REGRESSION AND TIME SERIES ANALYSIS OF KAZAKHSTAN'S PRIMARY REAL ESTATE MARKET

AISHA SMAGULOVA, ZARINA GABDULLINA

Capstone Project
Supervisor: Piotr Sebastian Skrzypacz
Second Reader: Dongming Wei

Abstract. This study employs regression and time series analysis to study and forecast primary real estate prices in Kazakhstan. By analyzing variables such as average wages, USD/KZT exchange rates, GDP per capita and population growth, the research quantifies their impact on property valuations. The methodology includes rigorous statistical techniques and residual analysis. The Regression model is created by regressing real estate prices on macroeconomic factors non-linearly to capture the relationship adequately. In addition, ARIMA and Cubic Spline regressions are employed to model the trend of housing prices and enhance forecasting accuracy. Comparative analysis indicates that time series models more effectively capture the underlying trend and demonstrate better forecasting abilities. Findings reveal significant correlations between economic indicators and real estate prices, providing a predictive framework for future market trends. This research offers valuable insights for policymakers, investors, and stakeholders in Kazakhstan's real estate sector, enhancing the understanding of market dynamics through a mathematical lens.

1. Introduction

The primary real estate market refers to newly built properties that are sold directly by developers to buyers. It plays a crucial role in shaping urban development, meeting housing demand, and driving economic growth. Analyzing the primary market is essential for understanding price trends, evaluating affordability, and making informed investment or policy decisions. In this study, three different models will be used to analyze data and make forecasts.

Recent literature supports the selection of macroeconomic variables such as Gross Domestic Product per capita of the country, average salary of the people, foreign exchange fluctuations, and population growth as significant predictors of housing price dynamics. Non-Kazakhstani housing market studies reveal that the GDP and unemployment are the strongest causal determinants of the housing prices, while the influence of other macroeconomic factors such as interest rate and inflation does not show consistency [11, 3]. Based on the findings of the past literature our study does not include inflation and interest rate as the determinants of the real estate prices. Similarly, in the context of Kazakhstan, housing prices are closely linked to the average wages of the household, GDP, and population showing linear relationship. On the other hand, exchange rate reflects non-linear relationship, as the construction costs mainly depend on the import prices [11].

Numerous studies have applied time series models to improve price prediction accuracy. Several works have compared forecasting methods such as ARIMA, GARCH, and Regime-Switching, with ARIMA demonstrating superior performance in capturing housing price dynamics in both developed and emerging markets [2, 8]. Additionally, when applied to regional housing data, ARIMA models often outperform alternatives like Vector Autoregression (VAR) in terms of forecasting accuracy[10]. These findings highlight the robustness and efficiency of ARIMA models for real estate price analysis. Following this evidence, our project applies ARIMA modeling to analyze and forecast real estate prices in Kazakhstan's primary housing market.

Also, some research has been done regarding the application of smoothing spline methods in modeling housing prices. Studies have shown that smoothing splines are particularly effective in capturing nonlinear relationships within housing data, outperforming linear models in terms of goodness of fit and predictive accuracy [1, 5]. In particular, spline smoothing has been found to perform well even with unevenly spaced data and a large number of explanatory variables, demonstrating both theoretical and empirical advantages over other nonparametric techniques such as kernel regressions or k-NN methods [1]. Additionally, penalized spline regression has been shown to better estimate price functions compared to linear hedonic models, revealing hidden nonlinear effects in housing attributes [5]. While most of these studies apply multivariate models, our project focuses on univariate smoothing spline regression to model and forecast housing price trends based solely on time. This will enable a focused and interpretable analysis of price dynamics in Kazakhstan's primary housing market.

2. Methods

2.1. **Linear Regression.** Multiple polynomial regression is an extension of multiple linear regression that allows for modeling nonlinear relationships between the dependent variable and one or more independent variables by including polynomial terms of the predictors.

The general form of a multiple polynomial regression model with one independent variable expressed as a cubic polynomial and others remaining linear is [9]:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1t}^2 + \beta_3 x_{1t}^3 + \beta_4 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t \tag{1}$$

where:

- y_t is the dependent variable at time t,
- x_{1t} is the independent variable with a polynomial relationship,
- $x_{2t}, \ldots, x_{kt}$ are additional independent variables entering the model linearly,
- β_0 is the intercept term,
- $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients,
- ε_t is the error term, assumed to be independently and identically distributed with mean zero and constant variance.

Polynomial terms are used to capture nonlinear patterns, e.g, curvature in the relationship between variables.

The parameters are estimated using Ordinary Least Squares (OLS), which minimizes the sum of squared residuals (SSR):

$$\min_{\beta_0,\ldots,\beta_k} \sum_{t=1}^{n} \left(y_t - \beta_0 - \sum_{j=1}^{k} \beta_j x_{jt} \right)^2 \tag{2}$$

Minimizing SSR with respect to each β_j (for $j = 0, 1, \ldots, k$) involves taking the partial derivative of the objective function with respect to each coefficient and setting the result equal to zero:

$$\frac{\partial SSR}{\partial \beta_j} = -2 \sum_{t=1}^{n} (y_t - \hat{y}_t) \cdot \frac{\partial \hat{y}_t}{\partial \beta_j} = 0. \tag{3}$$

Since the partial derivative of the fitted value $\hat{y}_t$ with respect to β_j is simply x_{jt}, we obtain the following system of normal equations:

$$\sum_{t=1}^{n} (y_t - \hat{y}_t)\, x_{jt} = 0, \quad \text{for all } j = 0, 1, \ldots, k. \tag{4}$$

This system of $k+1$ equations can be solved simultaneously to obtain the OLS estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. These estimates minimize the discrepancy between the observed values y_t and the values predicted by the model $\hat{y}_t$, under the classical assumptions of linear regression.

To examine the appropriateness of regression model for the data at hand, we require the fitted values $\hat{Y}_i$ and the residuals $e_i = Y_i - \hat{Y}_i$.

2.2. **Time Series Model.** To model the behavior of real estate prices over time, we employ the ARIMA model, which stands for Autoregressive Integrated Moving Average with parameters p, d, and q, where [4]:

- p is the order of the autoregressive (AR) part,
- d is the degree of differencing required to make the series stationary,
- q is the order of the moving average (MA) part.

General Equation of ARIMA(p, d, q):

$$\nabla^d y_t = \phi_1 \nabla^d y_{t-1} + \phi_2 \nabla^d y_{t-2} + \cdots + \phi_p \nabla^d y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \tag{5}$$

where:

- $\nabla^d y_t$ is the differenced series,
- ϕ_i are the autoregressive coefficients,
- θ_j are the moving average coefficients,
- ϵ_t is the white noise error term.

Parameter Estimation and Model Selection: To select the most appropriate ARIMA model, a visual inspection through the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, and criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) will be used.

After selecting the best model, residual diagnostics will be conducted to assess whether the assumptions of the ARIMA model were satisfied. This includes checking for autocorrelation in residuals, normality, and constant variance.

2.3. **Cubic Spline Regression.** A smoothing cubic spline is a nonparametric regression technique used to fit a smooth curve to data while balancing two goals[6]:

1) Goodness of fit to the data (minimizing squared errors),

2) Smoothness of the curve (penalizing excessive wiggliness via the second derivative).

The smoothing spline estimator $\hat{f}(x)$ is the solution to the following penalized least-squares minimization problem:

$$\hat{f} = f \in \mathcal{S} \left\{ \underbrace{\sum_{i=1}^{n} (y_i - f(x_i))^2}_{\text{Goodness of fit}} + \lambda \underbrace{\int_a^b [f''(x)]^2 \, dx}_{\text{Roughness penalty}} \right\} \tag{6}$$

$$f(x) = \sum_{j=1}^{d} \gamma_j B_j(x) \tag{7}$$

where:

- $\mathcal{S}$ is the space of natural cubic splines with knots at each unique x_i
- f satisfies the natural boundary conditions:

$$f''(a) = f''(b) = 0 \tag{8}$$

- $\lambda \geq 0$ is the smoothing parameter controlling the trade-off between fit and smoothness

Key Properties:

- **Knots**: Placed at each unique x_i value
- **Smoothness**: $f \in C^2$ (twice continuously differentiable)
- **Boundary behavior**: Linear beyond the range of the data

For this nonparametric model, the time will be used as the independent variable.

To model the relationship, the smooth.spline() function will be employed in R. This function automatically places knots at each unique x values and selects the smoothing parameter λ through generalized cross-validation (GCV).

This approach allows for flexible modeling of nonlinear relationships while controlling for overfitting through a roughness penalty based on the integrated squared second derivative of the fitted function.

3. Results and Discussion

3.1. **Linear Regression.** To examine the determinants of housing prices, a multiple linear regression model was estimated using wage, population, GDP per capita, and the exchange rate as explanatory variables. To define the relationship between the real estate prices and explanatory variables, we obtained the scatter plot for each of the independent variables against the price.

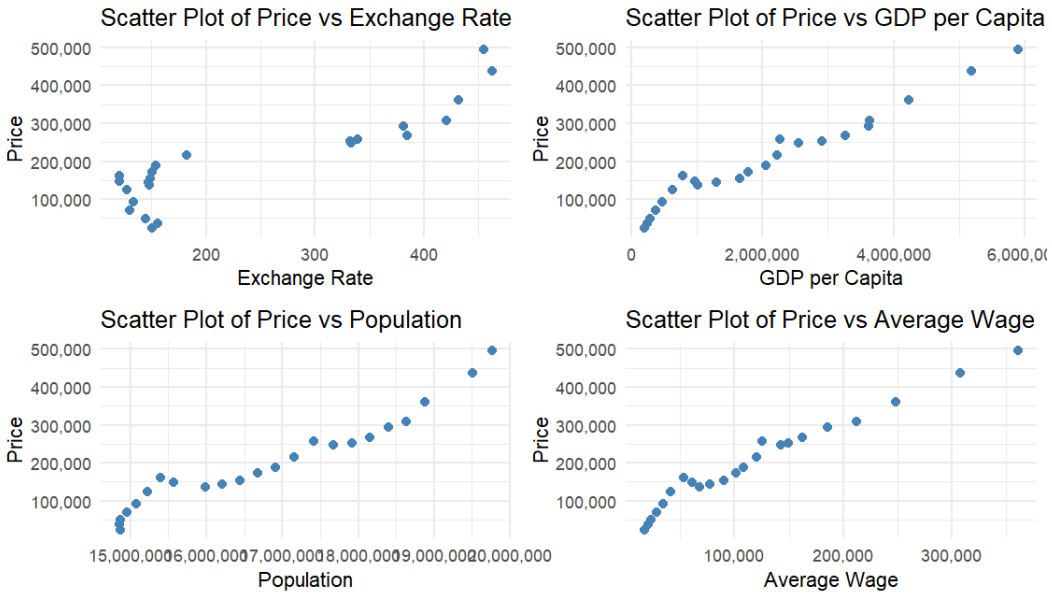

FIGURE 1. Scatter plot of the independent variables against the price

As illustrated in *Figure 1*, we assume that all explanatory variables except for the exchange rate exhibit a linear relationship with the price, while the exchange rate captures nonlinearity following the polynomial trend.

The use of a third-degree polynomial allows us to capture potential nonlinear effects of the exchange rate on housing prices. The model in *Figure 2* demonstrates an excellent fit, with an R^2 of 0.9854, indicating that approximately 98.5% of the variation in housing prices is explained by the model. The overall regression is statistically significant at the 1% level (p-value < 0.001).

```
Call:
lm(formula = price ~ poly(exchange_rate, 3, raw = TRUE) + wage +
    population + gdp_capita, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-24973  -9063  -3771   4413  33860

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            1.809e+04  3.457e+05   0.052  0.95893
poly(exchange_rate, 3, raw = TRUE)1 -5.468e+03  1.446e+03  -3.783  0.00163 **
poly(exchange_rate, 3, raw = TRUE)2  2.024e+01  5.289e+00   3.826  0.00149 **
poly(exchange_rate, 3, raw = TRUE)3 -2.381e-02  6.269e-03  -3.798  0.00158 **
wage                                   1.943e+00  5.858e-01   3.317  0.00436 **
population                             2.982e-02  2.373e-02   1.257  0.22692
gdp_capita                            -4.467e-02  4.096e-02  -1.091  0.29158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17420 on 16 degrees of freedom
Multiple R-squared:  0.9854,    Adjusted R-squared:  0.9799
F-statistic: 179.5 on 6 and 16 DF,  p-value: 9.237e-14
```

FIGURE 2. Multiple linear regression model

The estimated regression equation for the real estate prices is:

$$\begin{aligned} \text{price} = {} & 18090 - 5468 \cdot \text{exchange\_rate} + 20.24 \cdot \text{exchange\_rate}^2 \\ & - 0.02381 \cdot \text{exchange\_rate}^3 + 1.943 \cdot \text{wage} \\ & + 0.02982 \cdot \text{population} - 0.04467 \cdot \text{gdp\_per\_capita} + \varepsilon \end{aligned} \tag{9}$$

At lower levels of the exchange rate linear term dominates and the price falls. Then, since the real estate prices depend on the import price rises. At higher levels of the exchange rate, when the depreciation happens, the market collapse. As for the wage, 1000 tenge rise in the average salary leads to the increase of the price by 1,943 tenge. Population seems to be insignificant determinant of the price, however, the sign of the population aligns with the economic theory. Estimated coefficient of the GDP per capita reveals that increase of the GDP per capita by 10,000 tenge decreses the price of the real estate by 4,467 tenge, although this variable is not a significant factor affecting the price.

The plot of the residuals against the fitted values shows no clear pattern or systematic trend, which supports our assumptions that the appropriate model would be a curvilinear regression function. Also, it indicates that the model satisfies the constant variance of the errors.

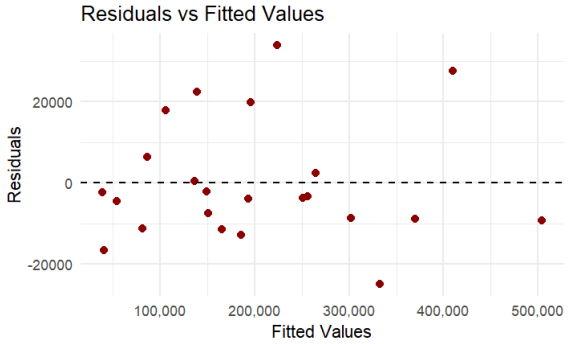

FIGURE 3. Residuals vs Fitted values

For checking the normality of the observations, we implemented the Q-Q plot. From the *Figure 4*, it can be observed that most of the residuals lie on the straight line, indicating that the normality holds.

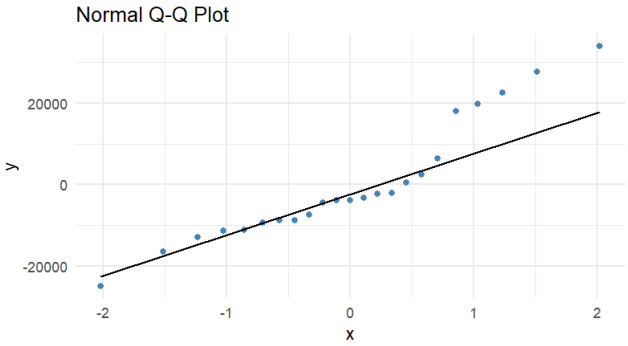

FIGURE 4. Q-Q Plot for regression model residuals

3.2. **Time Series Analysis.** First, we check if our time series data is stationary or not using the Augmented Dickey-Fuller and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test.

From *Figure 5A*, it is clear that the data is not stationary: the ADF test has a p-value greater than 0.05, while the KPSS test has a p-value less than 0.05. After first differencing, the Augmented Dickey-Fuller test did not reject the null hypothesis of non-stationarity, but the KPSS test showed stationarity. Given the short time series and supporting diagnostic checks, the series is treated as stationary after first differencing.

```
> adf.test(data$prices)

        Augmented Dickey-Fuller Test

data:  data$prices
Dickey-Fuller = -0.638, Lag order = 2,
p-value = 0.9635
alternative hypothesis: stationary

> kpss.test(data$prices)

        KPSS Test for Level Stationarity

data:  data$prices
KPSS Level = 0.83712, Truncation lag
parameter = 2, p-value = 0.01
```

(A) Before Differencing

```
> adf.test(diff_prices)

        Augmented Dickey-Fuller Test

data:  diff_prices
Dickey-Fuller = -1.5547, Lag order = 2,
p-value = 0.742
alternative hypothesis: stationary

> kpss.test(diff_prices)

        KPSS Test for Level Stationarity

data:  diff_prices
KPSS Level = 0.3149, Truncation lag
parameter = 2, p-value = 0.1
```

(B) After First Differencing

FIGURE 5. ADF and KPSS Stationarity Tests

The autocorrelation function (ACF) plot shows how current values of the time series are correlated with past values at different lags. In this plot, there is a strong correlation at lag 1, and the correlations for other lags are much smaller and mostly within the confidence bounds. This suggests that the time series is primarily influenced by the immediately previous value and does not have strong long-term memory.

The partial autocorrelation function (PACF) plot shows the direct relationship between current and past values, removing the influence of intermediate lags. In this plot, there is a significant spike at lag 1, but the remaining lags are not significant. This also indicates that only the most recent lag is important.

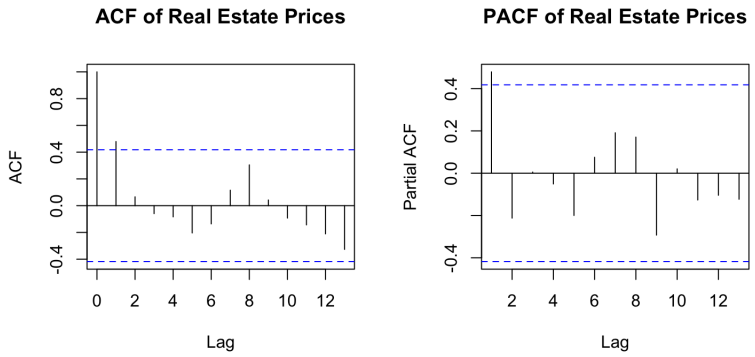

FIGURE 6. ACF and PACF plots

Based on both plots, a good model to consider would be an autoregressive model of order one, written as AR(1), or if the data was differenced to achieve stationarity, an ARIMA(1,1,0) model. This is because both plots suggest that only lag 1 is important for predicting future values.

From *Figure 7*, it's clear that assumption made after ACF and PACF plots analysis was confirmed, and ARIMA (1,1,0) model was selected as the best-fitting model based on the lowest AIC (Akaike Information Criterion).

```
> summary(best_model)
Series: ts_data
ARIMA(1,1,0) with drift

Coefficients:
         ar1      drift
      0.5261  22691.232
s.e.  0.1878   8001.159

sigma^2 = 378517319:  log likelihood = -247.6
AIC=501.2   AICc=502.53   BIC=504.47
```

FIGURE 7. Best Time Series Model

The model includes:

- One autoregressive (AR) term, meaning the current value is influenced by the previous one.
- First-order differencing to remove non-stationarity in the data.
- No moving average (MA) terms, as past forecast errors are not included. The inclusion of a drift term allows the model to capture a consistent upward trend over time, even after differencing.

The residual plot shows no clear pattern, suggesting the model captures the trend well. The ACF plot of residuals indicates minimal autocorrelation, supporting model adequacy. The residual histogram approximates normality, further validating model assumptions. Overall, the residuals appear to be white noise, indicating the ARIMA (1,1,0) with drift is an appropriate model for the data.

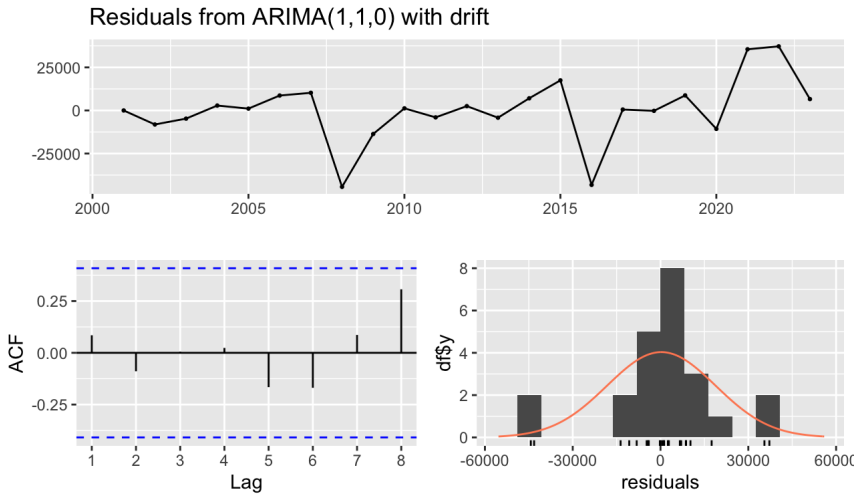

FIGURE 8. Residual Analysis of ARIMA(1, 1, 0)

ARIMA(1,1,0) with Drift Model Calculation

The selected time series model is ARIMA(1,1,0) with drift, represented as:

$$\nabla y_t = c + \phi_1 \nabla y_{t-1} + \epsilon_t \tag{10}$$

Where:

- $\nabla y_t = y_t - y_{t-1}$ is the first difference
- $c = 22\,691.232$ is the drift term (constant trend)
- $\phi_1 = 0.5261$ is the autoregressive coefficient
- $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$ is white noise

3.3. **Cubic Spline Regression.** *Figure 9* presents the cubic spline regression applied to the log-transformed real estate prices over time. Log transformation was applied to stabilize variance and ensure a smoother, more interpretable spline fit, especially in periods of rapid price growth.

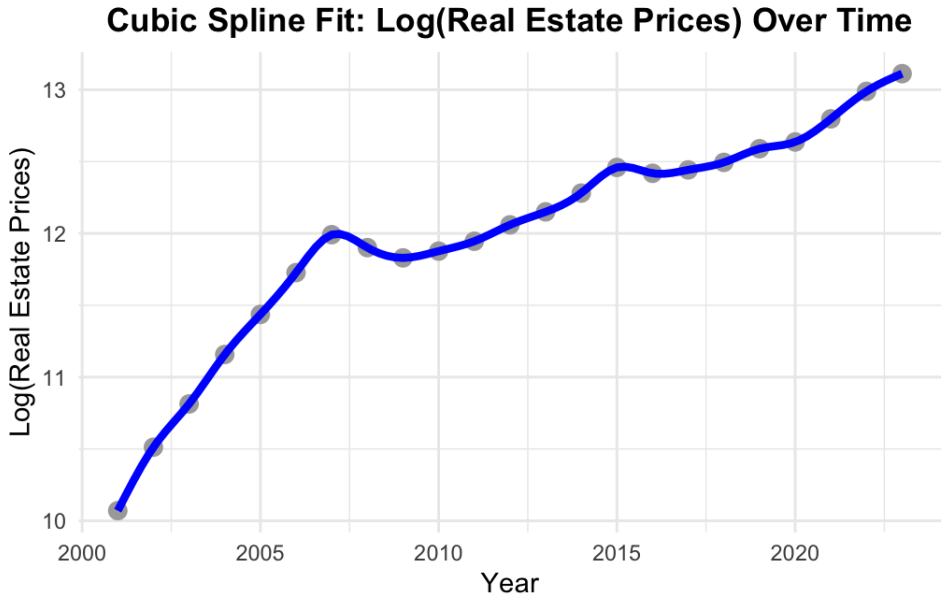

FIGURE 9. Cubic Spline Regression

The smooth blue spline curve effectively captures the nonlinear trend, highlighting periods of rapid growth (e.g., early 2000s and post-2020) and relative stability (e.g., 2009–2010 and 2016–2018).

```
        Durbin-Watson test

data:  residuals_log ~ lag(residuals_log, -1)
DW = 2.086, p-value = 0.7041
alternative hypothesis: true autocorrelation is not 0
```

FIGURE 10. Durbin-Watson Test

DW value is very close to 2, which suggests no autocorrelation in the residuals. P-value is high, so we fail to reject the null hypothesis. Residuals show no time-based autocorrelation, supporting the model's validity.

4. Forecast Calculation for 2024

4.1. Given Data.

4.1.1. *Data for Regression Models:*

- $population_{2024} = 20\,182\,003$
- $exchange\_rate_{2024} = 523.54$
- $wage_{2024} = 434\,982$
- $GDP\_per\_Capita_{2024} = 7\,736\,874.12$
- $\log(\text{real estate price})_{2023} = 13.11211$
- $\log(\text{real estate price})_{2024}^{(\text{extrapolated})} = 13.21732$

4.1.2. *Data for Time Series Model:*

- $y_{2023} = 494\,898$
- $y_{2022} = 437\,459$
- $\nabla y_{2023} = y_{2023} - y_{2022} = 57\,439$

4.2. Forecast Computation.

4.2.1. *Time Series Model (ARIMA).* Forecasted change in price for 2024:

$$\begin{aligned}\nabla\hat{y}_{2024} &= c + \phi_1 \nabla y_{2023} = 22\,691.232 + 0.5261 \times 57\,439 \\ &= 22\,691.232 + 30\,221.658 = 52\,912.89\end{aligned}$$

Forecasted price:

$$\hat{y}_{2024} = y_{2023} + \nabla\hat{y}_{2024} = 494\,898 + 52\,912.89 = 547\,810.89$$

4.2.2. *Regression Model.* The real estate price for 2024 was calculated by substituting the independent variable values into the multiple linear regression equation (9). The resulting forecasted price is:

$$\hat{y}_{2024}^{\text{regression}} \approx 387\ 712.8$$

4.2.3. *Cubic Spline Model.* An extrapolated value of the logarithm of real estate price for 2024 was obtained:

$$\log(\text{real estate price})_{2024} = 13.21732$$

Taking the exponential gives:

$$\text{real estate price}_{2024} = \exp(13.21732) \approx 549\,935.8$$

4.3. Forecast Result.

Metric	Regression	Time Series	Cubic Spline
Forecasted Price (2024)	387,713	547,811	549,936
Actual Price (2024)	500,198	500,198	500,198
Relative Error	22.4%	9.52%	9.9%

Table 1. Forecasted vs Actual Price for 2024

5. Conclusion

Comparing the forecast results of the three models, it is clear that ARIMA(1,1,0) produces the most accurate prediction with the lowest relative error. Although the cubic spline regression generated a similar result, spline models are generally less reliable for long-term forecasting, as they are not well-suited for extrapolating beyond the range of observed data. Linear regression model gives a good starting point for understanding how macroeconomic factors affect real estate prices in Kazakhstan, especially with its detailed treatment of the exchange rate. However, the large gap between the predicted and actual price for 2024 shows that some important factors might be missing. To improve its accuracy, the model could be expanded by adding variables like interest rates or government housing policies, and by taking into account delayed effects or how the variables interact with each other. It may also help to try different model types that can better reflect shifts in the market or changes in the economy, since the linear regression may have limited capacities.

References

[1] Bao, H. X. H., & Wan, A. T. K. (2004). *On the Use of Spline Smoothing in Estimating Hedonic Housing Price Models: Empirical Evidence Using Hong Kong Data*. Real Estate Economics, 32(3), 487–507.

[2] Choi, Y. (2022). *Forecasting Housing Prices: A Comparison of Time Series Models*. International Journal of Housing Studies, 45, 101–115.

[3] Cohen, V. & Kapaviciute, L (2017). The Analysis of the Determinants of Housing Prices. *Independent Journal of Management & Production, 8, 1*. http://www.redalyc.org/articulo.oa?id=449549996005

[4] Cryer, J. D., & Chan, K.-S. (2008). *Time Series Analysis: With Applications in R* (2nd ed.). Springer.

[5] Del Giudice, V., Manganelli, B., & De Paola, P. (2017). *Hedonic Analysis of Housing Sales Prices with Semiparametric Methods*. International Journal of Agricultural and Environmental Information Systems, 8(2), 65–77. https://www.igi-global.com/gateway/article/179584

[6] Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2021). *Regression: Models, Methods and Applications* (2nd ed.). Springer.

[7] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

[8] Johnson, R. (2021). *Housing Market Volatility and Time Series Models: A Comparative Study*. Journal of Real Estate Finance, 38(3), 205–220.

[9] Kutner, M., Nachtsheim, Ch., Neter, J. & Li, W. (2005). *Applied Linear Statistical Models* (5th edition). McGraw-Hill/Irwin.

[10] Wu, Z. (2025).Time Series Forecasting of Texas Housing Prices: A Comparison Between the ARIMA and VAR Models. *Theoretical and Natural Science, 80, 20-27*. https://doi.org/10.54254/2753-8818/2025.GL19918

[11] Ybyraev, Zh. & Becker, Ch. (2019). Real Estate Market Evolution and Monetary Policy in Kazakhstan. *Economic Research Initiatives at Duke (ERID) Working Paper No. 287*. https://dx.doi.org/10.2139/ssrn.3409328

6. R Code

```
# Multiple Linear Regression Model
# Load necessary libraries
library(forecast)
library(tseries)
library(ggplot2)
library(readxl)
library(scales)
library(gridExtra)

# Import data from excel
data2 <- read_excel("C:/Users/User/Downloads/data.xlsx")

# Scatter plot of independent variables against the price
var_labels <- c(
  exchange_rate = "Exchange Rate (USD/KZT)",
  gdp_capita = "GDP per Capita",
  population = "Population",
  wage = "Average Wage")
plot_list <- list()
for (var in names(var_labels)) {
```

```
  p <- ggplot(data, aes_string(x = var, y = "price")) +
    geom_point(color = "steelblue", size = 2) +
    labs(
      title = paste("Scatter Plot of Price vs", var_labels[[var]]),
      x = var_labels[[var]],
      y = "Price"
    ) +
    scale_x_continuous(labels = comma) +
    scale_y_continuous(labels = comma) +
    theme_minimal()
  plot_list[[length(plot_list) + 1]] <- p
}
grid.arrange( grobs = plot_list, ncol = 2, nrow = 2)

model <- lm(price ~ poly(exchange_rate, 3, raw = TRUE) +
                  wage + population + gdp_capita, data = data)

# Residuals vs Fitted values plot
data$residuals <- residuals(model)
data$fitted <- fitted(model)
p_fitted <- ggplot(data, aes(x = fitted, y = residuals)) +
  geom_point(color = "darkred", size = 2) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(
    title = "Residuals vs Fitted Values",
    x = "Fitted Values",
    y = "Residuals"
  ) +
  scale_x_continuous(labels = comma) +
  theme_minimal()
print(p_fitted)

# Q-Q plot
residual_data <- data.frame(Fitted = data$fitted, Residuals = data$residuals)
ggplot(residual_data, aes(sample = Residuals)) +
  stat_qq(color = "steelblue") +
  stat_qq_line(color = "black") +
  labs(title = "Normal Q-Q Plot") +
  theme_minimal()

#Time Series Model
# Load necessary libraries
library(forecast)
library(tseries)
library(ggplot2)

# Create the data frame
data <- data.frame(
  year = c(2001:2023),
```

```
  prices = c(23637, 36757, 49675, 70069, 92622, 123897, 161346, 147469,
             137279, 143903, 154123, 172770, 189124, 215531, 257644,
             247364, 253242, 266863, 293518, 307600, 361269, 437459, 494898)
)

# Convert to time series object
ts_data <- ts(data$prices, start = 2001, frequency = 1)

# Plot the time series
autoplot(ts_data) +
  ggtitle("Price Time Series") +
  ylab("Price") +
  xlab("Year")

# Check for stationarity
adf.test(ts_data)
kpss.test(ts_data)

# If not stationary, difference the series
diff_data <- diff(ts_data)
autoplot(diff_data) +
  ggtitle("Differenced Price Series") +
  ylab("Price Difference") +
  xlab("Year")

# Check ACF and PACF plots
acf(ts_data, main = "ACF of Original Series")
pacf(ts_data, main = "PACF of Original Series")
acf(diff_data, main = "ACF of Differenced Series")
pacf(diff_data, main = "PACF of Differenced Series")

# Automated model selection using auto.arima
best_model <- auto.arima(ts_data,
                         stepwise = FALSE, approximation = FALSE,
                         trace = TRUE, seasonal = FALSE)

# Show the selected model
summary(best_model)

# Check residuals
checkresiduals(best_model)

# Forecast next 5 years
forecast_values <- forecast(best_model, h = 5)
autoplot(forecast_values) +
  ggtitle("5-Year Price Forecast") +
  ylab("Price") +
  xlab("Year")
```

```
#Cubic Spline Model
# Install and load packages
install.packages("scales")
install.packages("lmtest")  # Only once
library(readxl)
library(ggplot2)
library(scales)
library(lmtest)

# Load the data
df <- read_excel(
  "/Users/aishasmagulova/Desktop/data2.xlsx",
  sheet = 1
)
# Clean the data
clean_data <- na.omit(data.frame(
  year = as.numeric(df$year),
  real_estate = as.numeric(df$'real estate prices')
)
clean_data <- clean_data[
  is.finite(clean_data$year) &
  is.finite(clean_data$real_estate),
]
# Fix scaling issues in real estate prices
clean_data$real_estate <- ifelse(clean_data$real_estate < 100,
                                 clean_data$real_estate * 1000,
                                 clean_data$real_estate)

# Log-transform the price variable
clean_data$log_real_estate <- log(clean_data$real_estate)

# Fit spline on log-transformed real estate prices
log_spline_fit <- smooth.spline(
  x = clean_data$year,
  y = log(clean_data$real_estate),
  tol = 1e-6,
  cv = TRUE
)

# Set prediction year to just 2024
pred_x <- 2024

# Make prediction (still using the existing spline fit)
log_pred <- predict(log_spline_fit, x = pred_x)

# Prepare predicted data
prediction_df <- data.frame(
  year = log_pred$x,
```

```
  log_real_estate = log_pred$y
)
# Combine for extrapolation
last_year <- max(clean_data$year)
last_value <- tail(
  clean_data$log_real_estate[clean_data$year == last_year],
  1
)
extrapolation_line <- data.frame(
  year = c(last_year, 2024),
  log_real_estate = c(last_value, log_pred$y)
)

ggplot(clean_data, aes(x = year, y = log_real_estate)) +
  geom_point(color = "darkgray", size = 3) +
  geom_line(
    data = data.frame(
      x = log_spline_fit$x,
      y = log_spline_fit$y
    ),
    aes(x = x, y = y, group = 1),
    color = "blue",
    linewidth = 1.5
  ) +
  labs(
    title = "Cubic Spline Fit: Log(Real Estate Prices) Over Time",
    x = "Year",
    y = "Log(Real Estate Prices)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(
      hjust = 0.5,
      face = "bold"
    )
  )

# Perform Durbin-Watson test
dw_result <- dwtest(
  residuals_log ~ lag(residuals_log, -1),
  alternative = "two.sided"
)
print(dw_result)
```