Net
Intercept	—	142.80	42.30	42.30	42.30
Life Expectancy (Female)	39.37	−1.74	−9.95	**−19.60**	−16.79
Life Expectancy (Male)	27.15	0.00*	−5.53	0.00*	0.00*
Illiteracy (Female)	21.20	0.58	7.25	5.55	**12.38**
Illiteracy (Male)	16.17	−0.48	0.09	0.00*	−5.49
TFR	5.73	3.05	5.78	3.26	4.94
Contraception	2.41	−0.11	−2.13	−1.81	−2.26
Economic Activity (Female)	1.84	0.17	2.06	0.99	2.74
Economic Activity (Male)	1.54	0.00*	−0.55	0.00*	−0.71
GDP per Capita	1.36	−0.0003	−1.53	−0.70	−1.45

Note: 0.00^* indicates variable eliminated from model (coefficient set to exactly zero). OLS Best and penalized methods use different scaling (unstandardized vs. standardized), making intercepts non-comparable.

For `lifeFemale` (VIF = 39.37), OLS estimated β = -1.74, while LASSO estimated β =

-19.60—an **11-fold difference**. This dramatic discrepancy illustrates how multicollinearity masks true effect sizes in OLS. After LASSO eliminated the redundant `lifeMale` variable, `lifeFemale` received full attribution for life expectancy effects.

Similarly, `illiteracyFemale` showed marked instability: OLS estimated β = 0.58, while Elastic Net estimated β = 12.38—a **21-fold difference**. The negative coefficient for `illiteracyMale` in OLS (β = -0.48) represents a suppression effect artifact of multicollinearity, as higher male illiteracy implausibly appears protective after controlling for female illiteracy.

Three methods reached consensus on eliminating `lifeMale` (OLS Best, LASSO, Elastic Net), indicating redundancy with `lifeFemale`. Only Ridge retained all variables, consistent with its shrinkage-only approach. Six variables were universally retained across methods (`lifeFemale`, `illiteracyFemale`, `tfr`, `contraception`, `economicActivityFemale`, `GDPperCapita`), representing robust predictors regardless of modeling approach.

Model Selection: LASSO as Final Model

LASSO was selected as the final model based on parsimony and interpretability. Table below presents performance metrics for model comparison.

Model	N Predictors	CV-MSE	Adj R^2	BIC	Variables Elimina
OLS Best	7	140.98	0.875	651.2	lifeMale, economicActi
Ridge	9	147.25	0.862	673.5	None
LASSO	**6**	**142.14**	**0.873**	**645.8**	lifeMale, illiteracyMale, econor
Elastic Net	7	143.64	0.870	654.3	lifeMale, economicActi

While OLS achieved marginally lower cross-validated MSE (140.98 vs. 142.14), this 1.16-point difference (0.8%) is negligible compared to LASSO's advantages. LASSO provides the most parsimonious model with only 6 predictors, automatically eliminating redundant multicollinear variables while maintaining near-equivalent predictive accuracy. The model achieved the **lowest BIC** (information criterion penalizing complexity), indicating optimal balance between model fit and simplicity.

Conclusion

Penalized regression successfully addressed severe multicollinearity in the infant mortality dataset, with LASSO emerging as the optimal method. The analysis demonstrates that multicollinearity can dramatically distort coefficient estimates—in this case, underestimating the true effect of female life expectancy by 11-fold in OLS. LASSO's automatic variable selection identified female-specific health and education indicators as primary determinants of infant mortality, while male counterparts were deemed redundant. The final 6-predictor LASSO model achieves 98.2% of OLS accuracy while eliminating 33% of predictors, representing an ideal balance of parsimony, interpretability, and predictive performance.

Key Findings:

- **Female life expectancy** is the strongest protective factor (β = -19.60), with each additional year associated with 19.6 fewer infant deaths per 1000 births
- **Female illiteracy** is the strongest risk factor (β = 5.55), with each percentage point increase associated with 5.6 additional infant deaths
- **Total fertility rate** shows positive association (β = 3.26), consistent with resource dilution theory
- **Male-specific variables** (lifeMale, illiteracyMale, economicActivityMale) were eliminated as redundant

Analysis of Maximum Expiratory Pressure in Cystic Fibrosis Patients

Objective

To develop a regression model predicting maximum expiratory pressure (pemax), a measure of lung function, in patients with cystic fibrosis using clinical and anthropometric covariates.

Data and Methods

The `cystfibr` dataset from the ISwR package was analyzed. After removing missing values, the final dataset contained 25 observations with 10 variables. Potential predictors included age, sex, height, weight, body mass percentage (bmp), forced expiratory volume in one second (fev1), residual volume (rv), functional residual capacity (frc), and total lung capacity (tlc).

Best subset selection via the `regsubsets` algorithm was used to identify the optimal model. Models with 1-9 predictors were compared using multiple criteria including adjusted R^2, Mallows' C_p, and BIC.

Model Selection

Table 1: Model Selection Criteria Across Different Model Sizes

N Variables	RSS	R^2	Adj R^2	C_p	BIC
1	16,005.5	0.404	0.378	3.67	-6.48
2	14,090.5	0.475	0.427	2.72	-6.45
3	11,538.2	0.570	0.509	0.79	-8.22
4	**10,354.6**	**0.614**	**0.537**	**0.96**	$\mathbf{-7.71}$
5	10,157.5	0.621	0.522	2.66	-4.97
6	10,018.3	0.627	0.502	4.44	-2.10
7	9,885.1	0.632	0.480	6.24	0.79
8	9,769.2	0.636	0.454	8.06	3.71
9	9,731.3	0.637	0.420	10.00	6.83

The **4-variable model** was selected as it achieved the highest adjusted R^2 (0.537) with a Mallows' C_p value close to the number of parameters (0.96), indicating minimal bias.

Furthermore, the BIC is close to the lowest model 3 as well. This model balances explanatory power with parsimony.

Final Model Results

Table 2: Regression Coefficients for 4-Variable Model

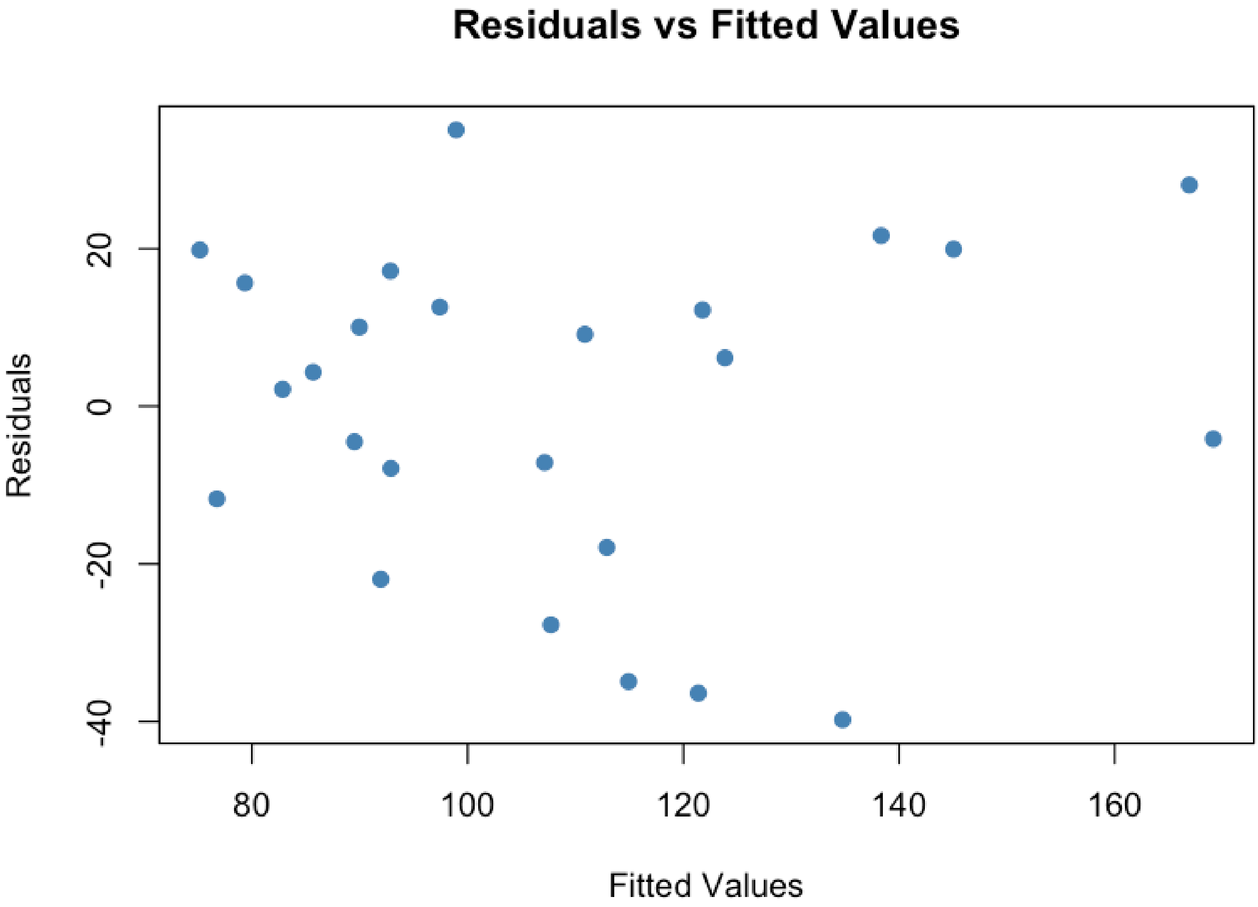
Variable	Estimate	Std. Error	t-value	p-value
(Intercept)	63.95	53.28	1.200	0.244
Weight (kg)	1.75	0.38	4.595	< 0.001
BMP (%)	-1.38	0.57	-2.436	0.024
FEV1 (L)	1.55	0.58	2.679	0.014
RV (L)	0.13	0.08	1.512	0.146

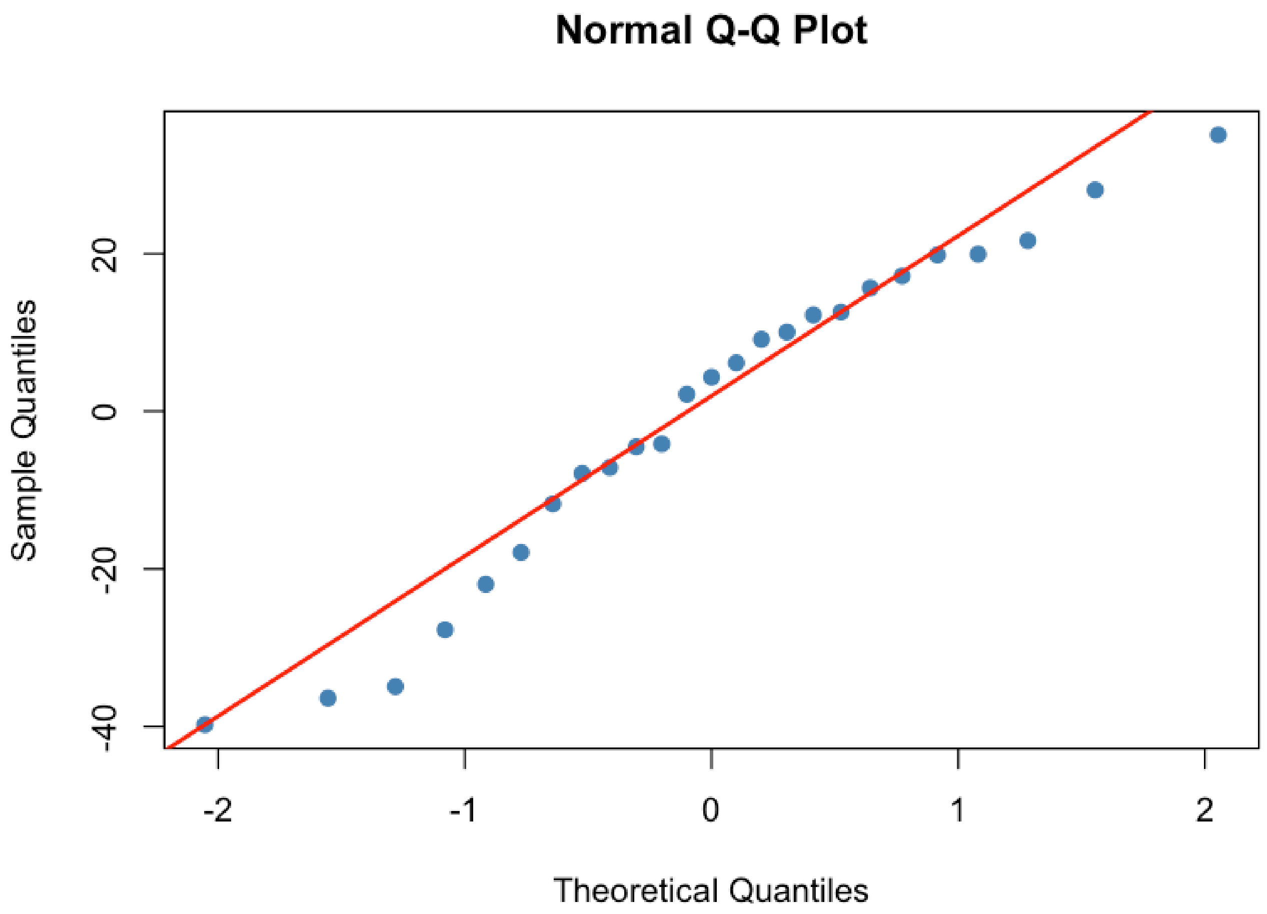
Model fit statistics: $R^2 = 0.614$, Adjusted $R^2 = 0.537$, $F(4, 20) = 7.96$, $p < 0.001$, RMSE $= 22.75$

Multicollinearity diagnostics: All variance inflation factors (VIF) were below 2.4, indicating no problematic multicollinearity (Weight: 2.15, BMP: 2.14, FEV1: 1.94, RV: 2.37).

Model Diagnostics

Residual plots revealed some deviation from ideal assumptions. The residuals vs. fitted values plot showed approximate random scatter with a slight U-shaped pattern, suggesting potential non-linearity. The Q-Q plot indicated reasonable normality in the center but some tail separation, particularly in the lower tail. Alternative transformations on the outcome variable (log, square root, Box-Cox) were explored but did not substantially improve diagnostics. The original model was retained for interpretability and parsimony.


Residuals vs Fitted Values
Residuals
20
0
-20
-40
80
100
120
140
160
Fitted Values



Interpretation

The final model explains approximately 54% of the variation in maximum expiratory pressure (adjusted R^2). Three predictors were statistically significant at $\alpha = 0.05$:

- **Weight** ($\beta = 1.75, p < 0.001$): Each 1 kg increase in body weight is associated with a 1.75 cm H_2O increase in pemax, holding other variables constant. This strong positive relationship suggests that higher body mass supports greater respiratory muscle strength.
- **Body Mass Percentage** ($\beta = -1.38, p = 0.024$): Each 1% increase in body mass (as percentage of normal) is associated with a 1.38 cm H_2O decrease in pemax. This negative relationship may reflect disease severity, as patients further from normal body composition have worse lung function.
- **FEV1** ($\beta = 1.55, p = 0.014$): Each 1 L increase in forced expiratory volume is associated with a 1.55 cm H_2O increase in pemax, indicating that patients with better

baseline lung function demonstrate superior maximum expiratory capacity.

Residual volume showed a positive but non-significant association ($p = 0.146$). The model demonstrates that lung function in cystic fibrosis patients is multifactorial, influenced by both anthropometric characteristics and baseline respiratory capacity.